A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users

Yaguang Liu[1], Lisa Singh[1] and Zeina Mneimneh[2]
[1]*Department of Computer Science, Georgetown University, 3700 O St., NW, Washington, DC, U.S.A.*
[2]*Survey Research Center, University of Michigan, 426 Thompson Street, Ann Arbor, Michigan, U.S.A.*

Keywords: Demographic Inference, Siamese Network, BERT, Deep Learning.

Abstract: In order for social scientists to use social media as a source for understanding human behavior and public opinion, they need to understand the demographic characteristics of the population participating in the conversation. What proportion are female? What proportion are young? While previous literature has investigated

et al., 2010; Jungherr et al., 2016; Bode et al., 2020) using social media data. Traditionally, many of these types of studies have used survey data, where the demographics of the survey respondents are self reported. As social science researchers begin using social media data instead of or in addition to survey data, they need to understand the characteristics of the population being studied. Because of the variability in features shared by users, the short length of the posts, and the noisiness of the domain, robust methods for demographic inference are challenging (Zhang et al., 2016). We study two traditionally important demographics for social science research, gender and age. Research in these areas is rich, and a number of methods have been proposed for inferring them (Hinds and Joinson, 2018; Ciot et al., 2013; Sakaki et al., 2014; *tual features* are available. Our first goal in selecting these demographics is to understand the strengths and weaknesses of different methods on the same data set across traditionally important demographics. Previous research has shown that there are linguistic differences between demographic groups (Jørgensen et al., 2015), further motivating this work.

More specifically, we investigate the following research questions: (1) Which demographics can be inferred effectively from text alone? (2) How useful are statistical features for demographic inference? (3) When are classic models sufficient for demographic inference and when are deep learning models substantially better? (4) For which demographics are words, phrases, and/or sentences most informative?

While there are different social media platforms

Liu, Y., Singh, L. and Mneimneh, Z.
A Comparative Analysis of Classic and Deep Learning Models for Inferring Gender and Age of Twitter Users.
DOI: 10.5220/0010559500480058
In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications (DeLTA 2021)*, pages 48-58
ISBN: 978-989-758-526-5

we could study, we chose Twitter for two reasons. First, Twitter is an information sharing site that enables users to engage in conversation about important topics or follow users of interest as opposed to just friends (Yu et al., 2021). Thus, analyzing Twitter data is important and likely more challenging than friendship networks. Second, Twitter data are publicly available. Because of data availability, we consider simplified versions of both gender and age. For gender, we consider the binary version of the task with only male and female since our ground truth data contains only those two classes. For age, we consider a binary task with two age categories and a multi-class version of the task with three age categories. Again, this is done to ensure that we have sufficient training data for each class.

We conduct an extensive analysis of the relationship between different types of features and different types of models. In addition to analyzing tradi-

To summarize, this paper makes the following contributions: 1) We construct a range of different types of features and show when they are useful for different demographic inference tasks. 2) We compare classic and deep learning models for two different demographic inference tasks and evaluate their performance. 3) We evaluate deep learning models incorporating different types of embeddings (both word embeddings and sentence embeddings) to understand which network constructions are most promising for the demographic inference task. 4) We make available a curated Wikidata set so other researchers have access to a reliable ground truth data set for this task.

The rest of the paper is structured as follows. In Section 2 we review relevant literature. In Section 3 we present our experimental design. Section 4 describes our dataset. In Section 5, we present our empirical evaluation. Section 6 presents our conclusions and discusses future work. Finally, we discuss ethical considerations associated with inferring demographics from Twitter data in Section 7.

2 RELATED LITERATURE

Researchers have been developing methods for inferring a number of different demographics, including age (Schler et al., 2006; Rosenthal and McKeown, 2011; Al Zamal et al., 2012; Chen et al., 2015), gender (Chen et al., 2015; Al Zamal et al., 2012; Sakaki et al., 2014; Taniguchi et al., 2015), race/ethnicity (Preoţiuc-Pietro and Ungar, 2018; Culotta et al., 2016), location (Ikawa et al., 2012; Tian et al., 2020), and education level (Culotta et al., 2015;

demographic inference. Therefore, in addition to using the previously proposed features in the classic machine learning methods, we also consider features from sequential pattern mining. Sequential pattern mining is a classic data mining technique for identifying patterns of ordered events within a data set (Agrawal and Srikant, 1995). It has been applied in many domains, and has been shown to be effective for text mining tasks (Pokou et al., 2016).

