information

Article

Demographics and Personality Discovery on Social Media: A Machine Learning Approach

Sarach Tuomchomtam and Nuanwan Soonthornphisaj *

Artificial Intelligence and Knowledge Discovery Laboratory, Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand; sarach.t@ku.th
* Correspondence: nuanwan.s@ku.th

Abstract: This research proposes a new feature extraction algorithm using aggregated user engagements on social media in order to achieve demographics and personality discovery tasks. Our proposed framework can discover seven essential attributes, including gender identity, age group, residential area, education level, political affiliation, religious belief, and personality type. Multiple feature sets are developed, including comment text, community activity, and hybrid features. Various machine learning algorithms are explored, such as support vector machines, random forest, multi-layer perceptron, and naïve Bayes. An empirical analysis is performed on various aspects, including correctness, robustness, training time, and the class imbalance problem. We obtained the highest prediction performance by using our proposed feature extraction algorithm. The result on personality type prediction was 87.18%. For the demographic attribute prediction task, our feature sets also outperformed the baseline at 98.1% for residential area, 94.7% for education level, 92.1% for gender identity, 91.5% for political affiliation, 60.6% for religious belief, and 52.0% for the age group. Moreover, this paper provides the guideline for the choice of classifiers with appropriate feature sets.

Keywords: demographic attributes; personality prediction; social media; machine learning

Citation: Tuomchomtam, S.; Soonthornphisaj, N. Demographics and Personality Discovery on Social Media: A Machine Learning Approach. *Information* **2021**, *12*, 353. https://doi.org/10.3390/info12090353

Academic Editor: Arkaitz Zubiaga

Received: 13 August 2021
Accepted: 24 August 2021
Published: 30 August 2021

1. Introduction

User demographic attributes and personality type (collectively called "private attributes") can be applied in several domains, for example, hate speech detection [1] and product recommendation [2] using additional demographic data. The ability to identify personality is useful for better understanding ourselves and others. For instance, we can choose an appropriate field of study that fits our personality or apply for a job that best fits our preferences. On the other hand, it can also be applied by recruiters to find appropriate applicants that fit the job description [3]. Persuasive mass communication is another benefit of personality discovery. It aims at encouraging large groups of people to believe and act on the communicator's viewpoint. It is used by governments to encourage healthy behaviors, by marketers to acquire and retain consumers, and by political parties to mobilize the voting population [4].

Myers–Briggs Type Indicator (MBTI) [5] is a well-established personality model that describes the characteristics of an individual using four dichotomous attributes: (1) Main focus or favorite world: people who prefer to focus on the outer world then have the *extraversion* characteristic (E). Otherwise, if they prefer their inner world, then they have the *introversion* characteristic (I). (2) The way people process their information: if they prefer to focus on basic information, then they have the *sensing* characteristic (S). If they prefer to interpret and add meaning or they seek creative solutions to problems, then they are *intuitive* people (N). (3) Decision-making method: if they use logic or fairness in making a decision, then they have a *thinking* characteristic (T). If they decide by first looking at the people and circumstances, then they are sensitive and have the *feeling* characteristic (F). (4) The way people deal with the outside world: if they prefer to be decisive and well organized then they have *judging* behavior (J). If they are flexible and willing to stay open

to new information or options, this means that they have the *perceiving* characteristic (P). Moreover, the combinations of MBTI types can be integrated to fit the personality type of individuals; they are ENFJ, ENFP, ENTJ, ENTP, ESFJ, ESFP, ESTJ, ESTP, INFJ, INFP, INTJ, INTP, ISFJ, ISFP, ISTJ, and ISTP. This model has been widely used in many practical applications despite its validity and reliability [6]. Additionally, it has been shown that MBTI attributes can be correlated with ones from the Big Five model [7,8].

To discover their personality types, users are required to explicitly provide their information to the MBTI instrument by filling out a multiple-choice questionnaire. The online assessment is time-consuming and may lead to a response bias problem. Previous studies have investigated the potential of machine learning algorithms in demographic and personality attribute classification. On Facebook, the features can be extracted from both texts (e.g., posts and comments) and activity (e.g., likes). Kosinski et al. [9] extracted users' liked pages as feature sets for predicting private attributes and personality traits. The singular value decomposition technique was deployed to solve the dimensionality problem by reducing the dimension of the user-like matrix before training with logistic and linear regression. On Twitter, the textual information and network relation (e.g., followings and followers) are applied as feature sets. For example, Aletras and Chamberlain [10] created a graph that represents user relations to predict socioeconomic attributes. They suggested that the combination of textual features and graph embeddings provides a significant improvement over the use of either alone. On Instagram, the features are extracted from images and sometimes from other information, such as likes. For example, Ferwerda and Tkalcic [11] proposed the use of both visual and content features extracted from pictures to predict the user's personality type. On Reddit, the features are mainly based on text content (e.g., posts and comments). Gjurković and Šnajder [12] proposed the use of text in the user's posts, comments, and other metadata for predicting MBTI personality types. Multilayer perceptron and logistic regression algorithm are applied and obtained 76% of the macro average of F_1 score.

We analyzed the concept of MBTI and found that the personality type model is related to social behaviors. Therefore, it is in our interest to explore the possibility of social media contributing to personality prediction. Therefore, Reddit, a well-known social media platform, was our main focus since its users are organized as members of communities (called "subreddits"). Each text post is attached with user information. Most Reddit users are anonymous. However, some users declare themselves by published short tags called "author flairs" next to their names. Our contributions are as follows.

1. We propose methods for extracting demographic and personality attributes from Reddit users using author flairs.
2. Multiple feature sets are also proposed and explored by machine learning algorithms to find the best-performing combinations.
3. To validate our experimental results, processed author flairs are applied as ground truth for the training and testing process.

