Mining Personal Traits in Social Media

Rounak Saha

Department Of Informatics, King's College London

Abstract

The inception of Social Media has led human beings to pen their day to day thoughts and activities on the internet. This large chunk of readily available data is rife with latent patterns which mathematical methods can take advantage of to predict personality traits of individuals. In this paper, we review data mining methods and their suitability for different user attributes. This paper will conclude by discussing sociological concerns with respect to predicting human personality traits computationally.

1 Introduction

Social Media in an important part of our daily lives in today's digital world. A concept to connect friends and families around the globe has evolved to become an integral part of our daily lives. Today, we rely on Social Media for a range of activities starting from generating and maintaining professional relationships to basic commodities of life such as News. Penning down our thoughts and daily activities on Social Media leads us to provide a lot of information about ourselves which can be used to infer personality characteristics we do not provide willingly.

The scientific domain of study which uses certain aspects of our representation on Social Media to predict personality traits which are otherwise not explicitly stated is known as data mining. Data mining is an interdisciplinary field between Mathematics and Computer Science. Mathematical Methods developed in this domain helps us to deduce missing parameters based on models of large datasets. Application of Data Mining Methods(**DMT**) on Social Networks will help us to model large datasets of user profiles. This in turn will be used to obtain personality characteristic of users which are not provided voluntarily.

Data Mining Methods(**DMT**) have a vast range of applications. The most popular applications are in the domain of targeted advertisements and government surveillance. However, the list of applications does not end there. In recent years, **DMT** has been used in conjunction with Social Media to track outbreaks of diseases such as Influenza [18]. Natural disasters have been tracked in the same fashion as well [47]. The usage of Social Media applications in political campaigns and brand building has been much spoken about of late.

This paper attempts to analyse three data mining methods in the context of predicting personality attributes on Social Media, namely, a) Bayesian Methods, b) Support Vector Machines and c) Label Regularisation. We argue that these methods are best suited to predict personality traits such as Gender, Ethinicity, Location of a User and Political Affiliation. We make observations in terms of limitations of these methods and identify techniques which can yield better results using these methods.

We first look at the evolution of Social Networks in Section 2. This section will highlight the ever changing landscape of Social Media. We will discuss how the evolution of Social Media platforms is linked with the evolution of Mobile Computing Technology. In Section 3, we establish the need to predict personality traits of individuals. We then discuss how computational predictions are more efficient and cost effective than using human labour. In Section 4, we discuss how data mining and personality traits are related to each other. In Section 5 we describe the obstacles we face in the process of Data Mining.

The use of mathematical methods to infer personality traits has long existed in the history of human race. In fact, this is very similar to methods employed by detectives and police officials to solve crimes. In Section 6, we take a look at how data mining methods were applied to content before the advent of Social Media. In Section 7.3, we discuss the three Mining Methods which are popularly used to deduce personality traits on Social Media.

Mining social media and the use of personal details of users has often raised sociological concerns. Data analysis has been critiqued to ignore multiple communities and ethical questions have been raised in the usage of data. In Section 8, we discuss sociological aspects of data mining and how we can ensure ethical standards. This paper concludes in Section 9, by discussing future directions of work in the domain of data mining social networks.

2 Evolution of Social Media

The year 1997, saw the advent of Web Logs. Perhaps, this was the first form of social media that we know of today. Web Logs commonly known as Blogs, provided users a platform to express themselves without much regulation, free of cost. Similarly, users could read Blogs written by other users. This led to an open environment where human beings could communicate with each other irrespective of time or location with the aid of just an internet connection. Blogs are a relevant form of social media till date. It has evolved a great extent from its initial format. Today, Blogs are a source of information in the domain of tourism [42] and food recipes [43] to name a few. One can argue about the authenticity of the information generated in these independantly run blogs. However, the fact that users interact with these blogs by using features such as comments and social shares is a reality.

The advent of blogs were followed with platforms such as MySpace and Orkut where users could now interact with one other directly and not just via content generated by one another. This was a step forward in connecting human beings across the globe. On Orkut, users could write short messages known as “Scraps” to each other. This led to the formation of networks of users interacting with each other. Such networks which existed digitally came to be known as Social Networks.

Frigyes Karinthy in [24] states that two entities in the physical world are connected with each other by a maximum degree of six. In other words, any Person X in the physical world can be connected with any other Person Y with a maximum of six people connecting them. This theory is more commonly known as “The Six Degrees of Separation”. In 2004, Facebook was founded and shortly after in 2006, Twitter came to existence. These are the two major social media platforms in today’s day and age. The tagline of Facebook as mentioned on their login page is, “Facebook helps you connect and share with the people in your life”. In 2016, Facebook revealed that the degree of separation between any two users on their network has shurnk to only 3.5 [7]. This prooves that Facebook does bring users on its network closer to each other.

It has been much argued that the social media platforms we think of as products and use extensively as a part of our daily lives are actually not the product. In [40], Rushkoff argues that the product of social media platforms are actually the users themselves. Social media platforms provide a range of features and a skeleton whose flesh is the content generated by users. The valuation of corporations such as Facebook and Twitter is due to the vast range of data they own generated by the users on their platform. The content we generate in forms of textual messages as well as multimedia content provide a rich source of information that can easily deduce multiple personality traits of users. In addition, features such as Geo Tagging, Hastags and other forms of multimedia make this task easier.