More recently, researchers have begun incorporating neural network models for inferring demographics. For example, Vijayaraghavan et al. (Vijayaraghavan et al., 2017) build a deep learning model using users' profile information, tweets, and images. Wang and colleagues (Wang et al., 2019) investigate using profile based features like name with character embedding and image embedding of profile pictures

within deep learning models and achieve state-of-the-art performance. A graph-based Recursive Neural Networks (RNN) using skip-gram embeddings is proposed by Kim et al. (Kim et al., 2017). The model incorporates not only the text of the user, but also the text of the user's network. In our scenario, we do not have access to the user's network, i.e. the followers' text. We want to consider newer methods that take advantage of pretraining, while recognizing the need to build models with limited training data that can be applied to larger social media data sets by social scientists. Therefore, in this paper we will use BERT (Devlin et al., 2018), a pretrained transformer network, that, to the best of our knowledge, has not been used this way for the demographic inference task.

While our analysis compares deep learning models incorporating word embeddings, we also explore the use of sentence embeddings. Many models have been proposed for sentence embedding. As an ex-

and sentence embeddings for this task.

Figure 1: Model Overview.

3 EXPERIMENTAL DESIGN

In this section, we present our experimental design. We begin with a problem formulation and an overview of the methodology. We then describe the feature construction and the model building in more detail. Specifics about the data set and the data preparation are presented in Section 4.1.

3.1 Problem Formulation

Suppose we are given a data set D containing a set of user profiles. Each user profile U_i contains public information shared by a user, including his/her biography and the public posts he/she shares. U_i also contains standard account information, e.g. number of followers. We represent all the information in U_i as a set of attribute-value pairings. Each attribute-value may be either a singleton, $(age, \{30\})$, or a set of values, $(location, \{\texttt{Chicago}, \texttt{NewYork}\})$. For each user U_i, we maintain a vector of feature values $\mathbf{X_i}$ derived from the attribute-value pairings and a class label y_i. Our goal is to build a classifier that uses $\mathbf{X_i}$ to infer a user demographic y_i. The demographics we attempt to predict are gender $(\texttt{male}, \texttt{female})$, binary age bin $(<= 45, > 45)$, and multi-class age bin $(<= 35, 35-55, > 55)$.

We construct a range of features. We categorize them into two groups: (1) statistical features and (2) textual features. We construct sixteen statistical features related to account usage, user network, tweet content, and tweet structure (see Table 1).

Textual features are derived from tweet text and user biographies. The types of features extracted from text vary depending upon the models being built. Figure 2 shows the different text features we consider for our two classes of models. For the classic models, we use unigrams, bigrams, or sequential patterns. We use word-level or sentence-level embeddings for the deep learning models. Recall, one of our main goals is to understand the impact of these different textual representations for our inference tasks.

For our ngram construction, we use the traditional approach of grouping a contiguous sequence of n

Table 1: Statistical features.

Category		Features
account usage statistics		number of tweets, days since first tweet, proportion of tweets posted on weekends, average number of tweets per day
network statistics		number of friends, number of followers
tweet statistics	tweet structure statistics	average number of words per tweet, average word length, vocabulary size of per tweet
	tweet content statistics	proportion of emojis in bio, proportion of hashtags in bio, proportion of punctuation in bio, proportion of emojis per tweet, proportion of hashtags per tweet, proportion of punctuation per tweet, proportion of real

terested in determining if sequential patterns that allow for gaps can further improve the performance of classic models.

We use embeddings as text features for our deep learning models. We use word embedding from GloVe (Pennington et al., 2014) and sentence embedding from BERT (Devlin et al., 2018) in different models. By considering different linguistic representations of data (bag of words, sequential patterns, word embeddings, and sentence embeddings), we can begin to gain insight into the types of linguistic features that are important and those that are not as necessary.

3.4 Learning Models

We now briefly present the classic and deep learning models used in this paper. Our goal is to conduct ex-

task of sequential pattern mining is to find all frequent sequential patterns in D (Pokou et al., 2016). In this paper, when we construct sequential pattern features we use the frequent sequential patterns as the text features.