2. Materials and Methods

2.1. Experimental Data

This section introduces the characteristics of Reddit posts. Figure 1 illustrates a Reddit post in the community, namely, "datingoverthirty". Author flairs are community-specific descriptors that some users apply to describe themselves to other members of the communities ("subreddits"). Figure 2 visualizes the element of the text post that consists of the author's name, attached with the short tag called author flair. We can see that the author clarifies her gender and age on the author flairs as "♀36", which means that she is a 36-year-old woman. However, author flair is not required by Reddit; hence, most users are anonymous to the system. We found that Reddit does not summarize user profiling, which is different from Facebook.

We obtained comments made in August, September, October of 2018 from the Pushshift website, which maintains a publicly accessible database of various Reddit data, includ-

ing submissions and comments. The obtained data contains 300,877,224 comments from 8,131,714 users in 177,116 communities. Each comment includes an author's name, author flair, community name, and text body. Note that we respect user privacy; therefore, data anonymization is performed on user identity before the experiment setup.

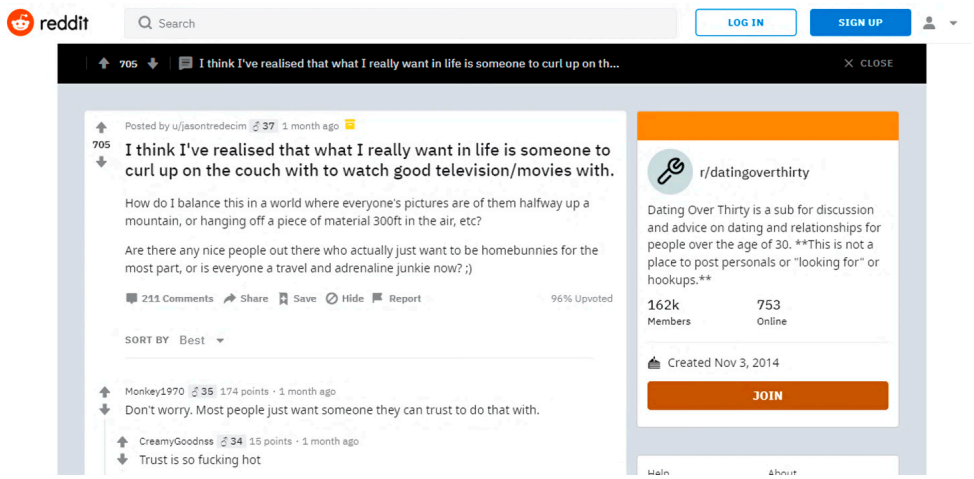

Figure 1. A Reddit post in the "datingoverthirty" community.

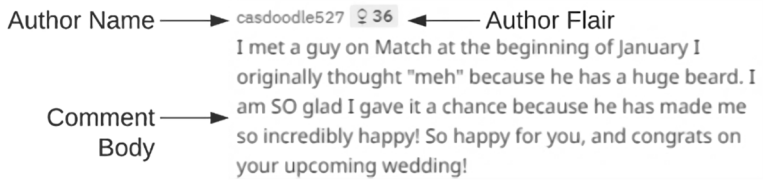

Figure 2. A Reddit comment.

2.2. Framework Overview

There are three main processes in our proposed framework, which are: (1) private attributes extraction, (2) feature extraction, and (3) private attribute prediction. The private attribute extraction takes in the user description dataset and outputs private attribute datasets. The feature extraction takes in the comment text dataset and outputs the extracted feature sets. The private attribute prediction takes in the feature sets and the attributes for the classification. Figure 3 illustrates our framework. We conducted our experiments using 64-bit Python 3.6 on a Linux system with Intel Xeon Gold 6130 and 250 GB of memory.

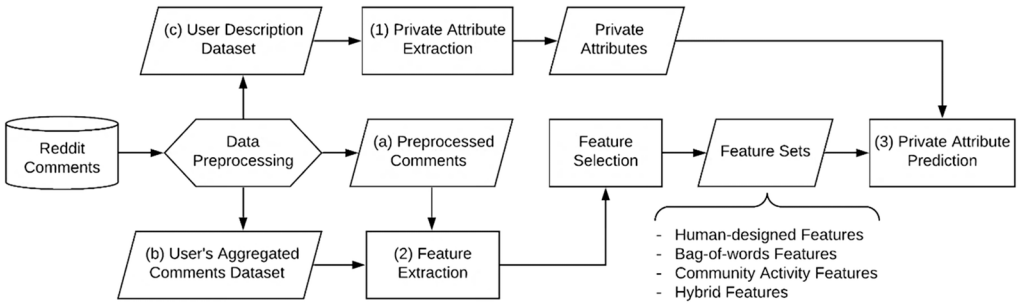

Figure 3. Framework diagram.

2.3. Data Preprocessing

Given the Reddit comment dataset, the framework generates three new datasets, which are: (a) preprocessed comments, (b) user's aggregated comments dataset, and (c) user description dataset. Note that dataset (a) and (b) will be used for feature extraction. Dataset (c) will be used for private attribute extraction.

To pre-process the user's comments, we first converted the body text to lower cases, then tokenization was performed by replacing URLs, user names, community names, HTML characters, elongated words, and numbers. The token "xxeos" marks the end of the sliding window for n-gram feature extraction. Table 1 depicts the replacement tokens and their descriptions. We then expanded word contractions and removed unwanted punctuation and extra white spaces in the comment's body.

Table 1. Replacement tokens and their descriptions.

Replacement Token	Description
xxurl	URL
xxuser	Author name
xxsub	Community name
xxrep	Repeated word
xxelon	Elongated word
xxd	One-digit number
xxdd	Two-digit number
xxddd	Three-digit number
xxdddd	Four-digit number
xxddddd	Five-digit number
xxeos	The end of a comment

Table 2 shows a sample of the pre-processed comments. Finally, the user's engagements on Reddit were obtained by aggregating all comments posted by each user as a document. A sample of the pre-processed comment text dataset is shown in Table 3.

Table 2. A sample of the pre-processed comments. Bold font indicates the replacement tokens used to replace some segments in the comment.