The first known form of social media was only available to a priviledged section of the society. This was so because internet connection was not an easily availble commodity and also computers were fairly expensive. Social media sites such as Orkut and MySpace could be acessed only through web browsers. Today, one does not even need a computer to access social media platforms. A low range smart phone with an internet connection can provide access to these platforms. In addition, the advent of mobile applications have made it easier for us to keep our social media presence in tips of our fingers. This was only possible due the proliferation of mobile computing in recent times. The growth of Facebook as a social media platform can be observed in Figure 1. The number of users on various social media platforms as of June 2017 can be observed in Figure 2.

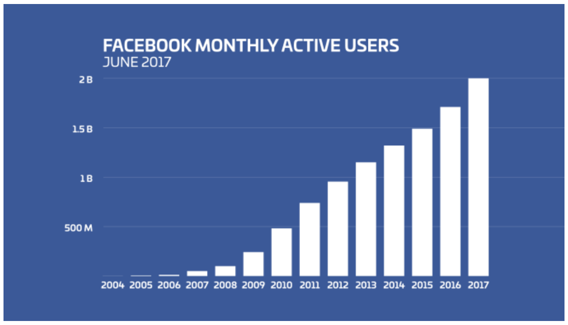

Figure 1: Growth of Facebook Users over the years. [14]

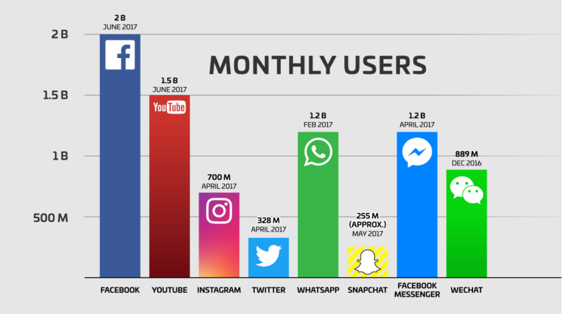

Figure 2: Number of Users of Various Social Media Platforms as of June 2017. [14]

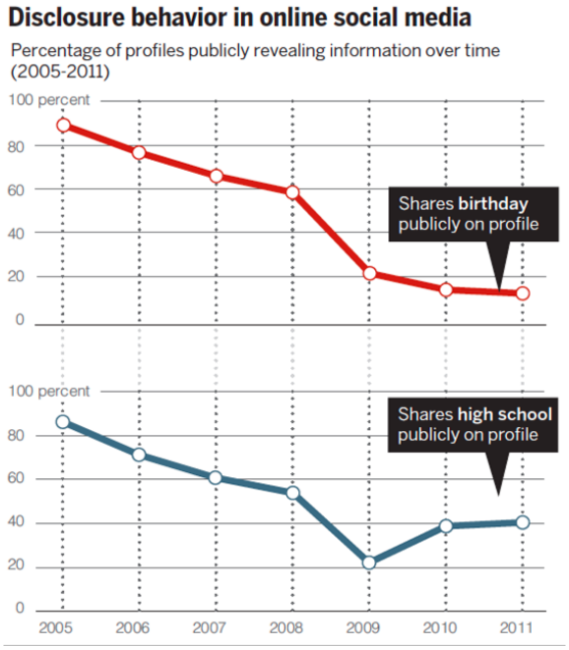

Figure 3: Decline in the number of users making information publicly available. [1]

This vast amount of data penned by millions of users from the inception of Blogs in 1997 till date contains clues and logical expressions that data mining methods make use of to predict personality traits from social networks. However, one many argue that if such an amount of data is present on the internet which can be easily acessed, what is the need of emplyoing complex mathematical methods? The reason to do so is two fold. First, not all data is made publicly availble by users on these paltforms. Hence, we may not have acess to valuable data such as gender and age which are very important in terms of targeted marketing. Figure 3 shows that users have stopped revelaing information about themselves publicly over the years on social media. In addition, personal opinions such as political affiliations is not present as data on social media platforms but derived from content generated by users. Second, even if the data was readily availble, it maybe erroneous in nature. Human beings would have trustred erroneous data which algorithms can bypass. Algorithms also work at a much greater speed than human beings.

3 Computational Prediction of Personality Traits

The need to predict personalities traits of human beings has been essential for years. A quintessential example would be the process of hiring a new employee for an organisation. Employers would want to know personality characteristics of a new hire to ensure that the candidate fits the role. In the absence of scientific methods, orgsanitions would employ third parties to learn more about a new employee. Over the years, this process has evolved. Aptitude and physcometric tests often reveal a lot of personality traits about the examinee. Such tests are designed in such a way that they judge an individuals ability to fit the role on a number of factors. Our approach in the field of data mining personality traits in social media is quite similar. We look for features in content generated by users and try to classify the user according to certain criterion.