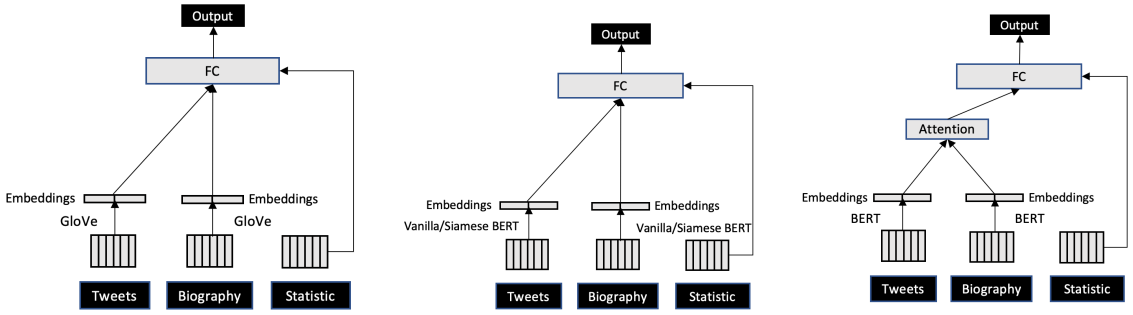
To explain why sequential pattern mining could be useful, assume we have the following two tweets from a user. 1) "The Mac is big and bright." 2)"I like the Mac which is bright." If we construct bigrams for this example, we get the following bigrams: "the mac", "mac is", "is big", "big and", "and bright", and the second tweet is parsed into "I like", "like mac", "mac which", "which is", "is bright". However, the two word phrase that contains the most similar content is "mac bright". Because that feature will be captured with sequential patterns, but not bigrams, we are interested in determining if sequential patterns that allow for gaps can further improve the performance of classic models.

limit the length of the sequential patterns to size two so that we can directly compare this approach to one involving bigrams.

3.4.2 Deep Learning Models

We consider different architectures for the deep learning models. The difference between the architectures has to do with the construction of the embedding spaces and the underlying data used, as well as the inclusion of an attention layer in some of the models. Figure 3 shows the components of each model.

Word Embedding Model: Previous literature that employed deep learning models for demographic inference used character or word embeddings for the embedding layer of the neural network (Kim et al., 2017; Wang et al., 2019). We do the same. We use



(a) Word Embedding Model (b) Vanilla/Siamese BERT Model (c) Attention-based Model

Figure 3: Illustration of different models.

the pretrained GloVe model (Pennington et al., 2014) as the embedding layer. In this model, each word is mapped into a vector and the posts/tweets of a user

The rest of the architecture is the same as that of the word embedding model (see Figure 3b). **Siamese-Network Model:** While sentence embeddings help

uncased BERT-Base model to generate sentence embedding for each tweet by averaging the BERT output layer without fine-tuning.[3] We represent tweet features as a vector by summing the tweet embeddings.

sentence pairs annotated with textual entailment information. With the fine-tuning, we are able to better represent similar sentences with similar embeddings. The classification objective function is defined as

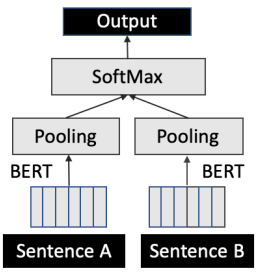


Figure 4: Siamese Network architecture.

[1]We have also tried to average it for all of our deep learning models. However, due to the sparsity of the data set, averaging leads to less variation in the features for different users. Summing has a higher variance and therefore, a better overall performance.

[2]We pause to mention that we considered some different configurations and found this one to be effective. We leave a more extensive analysis of other configurations for future work.

[3]The uncased BERT-Base model was pretrained on the BookCorpus dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers).

Table 2: Cosine similarity comparison using different embedding strategies between tweet 1 and tweet 2.

Tweet 1	Tweet 2	Similarity (Vanilla BERT)	Similarity (Siamese-Network BERT)
No, I don't know	I do not know	0.753	0.955
I am a powerful guy	I am a energetic man	0.818	0.848

follows:

$$o = softmax(W^{(t)}(a, b, |a-b|)) \quad (1)$$

where $o \subset R^k$ and $a \subset R^n$ and $b \subset R^n$ are sentence embeddings with the element-wise difference $|a-b|$. They are multiplied with the weight $W^{(t)} \subset R^{3n \times k}$. Here, R is the real numbers and n is the dimension of the sentence embeddings and k is the number of labels.

Table 3: Ground truth data distribution for gender and age.