Author Name	Community	Comment Body
######	AMD_Stock	this is all i could find, fwiw **xxurl** it does line up with the expected end...
######	TributeMe	this is my favorite yet. if you pm me more, i will tribute it. **xxurl**
######	news	the trolls from t d constantly brigade astroturf **xxsub** in a bid to control t...
######	squirting	source **xxurl** gif starts at **xxd xxdd xxdd**
######	AskReddit	somebody already did your job. something for you to read **xxurl**
######	Windows10	**xxurl xxelon xxurl** basically, instead of white, we get that subdued color...
######	tifu	thank you for submitting to **xxsub, xxuser**. your submission, tifu by oblit...
######	sydney	i figure this clears **xxelon xxuser xxelon xxuser, xxuser xxelon xxuser** . . .
######	CxTV	this thread was crossposted from **xxurl** made by **xxuser**. to mute xpost . . .
######	woooosh	far from heaven, **xxsub**! **xxsub** is **xxelon** better

Table 3. A sample of the user's aggregated comments dataset obtained from the users.

Author Name	Aggregated Comments
######	sure, but this is not **xxsub**, it is **xxsub**. posts need to demonstrate they...
######	**xxurl xxeos xxurl xxeos** philosophy mainly, or fiction that tends to be phil...
######	it is happened before, it will happen again not mine, **xxuser** s **xxeos** i wan...
######	from **xxurl** appropriate swim attire required, cotton shorts or shirts, spor...
######	you could try posting to **xxsub xxeos** i do not really have a time limit on da...
######	now i am just confused **xxurl** sh e **xxddd** db **xxdd xxeos** angry birds from outer...
######	you are cordially invited to **xxsub xxeos** is her cousin stan from south park...
######	here you go man. sorry, fell asleep. **xxurl** gclid cj **xxd** kcqjw **xxdd** bbrd a...
######	does not look like the **xxsub** has a chat. maybe try pming the mods and sugge...
######	hkj reviews of aa nimh chargers **xxurl** this sofirn looks ok not many chargers...

The user description dataset contains the author's name, community name, flair CSS class, and flair text. We then removed the duplicated descriptions for each author–community pair. Table 4 shows a sample of the preprocessed user description dataset.

Table 4. A sample of the preprocessed user description dataset.

Author Name	Community Name	Flair Class	Flair Text
######	AskALiberal	-	Centrist Democrat
######	atheism	no-knight	Atheist
######	ConservativesOnly	-	McCarthy did nothing wrong
######	Conservative	Conservative	Conservative
######	sexover30	male	♂50
######	Judaism	Orange	converting Conservative
######	Christianity	chirho	Christian (Chi Rho)
######	Christianity	coeusa	Episcopalian (Anglican)
######	datingoverthirty	male	♂Forty Minus One
######	datingoverthirty	female	♀32

2.4. Private Attribute Extraction

2.4.1. Gender Identity

For gender identity, we looked for users who identify themselves as male or female in gender-related communities and found that multiple patterns are representing male and female values, for example, users with "male" and "female" flair class or flair text with these regular expression patterns: "[♂♀] ?\d{2}" and "\d{2}[MF].+", such as "♂34" (34-year-old male) and "23/F/5′10" (23-year-old female with the height of 5′10"). We post-processed all variations into uniform values of "male" and "female". We performed random under-sampling on the male class to reduce the size of the dataset due to hardware limitations in our experiment.

2.4.2. Age Group

For age groups, we looked for users who specified themselves with two-digit ages in their descriptions, which happened to be the same patterns as gender identity, in age-related communities. We excluded users at the age of sixty-five and above because they were virtually non-existent on the website. The age attributes were segmented into four classes often used in demographic targeting, including teenagers (15–19), young adults (20–34), younger middle-aged (35–49), and older middle-ages (50–64). For the teenage group, we also looked for users who specified their age with "^(\d{2})$" flair pattern. We performed random under-sampling on the teenager class.

2.4.3. Residential Area

To extract the user's residential areas, we focused on national and continental communities. We segmented them into eight regions, including North American, European, South American, South Asian, Southeast Asian, East Asian, Middle Eastern, and African. We performed random under-sampling on the European class to match the North American class to reduce the size of the dataset.

2.4.4. Education Level

For education level, we focused on three groups: high school, undergraduate, and graduate. Hence, we looked for degree names or fields of study in communities related to education. However, we found a low number of users for the high school class; therefore, we used numeric age descriptions in the "teenagers" community, which is the largest high school community on the website, as additional information.

2.4.5. Political Affiliation

For political affiliation, there are a lot of factions both in the real world and on the website, for example, socialist, center-left, libertarian, and far-right. We only focused on liberals and conservatives, which are the biggest and clearest political groups on Reddit. We do this by looking for users with liberal and conservative flairs in mostly North American political communities.

2.4.6. Religious Belief

For religious beliefs, similar to political affiliation, there are a lot of religious factions. We looked for the six biggest beliefs in the world in communities discussing religious topics. These are Atheist, Christian, Muslim, Jewish, Buddhist, and Hindu.

2.4.7. Personality Type

For personality type, the framework searched for users whose author's flairs were one of the sixteen MBTI personality types with "^([EI][SN][TF][JP])$" pattern in community discussions about MBTI personality types. Table 5 shows the lists of communities used in the extraction.

Table 5. Communities that were explored for extracting demographic attributes.