Demographics		Category	Count
Gender		Male	10041
		Female	4274
Age	Bin 2	<=45	9689
		>45	4626
	Bin 3	<=35	6695
		35-55	4068
		>55	3552

4 DATA PREPARATION

This section begins by describing the ground truth data we use. We then discuss our approach for data labeling and data preprocessing.

model on subsets of information we expect to be more informative. Our attention-based models incorporate an attention mechanism for the user biography and the tweets. We accomplish this by multiplying the feature vectors by a modality weight in the attention layer. The attention over different modal features are computed as follows:

$$\alpha = softmax(W^{(1)} tanh(W^{(0)} M + b(0)) + b(1)) \quad (2)$$

where tweet and bibliographic features are concatenated to form a matrix $M \subset R^{n \times 2}$, $\alpha \subset R^{1 \times n}$, and $b(0)$ and $b(1)$ are the bias terms. Figure 3c depicts the attention model.

posted at least 20 tweets. After removing inactive or private accounts, and those accounts with less than 20 tweets,[5] we are left with 14,315 accounts[6] and 8.8 million tweets for age and gender. Table 3 shows the gender and binned age distributions. For gender, we can directly use the values provided by Wikidata. We have 10,041 male users and 4,274 female users. Because we have many more male users, we randomly sample from the group in order to have more balanced training data. While this may not be important for all

[4]https://github.com/siznax/wptools/wiki

[5]We have empirically found this to be the lowest number of posts that lead to reliable results.

[6]These data can be found at https://portals.mdi.georgetown.edu/public/demographic-inference-wikidata

of our models, it is important for a number of the classic ones.

In the case of age, we need to bin the continuous variable because the number of samples for some of the distinct values is too small. We use two different bin groupings, 2-bin, 3-bin. According to the Levinson adult development model (Levinson, 1986), age 45 defines a new era of adulthood. Therefore, this is what we use for our 2-bin model. For our three bin model, we worked with social science experts to identify meaningful bins that were also relatively balanced.

4.2 Data Preprocessing

We identify English tweets using the language attribute provided by the Twitter API for each tweet. To capture the different writing styles and content, we

Recall that the four classic methods in our experimental evaluation are logistic regression (LR), support vector machine (SVM), Multinomial Naive Bayes (MultiNB), and decision trees (DT). Based on a sensitivity analysis, we have a threshold that removes ngrams with a frequency support less than 0.003. For the sequential pattern models (SPM), the minimum frequency support is also set to be 0.003. The maximum length for a pattern is set to be 2 since we only consider unigrams and bigrams for classic models. For the deep learning models, the learning rate is set to be 0.0001. We use 4 NVIDIA Tesla P4 GPUs with each having 2560 CUDA Cores and 6 GBs of memory.

For all of our experiments, we use 10-fold cross validation for training and have a separate holdout set. We show the average 10-fold cross validation results, as well as the results from the holdout test set. We conducted an extensive sensitivity analysis for each model (see Appendix) and present the results for the best parameter settings for each configuration. Both the training data and testing data are balanced to avoid training and evaluation inaccuracies that could result from imbalanced data. The evaluation metric we present is the Macro-F1 measure.

5.2 Experimental Results

Table 4 presents a comparison of all the methods and feature combinations. The table is divided into seven groups: the classic models using unigram text features (Unigram-), the classic models with unigram and bigram features (Bigram-), the classic model with unigram and sequential pattern min-

similarity and these improvements are within the interval range. Using sequential pattern features within the classic models does not seem to improve the classic models. Among all of the classic models, the best one is logistic regression with bigrams, achieving a F1 score of 0.836.

The strongest models for gender are the deep learning models. We see that all the models except word embeddings perform better than the classic models with improvements ranging from 3% to 7% when compared to the best classic models in each feature group. The Word Embedding model has a comparable F1 score to the best classic models. The Vanilla BERT Sentence Embedding models perform 3% to 4% better than the Word Embeddings model. The Siamese Network models are the best performers, and the Siamese Network model with Attention

Table 4: F1 score for gender and age.