Attribute	Communities
Gender Identity	40something, AskMen, AskMenOver30, AskWomen, AskWomenOver30, DatingAfterThirty, DirtySnapchat, GWABackstage, LGBTeens, OkCupid, RelationshipsOver35, Tinder, amiugly, asktransgender, askwomenadvice, assholegonewild, childfree, datingoverthirty, keto, loseit, sexover30, xxketo
Age Group	40something, DatingAfterThirty, LGBTeens, OkCupid, RelationshipsOver35, Tinder, childfree, datingoverthirty, keto, loseit, sexover30, teenager, xxketo, teenager
Residential Area	AskAnAmerican, Africa, Arabs, Argentina, Brazil, Cambodia, Chile, China, Colombia, Europe, India, Indonesia, Japan, Korea, Laos, Malaysia, Thailand
Education Level	GradSchool, college, teenager
Political Affiliation	AskALiberal, CanadaPolitics, Conservative, ConservativesOnly, Republican, True_AskAConservative, askaconservative, liberalgunowners, ukpolitics
Religious Belief	AskAChristian, AskReligion, Christianity, DebateAChristian, DebateAnAtheist, DebateReligion, Judaism, OpenChristian, TrueChristian, atheism, excatholic, exchristian, survivor
Personality Type	MBTI, ENFJ, ENFP, ENTJ, ENTP, ESFJ, ESFP, ESTJ, ESTP, INFJ, INFP, INTJ, INTP, ISFJ, ISFP, ISTJ, ISTP

2.5. Feature Extraction

We began by performing feature extraction, followed by feature selection. After that, we experimented with multiple classification algorithms and a couple of techniques to address the imbalance problem. Then, we evaluated the performance of each approach.

2.5.1. Human-Designed Features

We used Linguistic Inquiry and Word Count (LIWC), introduced by Tausczik and Pennebaker [13], in our experiment for the human-designed features. These are predefined categories of words that can be created as a frequency vector for a document. We also experimented with the term frequency-inverse document frequency (tf-idf) version of LIWC.

2.5.2. Bag-of-Words (BoW) Features

This is a text representation model that considers the term occurrences in the aggregated comments. We used uni-grams and bi-grams for bag-of-words features. We also experimented with both stemmed and non-stemmed words for the n-grams. Finally, we calculated tf-idf, then selected the best 20,000 n-grams based on their ANOVA F-values.

2.5.3. Community Activity (CA) Features

These features indicated user engagement in the communities on the website. This was inferred from the number of comments made in each community as activity features. Let $f_{c,u}$ be the number of comments posted by user u in community c. Let C be the set of communities in the dataset. Community activity features of user u, denoted as CA_u, can be described as follows:

$$CA_u = \{f_{c,u} \text{ for c in C}\}$$

From our statistical analysis, we found that users commented in 82 communities at the 95th percentile. Hence, for each private attribute, we also created a feature set of the best 100 communities based on their ANOVA F-values from 53,966 communities.

Nevertheless, CA_u only represented user interests; therefore, we also experimented with a weighted feature set, denoted as CA_Wtg_u, that considered the normality of other users in the dataset, which is the same concept as tf-idf. Let U be the users in the dataset. The weighted community activity of user u can be described as follows.

$$CA\_Wtg_u = \left\{ f_{c,u} \times \log\left(\frac{|U|}{|u \in U : c \in u|} \right) \text{ for c in C} \right\}$$

Algorithm 1 shows the feature extraction and selection algorithm for the features. *UseWeighted* is a Boolean parameter indicating whether to transform into a weighted vector or not. *SelectKBest* is a Boolean parameter indicating whether to perform feature selection or not. *K* is an integer parameter indicating the number of desired features. The time complexity of this algorithm is *O(n)* with *n* as the number of comments in *PreprocessedComments*.

Algorithm 1. The proposed feature extraction and selection algorithm for community activity features.

```
function ExtractActivityFeatures(PreprocessedComments, UseWeighted, SelectKBest)
  ActivityFeatures[][] = A two-dimensional array
  for each Comment in PreprocessedComments do
    User = Comment's author
    Community = Community of comment's post
    ActivityFeatures[User][Community] += 1
  end for
  if UseWeighted then
    ActivityFeatures = CalcuateWeight(ActivityFeatures)
  end if
  if SelectKBest then
    ActivityFeatures = FTest(ActivityFeatures, 100)
  end if
  return ActivityFeatures
end function
```

2.5.4. Hybrid Features (HF)

We created a combination of bag-of-words and community activity as hybrid features. We also experimented with the addition of the human-designed features and a version with 10,000 features to study the robustness of the features. Algorithm 2 shows the proposed feature extraction and selection for the features. *UseLIWC* is a Boolean parameter indicating whether to add LIWC features to the vector. The time complexity of this algorithm is

O(nmk) where *n* is the number of users, *m* is the number of features in *NgramFeatures*, and *k* is the number of features in *ActivityFeatures*.

Algorithm 2. The proposed feature extraction and selection algorithm for hybrid features.

```
function ExtractHybridFeatures(LIWCFeatures, NgramFeatures, ActivityFeatures, UseLIWC, UseWeighted, SelectKBest)
    HybridFeatures[][] = A two-dimensional array
    Users = Users in NgramFeatures and ActivityFeatures
    for each User in Users do
        for each Feature in NgramFeatures do
            FeatureValue = NgramFeatures[User][Feature]
            HybridFeatures[User][Feature] = FeatureValue
        end for
        for each Feature in ActivityFeatures do
            FeatureValue = ActivityFeatures[User][Feature]
            HybridFeatures[User][Feature] = FeatureValue
        end for
    end for
    if UseLIWC then
        for each User in Users do
            for each Feature in LIWCFeatures do
                FeatureValue = LIWCFeatures [User][Feature]
                HybridFeatures[User][Feature] = FeatureValue
            end for
        end for
    end if
    if UseWeighted then
                HybridFeatures = Tfidf(HybridFeatures)
    end if
    if SelectKBest then
        HybridFeatures = FTest(HybridFeatures, 10000)
    end if
    return HybridFeatures
end function
```

2.6. Feature Selection

Filter-based feature selection was performed on all features except for human-designed features to maximize the performance and reduce overfitting. Table 6 shows the list of feature sets used in our experiment. We used the one-way ANOVA F-test to test the relationship between predictor and response then selected the features with the highest F-value. Let f_i be the average value of feature i, $\overline{x}$ be the average value of feature averages, x_i be a value of feature i, f be the average value of the feature, and DF be the degree of freedom. F-value can be calculated as follows.

$$\text{Sum of squares between features} = SS_{between} = \sum (f_i - \overline{x})^2$$

$$\text{Sum of squares within feature} = SS_{within} = \sum (x_i - f)^2$$

$$\text{F-value} = \frac{SS_{between} \div DF_{between}}{SS_{within} \div DF_{within}}$$

Table 6. Feature sets used in private attribute prediction.

	Feature Set	#Features	Description
Baseline	LIWC	64	Human-designed LIWC frequency features.
	LIWC_Tfidf	64	Human-designed LIWC tf-idf features.
	BoW_Ngrams	20,000	Uni-grams and bi-grams tf-idf features.
	BoW_Stemmed	20,000	Stemmed uni-grams and bi-grams tf-idf features.
Proposed	CA_Freq	53,966	Community activity frequency features.
	CA_Freq_100	100	100 k-best community activity frequency features.
	CA_Wgt_100	100	100 k-best community activity weighted features.
	HF	20,100	Hybrid tf-idf features.
	HF_LIWC	20,164	Hybrid tf-idf features with LIWC tf-idf features.
	HF_10k	10,000	Top 10k hybrid tf-idf features.

2.7. Classification Algorithms

To see the potential of our proposed community and hybrid features, we performed experiments using 10-fold cross-validation on several classifiers, including multinomial naive Bayes, support vector machine, random forest, multi-layer perceptron, and majority class classifier. These classifiers will be trained with the feature sets using the extracted attributes as labels.

- Majority class classifier (MCC) always classifies the most frequent class in the dataset. This classifier is often used as the baseline against machine learning models to demonstrate their superior decision-making.
- Multinomial naïve Bayes (NB) is a popular conditional probabilistic classifier. We used one of the classic variants used in text classification with Laplace smoothing.
- Support vector machine (SVM) [14] creates a discrimination hyperplane between two sets of data points. We used linear SVM with the L2 penalization and squared hinge as the loss function. We used the one-vs-rest strategy for multi-class datasets.
- Random forest (RF) [15] is a majority-voting classifier that consists of multiple decision trees, each trained with a different dataset. We created a random forest with 100 decision trees with the maximum features equal to the square root of the original number of features.
- Multi-layer perceptron (MLP) is a fully connected artificial neural network. We used two hidden layers, each with 64 units with the rectified linear unit (ReLU) activation. We held out 10% of the training data to use as the validation set for early stopping.

2.8. Imbalance Problem

We experimented with a couple of resampling methods, including random over-sampling (RO) and synthetic minority over-sampling technique (SMOTE) [16], to address the imbalance problem and study their effects on the performance.

3. Results

We extracted seven attributes from 45,751 unique users. These were 17,589 users for gender identity, 4136 users for age group, 17,446 users for residential area, 3499 users for education level, 810 users for political affiliation, 2709 users for religious belief, and 4723 users for personality type. Table 7 shows the number of users in each class of the datasets.

Table 7. The number of users of each class for each attribute. An asterisk indicates that random under-sampling was performed on that class.

Private Attribute	#Users	Class	#Users
Gender Identity (Gen.)	17,589	Male	8797 *
		Female	8792
Age Group (Age)	4136	Young Adult	1791
		Teenager	1790 *
		Younger Middle-Aged	501
		Older Middle-Aged	54
Residential Area (Res.)	4723	North American	4967
		European	4965 *
		South American	2701
		South Asian	1770
		Southeast Asian	1738
		East Asian	799
		Middle Eastern	477
		African	29
Education Level (Edu.)	3499	High School	1787
		Graduate	1046
		Undergraduate	666
Political Affiliation (Pol.)	810	Conservative	475
		Liberal	335
Religious Belief (Rel.)	2709	Atheist	1730
		Christian	857
		Muslim	50
		Jewish	36
		Buddhist	20
		Hindu	16
Personality Type (Per.)	4723	INTP	1196
		INTJ	1078
		ENFP	529
		INFJ	504
		ENTP	374
		INFP	329
		ISTP	259
		ISTJ	94
		ENTJ	87
		ESTP	58
		ISFJ	52
		ISFP	49
		ENFJ	47
		ESFP	41
		ESFJ	13
		ESTJ	13
Introversion/Extraversion (I/E)	4723	Introversion	3561
		Extraversion	1162
Sensing/Intuition (S/N)	4723	Intuition	4144
		Sensing	579
Thinking/Feeling (T/F)	4723	Thinking	3159
		Feeling	1564
Judging/Perception (J/P)	4723	Perception	2835
		Judging	1888

3.1. Classification Performance

We obtained quite impressive and promising results for both demographic attributes and personality types. Table 8 shows the best macro average F_1 scores for each demographic attribute. We found that our proposed CA_Freq_100 feature set obtained the best performance measured in terms of F_1 score. The F_1 score of residential area prediction reached 98.1%. Applying another proposed feature set (CA_Wgt_100), the education level prediction gets an F_1 score of 94.7%. The gender identity prediction using the HF feature set obtained 92.1%. For political affiliation, we received an F_1 score of 91.5% by using the CA_Wgt_100 feature. The religious belief prediction performance was 60.6% using CA_Wgt_100, and the age group at 52.0% with the HF features. We can conclude that our proposed feature sets, CA and HF, provided the best performance contribution for predicting all demographic and personality attributes.

Table 8. The best macro average F_1 scores of the feature sets for the attributes. Bold values indicate the highest performance.