	Gender		Age (2 bins)		Age (3 bins)	
Model	**95% CI**	**Test**	**95% CI**	**Test**	**95% CI**	**Test**
Unigram-LR	0.835±0.006	0.834	0.811±0.007	0.796	0.674±0.014	0.673
Unigram-SVM	0.831±0.007	0.822	0.790±0.007	0.778	0.645±0.019	0.646
Unigram-MultiNB	0.733±0.012	0.734	0.742±0.007	0.723	0.574±0.017	0.576
Unigram-DT	0.787±0.009	0.793	0.764±0.009	0.767	0.602±0.016	0.587
Bigram-LR	0.825±0.005	0.836	0.819±0.011	0.821	0.679±0.011	0.685
Bigram-SVM	0.819±0.008	0.829	0.789±0.006	0.800	0.640±0.014	0.635
Bigram-MultiNB	0.741±0.009	0.745	0.754±-0.010	0.757	0.594±0.011	0.597
Bnigram-DT	0.786±0.007	0.805	0.761±0.015	0.773	0.601±0.016	0.591
SPM-LR	0.836±0.007	0.821	0.815±0.008	0.817	0.667±0.011	0.685
SPM-SVM	0.834±0.007	0.827	0.792±0.007	0.816	0.630±0.007	0.653
SPM-MultiNB	0.736±0.009	0.740	0.749±0.011	0.745	0.582±0.013	0.581
SPM-DT	0.791±0.008	0.770	0.764±0.010	0.779	0.607±0.012	0.587
Word_emd MLP	0.840±0.011	0.838	0.813±0.008	0.819	0.655±0.014	0.680
Bert_emd MLP	0.872±0.011	0.869	0.827±0.014	0.837	0.681±0.011	0.683

that their results are comparable to the best classic models. The Siamese BERT Sentence Embeddings with attention is the best deep learning model, and its performance is 2.5% better than logistic regression. For the 3-bin case, logistic regression is again higher than other classic models. The best deep learning model is the Siamese Network model with Attention. Once again, it is comparable to the best classic model. Overall, the worst classic models are around 10% lower than the best models. The worst neural network model is only 2% worse than the best one. The best classic model and the best neural network model are comparable with F1 scores within 2% of each other. This is a case where the simpler model is sufficient.

w/o statistical	0.871

The ablation results showing the F1 score for gender using our Siamese model is presented in Table 5. Compared to the full model, we see that removing the tweet text reduces the F1 score by over 17%. Removing the biography data or the statistical features does not have as significant an impact for inferring gender. It is likely that the tweet text is capturing important components of the other features when they are all used together.

6 CONCLUSIONS AND FUTURE WORK

In this paper we investigated the demographic inference on Twitter by using a large number of text features with a variety of classic and deep learning models to infer gender and age. Returning to the questions posed in the introduction, we found that (1) both of the demographics can be inferred effectively from text data using the proposed models, with the binary demographic inference tasks having an F1 score above 80%; (2) sequential patterns perform similarly to the unigrams and bigrams model for gender and age, (3) statistical features have the least impact on the overall performance of the model; (4) classic models are sufficient for age inference, but not as strong as the deep learning models for gender; and (5) the Siamese network architecture with attention to tweets and bi-

While automated methods can be valuable for understand differences in attitudes and beliefs among those on social media, error does exist in these models and there are possible equity and justice related consequences to imbalances in these errors. Getting informed consent in a social media domain is complicated when considering a data stream with millions of users. Whether or not public social media data should be used for research is an open question that Institutional Review Boards (IRBs) are not handling consistently. What is clear is that any usage of these data should be to advance research and should not compromise reasonable expectations of privacy. We do have an IRB exemption for this research from our institution.

Because our base data set is a Wikimedia data set, the handles we have were shared publicly by the individuals who created their Wikipedia pages. Therefore, we will share them with other researchers working on similar projects. However, we will not publicly post them because of Twitter's privacy policy and ethical concerns. Finally, we know that our sample data set is not representative of the general population. We do balance all of our data sets for training our models and will continue to try to improve our ground truth data so that it is more representative, thereby creating more general purpose inference models.

ACKNOWLEDGEMENTS

This work is funded by National Science Foundation awards #1934925 and #1934494, the National Collaborative on Gun Violence Research (NCGVR) and the Massive Data Institute (MDI) at Georgetown Univer-

Probabilistic inference of twitter users' age based on what they follow. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Chen, X., Wang, Y., Agichtein, E., and Wang, F. (2015). A comparative study of demographic attribute inference in twitter. In *AAAI Conference on Weblogs and Social Media(ICWSM)*.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of twitter users in non-english contexts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Culotta, A., Kumar, N., and Cutler, J. (2015). Predicting the demographics of twitter users from website traffic data. In *Association for the Advancement of Artificial Intelligence*.

Culotta, A., Ravi, N., and Cutler, J. (2016). Predicting twitter user demographics using distant supervision from

website traffic data. *Journal of Artificial Intelligence Research*.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.

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

Dredze, M. (2012). How social media will change public health. *IEEE Intelligent Systems*.

Hinds, J. and Joinson, A. (2018). What demographic attributes do our digital footprints reveal? a systematic review. *PloS one*.

Ikawa, Y., Enoki, M., and Tatsubori, M. (2012). Location inference using microblog messages. In *International Conference on World Wide Web*.

Jørgensen, A., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social

AAAI Conference on Weblogs and Social Media (ICWSM).

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Conference on empirical methods in natural language processing (EMNLP)*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pokou, Y., Fournier-Viger, P., and Moghrabi, C. (2016). Authorship attribution using small sets of frequent part-of-speech skip-grams. In *International Flairs Conference*.

Preoţiuc-Pietro, D. and Ungar, L. (2018). User-level race and ethnicity predictors from twitter text. In *Conference on Computational Linguistics*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., and Rosenquist, J. (2011). Understanding the demographics of twitter users. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.

Nguyen, D., Gravel, R., and Trieschnigg, D.and Meder, T. (2013). " how old do you think i am?" A study of language and age in twitter. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.

Nguyen, D., Smith, N., and Rose, C. (2011). Author age prediction from text using linear regression. In *ACL-HLT workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pennacchiotti, M. and Popescu, A. (2011). A machine learning approach to twitter user classification. In

(2014). Twitter user gender inference using combined analysis of text and image processing. In *Workshop on Vision and Language*.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. (2006). Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs*.

Sinnenberg, L., Buttenheim, A., Padrez, K., Mancheno, C., Ungar, L., and Merchant, R. (2017). Twitter as a tool for health research: A systematic review. *American Journal of Public Health*.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*.

Taniguchi, T., Sakaki, S., Shigenaka, R., Tsuboshita, Y., and Ohkuma, T. (2015). A weighted combination of text and image classifiers for user gender inference. In *Workshop on Vision and Language*.

Tian, H., Zhang, M., Luo, X., Liu, F., and Qiao, Y. (2020). Twitter user location inference based on representation learning and label propagation. In *Proceedings of The Web Conference*.

Vijayaraghavan, P., Vosoughi, S., and Roy, D. (2017). Twitter demographic classification using deep multi-modal multi-task learning. In *Annual Meeting of the Association for Computational Linguistics*.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*.

Wang, Z., Hale, S., Adelani, D., Grabowicz, P., Hartman, T., Flock, F., and Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*.

Williams, A., Nangia, N., and Bowman, S. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Williams, J., Singh, L., and Mezey, N. (2019). # metoo as

Table 6: Best parameter settings for gender.

Features	Gender
Unigram-LR	C:0.5, penalty:none
Unigram-SVM	kernel: linear
Unigram-NB	alpha:0.5, fit-prior:False
Unigram-DT	criterion:gini, max-depth:11
Bigram-LR	C:0.5, penalty:non
Bigram-SVM	kernel:linear
Bigram-NB	alpha:0.5, fit-prior:False
Bigram-DT	criterion:entropy, max-depth:11
SPM-LR	C:0.5, penalty:none
SPM-SVM	kernel:linear
SPM-NB	alpha:0.5, fit-prior:False
SPM-DT	criterion:entropy, max-depth:11
Word_emd MLP	epoch 500, lr 0.001
Bert_emd MLP	epoch 1500, lr 0.0001
Siamese_emd MLP	epoch 1500, lr 0.0001
Siamese_emd Attention	epoch 1500, lr 0.0001

Table 7: Best parameter settings for age (2-bin).

Unigram-NB	alpha:0.5, fit-prior:False
Unigram-DT	criterion:gini, max-depth:11
Bigram-LR	C:0.5, penalty:non
Bigram-SVM	kernel:linear
Bigram-NB	alpha:0.5, fit-prior:False
Bigram-DT	criterion:entropy, max-depth:11
SPM-LR	C:0.5, penalty:none
SPM-SVM	kernel:linear
SPM-NB	alpha:0.5, fit-prior:False
SPM-DT	criterion:entropy, max-depth:11
Word_emd MLP	epoch 500, lr 0.001
Bert_emd MLP	epoch 1500, lr 0.0001
Siamese_emd MLP	epoch 1500, lr 0.0001
Siamese_emd Attention	epoch 1500, lr 0.0001

Table 6 - 8 shows the best parameters for gender, binary age, and muli-class age, respectively.

We use 10-fold cross validation with the dataset and determine the best parameters by evaluating the F1 score for each model. We then apply those parameters to our holdout test data set.