	Feature Set	Gen.	Age	Edu.	Res.	Pol.	Rel.	Per.	E/I	S/N	T/F	J/P
Baseline	None (MCC)	0.333	0.151	0.055	0.226	0.370	0.130	0.025	0.430	0.467	0.401	0.375
	LIWC	0.757	0.361	0.293	0.546	0.608	0.255	0.078	0.533	0.486	0.602	0.551
	LIWC_Tfidf	0.781	0.362	0.336	0.553	0.592	0.231	0.064	0.437	0.467	0.578	0.496
	BoW_Ngrams	0.895	0.480	0.702	0.791	0.728	0.447	0.222	0.595	0.529	0.702	0.624
	BoW_Stemmed	0.895	0.478	0.707	0.791	0.729	0.467	0.231	0.591	0.545	0.706	0.636
Proposed	CA_Freq	0.892	0.508	0.868	0.956	0.896	0.490	0.545	0.810	0.768	0.835	0.863
	CA_Freq_100	0.840	0.477	0.901	**0.981**	0.907	0.495	**0.644**	**0.871**	**0.859**	0.863	**0.886**
	CA_Wgt_100	0.880	0.498	**0.947**	0.979	**0.915**	**0.606**	0.562	0.868	0.836	**0.871**	0.878
	HF	**0.921**	**0.520**	0.854	0.907	0.877	0.531	0.511	0.760	0.721	0.808	0.801
	HF_LIWC	0.920	0.517	0.855	0.907	0.873	0.557	0.511	0.761	0.722	0.808	0.801
	HF_10k	**0.921**	**0.520**	0.856	0.907	0.877	0.562	0.515	0.759	0.722	0.809	0.816

For personality datasets, we found that our proposed feature sets significantly outperformed the baseline ($p < 0.001$). To the best of our knowledge, our methods achieved the highest performance on MBTI personality prediction for Reddit datasets. We compared our results with the work done by Gjurković and Šnajder [12], which experimented with a similar Reddit dataset. Despite having fewer instances (4723 vs. 9111) and features (100 vs. 11,140), using our proposed feature sets displayed significantly better performance. However, we were not able to experiment with their published dataset due to the lack of data for our feature extraction methods. Table 9 shows the performance comparison between [12] and our proposed methods.

Table 9. The performance comparison between previous work and our proposed methods. Bold font indicates better performance. Parentheses indicate the algorithm obtained the performance. LR stands for logistic regression.

	Our Methods	Gjurković and Šnajder [12]
Personality Type	**64.4% (NB)**	41.7% (MLP)
Introversion/Extraversion	**87.1% (MLP)**	82.8% (MLP)
Sensing/Intuition	**85.9% (RF)**	79.2% (MLP)
Thinking/Feeling	**87.1% (RF)**	67.2% (LR)
Judging/Perception	**88.6% (MLP)**	74.8% (LR)

We evaluated the performance of NB, MLP, RF, and SVM by comparing our proposed feature sets and the feature set proposed by [10]. In Table 10, 10 feature sets are evaluated on 11 private attribute predictions. We found that our proposed feature sets outperformed all baseline feature sets. For community activity features, RF mostly performed best, except MLP for E/I and J/P datasets. For personality prediction, we found that the NB learned

from the community activity feature (CA_Freq_100) obtained the best performance. Gender and age prediction could be achieved by using MLP learned from the hybrid feature set.

Table 10. The best classifier of 10 compared feature sets for 11 private attribute predictions. Bold indicates the best performance on each attribute.

	Feature Set	Gen.	Age	Edu.	Res.	Pol.	Rel.	Per.	E/I	S/N	T/F	J/P
Baseline	LIWC	MLP	RF	NB	RF	NB	NB	NB	NB	SVM	NB	NB
	LIWC_Tfidf	MLP	RF	MLP	RF	RF	RF	RF	RF	MLP	RF	RF
	BoW_Ngrams	MLP	MLP	MLP	MLP	MLP	MLP	SVM	MLP	MLP	SVM	MLP
	BoW_Stemmed	MLP	MLP	MLP	MLP	MLP	MLP	SVM	MLP	MLP	SVM	SVM
Proposed	CA_Freq	RF	NB	RF	RF	RF	SVM	RF	RF	RF	RF	RF
	CA_Freq_100	MLP	NB	RF	**RF**	RF	NB	**NB**	**MLP**	**RF**	RF	**MLP**
	CA_Wgt_100	MLP	RF	**RF**	RF	**RF**	**RF**	RF	RF	RF	**RF**	RF
	HF	**MLP**	**MLP**	SVM	SVM	SVM	MLP	SVM	SVM	SVM	SVM	SVM
	HF_LIWC	MLP	MLP	SVM	SVM	SVM	MLP	SVM	SVM	SVM	SVM	SVM
	HF_10k	**MLP**	**MLP**	SVM	SVM	SVM	MLP	SVM	MLP	SVM	SVM	RF

3.2. Training Time

Table 11 shows the training time of the best algorithms (shown in Table 10) in seconds using different feature sets for attribute prediction. We found that our proposed feature sets required a shorter training time compared to the baseline feature sets ($p < 0.001$). For community activity features, CA_Freq_100 used the shortest training time followed closely by CA_Wgt_100 because of its small size. The hybrid feature sets had a longer training time due to their complex extraction processes. However, the stemmed version of the comment text feature set (BoW_Stemmed) had a significantly higher training time than the non-stemmed counterpart (BoW_Ngrams).

Table 11. The training time (in seconds) of the ten feature sets for attribute prediction. Bold indicates the shortest training time.

	Feature Set	Gen.	Age	Edu.	Res.	Pol.	Rel.	Per.	E/I	S/N	T/F	J/P
Baseline	LIWC	539	83	527	73	47	138	121	101	105	112	115
	LIWC_Tfidf	561	86	529	75	47	137	120	104	108	117	119
	BoW_Ngrams	667	120	788	115	61	158	150	121	130	153	135
	BoW_Stemmed	2426	366	2461	345	239	667	510	462	479	519	495
Proposed	CA_Freq	228	61	388	70	16	43	84	92	63	73	69
	CA_Freq_100	**49**	**8**	**39**	**8**	**2**	**5**	**10**	**9**	**9**	**10**	**10**
	CA_Wgt_100	103	17	83	15	3	10	22	18	18	20	21
	HF	1004	154	879	147	83	223	213	170	175	206	194
	HF_LIWC	1788	280	1618	249	146	421	322	321	333	371	362
	HF_10k	996	140	834	128	72	208	153	156	165	185	180

3.3. Robustness

We evaluated the robustness of all algorithms learned from different feature sets by measuring the difference between the training and testing performance. The overfitting rate was calculated by the following equation. Note that the lower overfitting rate is more desirable.

$$\text{Overfitting rate} = \text{F}_{1,\text{Train}} - \text{F}_{1,\text{Test}}$$

From Table 12, we found that most of the baseline feature sets (LIWC and BoW) were over-fitted. Our proposed feature sets had a very low overfitting rate. This means that our proposed feature sets are desirable for learning algorithms. For the community activity, we found that the CA_Freq_100 feature set was the most fitted. We also found that the hybrid feature sets fit better than the LIWC and n-gram feature sets. Unsurprisingly, the HF_10k feature set fit better than the regular one (HF).

Table 12. Overfitting rate of learning algorithm using different feature sets on private attribute prediction. Bold indicates the most fitted. Lower values indicate better performance.

	Feature Set	Gen.	Age	Edu.	Res.	Pol.	Rel.	Per.	E/I	S/N	T/F	J/P
Baseline	LIWC	0.245	0.638	0.728	0.454	0.454	0.776	0.938	0.554	0.532	0.434	0.490
	LIWC_Tfidf	0.236	0.638	0.696	0.447	0.408	0.769	0.935	0.563	0.532	0.422	0.503
	BoW_Ngrams	0.164	0.615	0.493	0.306	0.388	0.705	0.800	0.475	0.489	0.388	0.412
	BoW_Stemmed	0.164	0.616	0.490	0.311	0.376	0.723	0.808	0.465	0.482	0.388	0.414
Proposed	CA_Freq	0.136	0.504	0.171	0.141	0.131	0.547	0.552	0.285	0.293	0.256	0.262
	CA_Freq_100	**0.080**	**0.179**	**0.048**	**0.032**	**0.060**	0.302	**0.218**	**0.071**	**0.077**	**0.086**	**0.063**
	CA_Wgt_100	0.107	0.459	0.050	0.038	0.070	**0.276**	0.403	0.122	0.149	0.116	0.101
	HF	0.127	0.585	0.287	0.174	0.286	0.663	0.667	0.382	0.398	0.294	0.255
	HF_LIWC	0.133	0.586	0.301	0.195	0.269	0.688	0.706	0.404	0.362	0.325	0.254
	HF_10k	0.118	0.574	0.239	0.131	0.252	0.659	0.584	0.307	0.320	0.268	0.232

3.4. Imbalance Problem

From the information of our datasets shown in Table 13, we found that class imbalance occurred in all private attribute datasets. Therefore, two oversampling techniques were explored to see their potential on our proposed feature sets. Table 14 shows F_1 scores obtained from random oversampling (RO) and SMOTE compared to the performance obtained from the original datasets (Table 8), denoted as the "None" technique (which means no oversampling method was deployed on that dataset).

Table 13. The number of classes, instances, and the imbalance ratio of the datasets.

Private Attribute Dataset	Classes	Instances	Imbalance Ratio
Gender Identity	2	17,589	1.00
Age Group	4	4136	33.17
Education Level	3	3499	2.68
Residential Area	8	17,446	171.28
Political Affiliation	2	810	1.42
Religious Belief	6	2709	108.12
Personality Type	16	4723	92.00
Introversion/Extraversion	2	4723	3.06
Sensing/Intuition	2	4723	7.16
Thinking/Feeling	2	4723	2.02
Judging/Perception	2	4723	1.50

Table 14. The F_1 scores of oversampling techniques for private attribute prediction. Bold indicates the highest performance of each attribute.

Feature Set	Technique	Gen.	Age	Edu.	Res.	Pol.	Rel.	Per.	E/I	S/N	T/F	J/P
CA_Freq	None	0.892	0.508	0.868	0.956	0.896	0.490	0.545	0.810	0.768	0.835	0.863
	RO	0.892	0.520	0.903	0.969	0.917	0.454	0.521	0.809	0.773	0.848	0.865
	SMOTE	0.891	0.527	0.896	0.965	0.905	0.481	0.530	0.792	0.751	0.835	0.857
CA_Wgt_100	None	0.880	0.498	0.947	0.979	0.915	0.606	**0.562**	**0.868**	**0.836**	**0.871**	**0.878**
	RO	0.868	0.513	0.955	**0.984**	0.917	0.598	0.560	0.861	0.815	0.871	0.871
	SMOTE	0.867	0.508	**0.959**	0.983	**0.919**	0.561	0.518	0.851	0.786	0.867	0.867
HF	None	**0.921**	0.520	0.854	0.907	0.877	0.531	0.511	0.760	0.721	0.808	0.801
	RO	0.918	**0.563**	0.907	0.916	0.880	**0.691**	0.558	0.816	0.775	0.825	0.815
	SMOTE	0.918	0.551	0.906	0.919	0.884	0.685	0.542	0.824	0.778	0.829	0.815

Experimental results shown in Table 14 revealed that using our proposed feature sets (CA_Wgt_100) without oversampling techniques reached the highest performance for personality prediction tasks (Per., E/I, S/N, T/F, J/P). Note that the personality type

prediction task (Per.) was the most difficult problem since it contained sixteen classes that came from the combination of [E/I][S/N][T/F][J/P] (see Table 7 for details). For the CA_Freq and CA_Wgt_100 feature set, we found that RO and SMOTE had a small contribution to the F_1 score for education and political belief prediction. RO method improved classification performance on age group, residential area, and religious belief prediction.

4. Discussion

4.1. Demographic Attributes

We have shown the predictive analysis of our work in the previous section. However, we also wanted to discuss descriptive results to better understand user behavior. We did this by looking for informative word features with high F-test values in each dataset. For gender identity prediction, we found effective word features related to relationships such as "SO" (significant other), "boyfriend", and "my husband". We also found that some lifestyle and news communities, such as "gaming", "technology", and "worldnews", can be used to imply the gender of the user who interacts with them.

We discovered communities related to lifestyle activities, such as "beetle", "Curling", "bicycleculture" that could be used as a data source for age group prediction. For the residential area dataset, we could predict the residential area with a high F_1 score of 98.1%. For informative words, we found words corresponding to their languages, for instance, "el" (Spanish for "the") or "de" (Spanish for "of"). For education level prediction, we found words explicitly related to the topic, such as "PhD", "grad", "student", and "college". We also discovered communities directly related to education other than the ones we extracted from, such as "AskAcademia", "csMajors", "gradadmissions", and "CollegeRant".

For the political affiliation dataset, we discovered that the most informative words were related to accusations, such as "FBI" or "witnesses" since we obtained the experimental data during the nomination of US Supreme Court Justice Brett Kavanaugh. New communities related to controversial discussions were discovered, such as "AskScience" and "debatereligious". This implies that those users like to express their world views on controversial issues. For the religious belief dataset, we found words corresponding to religious teachings, for example, "Quran" (the text of Islam) or "Allah" (the god of Islam).

4.2. Personality Types

Our proposed feature set, CA_Freq_100 with NB, significantly outperformed the research work done by Gjurković and Šnajder [12] on personality prediction at 64.4% (with over 22.7% improvement). We also performed a feature analysis and found words mentioning their personality types and MBTI-related communities as the most effective features.

One interesting question is "Can personality type be inferred from the demographic attributes?" We answered this question by setting up the experiment to see the predictability power of demographic feature sets. First, we derived a new dataset from the personality data set consisting of 4723 users by integrating their six demographic attributes. Then, models obtained from each demographic dataset were deployed to predict the missing demographic value found in the new dataset. After that, logistic regression was trained by the set of six demographic attributes to predict the personality types. We found that the macro average F_1 score of the model was very low and close to that of MCC, with a 2.3% difference. As shown in Table 15, we found that using logistic regression learned from six demographic attributes obtained worse performance compared to the baseline feature set (LIWC). Our experimental results implied that people's personality types were independent of their demographic attributes.

Table 15. The F_1 scores of personality type prediction using different feature sets.

Feature Set	F_1 Score
MCC	0.025
LIWC_Tfidf + RF	0.064
Demographic attributes + LR	0.048
CA_Freq_100 + NB	0.644

5. Conclusions

We have done an empirical analysis of our proposed feature sets for private attribute prediction covering classification performance, training time, and imbalance problems. From experimental results, we can conclude that user engagement on Reddit shows promising results for the discovery tasks. Although much research has been done on large platforms, such as Facebook and Twitter, we have shown that Reddit is a potential source of demographic and personality study as well. Our results show that we can predict MBTI personality type with an F_1 score of 64.4% with a dataset of 4723 users. Our proposed feature sets applied with machine learning algorithms provided an impressive performance. We obtained 98.1% for residential area, 94.7% for education level, 92.1% for gender identity, 91.5% for political affiliation, 60.6% for religious belief, and 52.0% for age group.

For future work, we plan to explore ways of extracting other demographic attributes using the same technique. For the proposed feature sets, feature transformation and decomposition can be performed to study the change in performance. Imbalance problems can also be further investigated for textual features, which are known to be more difficult to handle than numeric ones.

Author Contributions: Both authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Faculty of Science and Department of Computer Science, Kasetsart University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://files.pushshift.io/reddit/comments/ (accessed on 23 August 2021).

Acknowledgments: We thank the Office of Computer Services, Kasetsart University for facilitating a high-performance computing platform beneficial to our experiments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Smedt, T.D.; Pauw, G.D.; Ostaeyen, P.V. Automatic Detection of Online Jihadist Hate Speech. *arXiv* **2018**, arXiv:1803.04596.
2. Zhao, W.X.; Li, S.; He, Y.; Wang, L.; Wen, J.-R.; Li, X. Exploring Demographic Information in Social Media for Product Recommendation. *Knowl. Inf. Syst.* **2016**, *49*, 61–89. [CrossRef]
3. Neal, A.; Yeo, G.; Koy, A.; Xiao, T. Predicting the Form and Direction of Work Role Performance from the Big 5 Model of Personality Traits. *J. Organ. Behav.* **2012**, *33*, 175–192. [CrossRef]
4. Matz, S.C.; Kosinski, M.; Nave, G.; Stillwell, D.J. Psychological Targeting as an Effective Approach to Digital Mass Persuasion. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 12714–12719. [CrossRef] [PubMed]
5. Myers, I.B. *Gifts Differing: Understanding Personality Type*; CPP Books: Palo Alto, CA, USA, 1993; ISBN 978-0-89106-064-2.
6. Barbuto, J.E. A Critique of the Myers-Briggs Type Indicator and Its Operationalization of Carl Jung's Psychological Types. *Psychol. Rep.* **1997**, *80*, 611–625. [CrossRef]
7. McCrae, R.R.; Costa, P.T. Reinterpreting the Myers-Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality. *J. Pers.* **1989**, *57*, 17–40. [CrossRef]
8. Furnham, A. The Big Five versus the Big Four: The Relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI Five Factor Model of Personality. *Personal. Individ. Differ.* **1996**, *21*, 303–307. [CrossRef]

9. Kosinski, M.; Stillwell, D.; Graepel, T. Private Traits and Attributes Are Predictable from Digital Records of Human Behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [CrossRef]
10. Aletras, N.; Chamberlain, B.P. Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In Proceedings of the 29th on Hypertext and Social Media, Baltimore, MD, USA, 9–12 July 2018; ACM: New York, NY, USA, 2018; pp. 20–24.
11. Ferwerda, B.; Tkalcic, M. Predicting Users' Personality from Instagram Pictures: Using Visual and/or Content Features? In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, 8–11 July 2018; ACM: New York, NY, USA, 2018; pp. 157–161.
12. Gjurković, M.; Šnajder, J. Reddit: A Gold Mine for Personality Prediction. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 87–97.
13. Tausczik, Y.R.; Pennebaker, J.W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **2010**, *29*, 24–54. [CrossRef]
14. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
15. Ho, T.K. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
16. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]