Another example is in the field of targeted marketing. A local business X may want to promote their products to individuals of a particular gender, ethinicity and age. In addition, the business would only want to promote their product in a certain radius of their location. In such a scenario, social media can lend us a set of users. The content generated from this set of users can be used to predict personality traits of these users. Once we infer the personality traits, we can show the advertisment to only those users who meet the constraints. In the absence of social media, the availbility of this set would not have been possible. Even in a scenario when we get this set of data, in the absence of computational methods, predicting these personality traits would be time consuming and the cost to do so would be so huge that business X would be able to afford it.

In [48], the authors compared personality judgements made by data mining algorithms with human efforts. The authors find that algorithms to predict personality traits of users are extremely fast. In addition, they note that the accuracy to judge personality traits of users is much higher in the case of computational methods. In fact, the authors make an observation that human beings sometimes make incorrect decision about their own personalities. Two reasons are provided by the authors for the lack of accuracy in human prediction of personality traits. First, human beings do not have a vast amount of memory to store information that is required to make predictions. Second, human beings are often biased which hinders them from making rational judgements.

Data mining methods can also be applied to infer political sentiments from social media. Traditional methods applied to infer political sentiments were in the form of surverys. In [33], the authors use publicly availble data on Twitter to detect political sentiments. Analysis of such sentiments is a very difficult task in the absense of computational methods. In addition, the results may not be accurate as surverys maybe biased or even rigged.

4 Data Mining and Personality Traits

In this Section, we will first discuss a generalised approach to data mining. The various types of data mining techniques and its corresponding approaches will be presented in detail. We will then discuss how personality traits of users can be deduced using data mining methods by taking advantage of distinguishing patterns in user generated content. All definitions and terminolgies in this Section are based on [46] and [19].

A. Data Mining

Data mining methods can be broadly classified into three categories. In this paper, we will mostly concentrate on supervised and semi-supervised methods of data mining.

1. **Supervised Approach** : In this method, an intial set of data is used to train the classifer. All data points or exemplars in this set is annotated with a label to define its state in the entire corpus. For example, if a set of data with X number of users is used to train a classifier to predict gender, every data point in this set should be labelled as Male or Female.

2. **Semi-Supervised Approach** : In this approach, a certain ratio of data points are labelled while the rest of not labelled. Using the already labelled data, algorithms try to learn the labels of the rest of the data points.

3. **Unsupervised Approach** : In this approach, no data points are labelled. It completely depends on the algorithm to learn the labels and make classifications. The approach to do so is to form groups or clusters which are able to make clear distinctions.

The first stage of data mining in a supervised or semi supervised method to make any form of predictions or classifications is known as Training. In this step, our first goal is to collect a fairly large corpus of data. This corpus should be well represented to ensure the training mechanism includes all possible combinations. For example, we are trying to predict if users are single or in a relationship, this corpus should have equal amounts of users representing both categories. The next stage is to extract distinguishing features from this set of data. For example, users in a relationship would have different interests than users who are single. These features are fed as input to our data mining methods. Such methods are responsible to learn these distinguishing features.

Once our classifier has been trained, it can be subjected to new data. Upon extraction of features, the classifier should now be able to make accurate predictions. A generic description of this mechanism is provided in Figure 4.

B. Data Mining Personality Traits

In 1993, Lewis Goldberg formulated the Five Factor Model [23]. This model highlights the various factors that contribute to a human beings personality traits. The factors in this model are described as follows :

1. **Openness** : The extent to which an individual is seeking out for new forms of experiences.

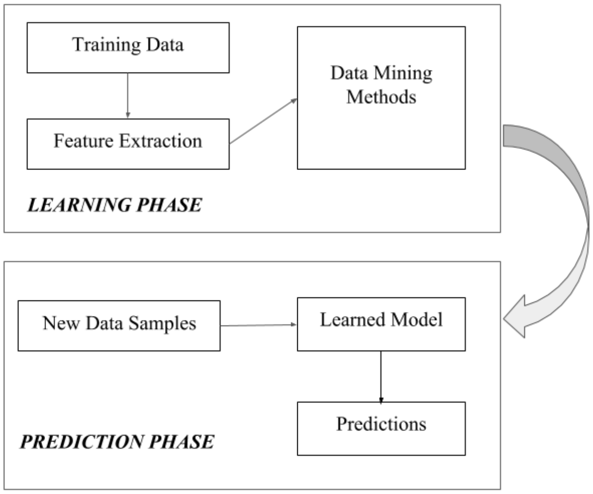

Figure 4: Generic Block Diagram of Supervised or Semi-Supervised Data Mining Approach.

2. **Conscientiousness** : The degree of discipline in one's life.

3. **Extraversion** : The extent to which an individual can acclimatise in external enviornments in the presence of other individuals.

4. **Agreeableness** : Co-operative nature of an individual.

5. **Neuroticism** : The degree to which an individual is susceptible to negative feelings.

In the purview of Social Media, we can extract socio-linguistic features from user generated content and match it against patterns which act as benchmarks of personality traits as described in the Five Factor Model. This in turn, would help us to predict personality traits succsefully with the help of data mining methods. It also enables us to predict traits in large scales as described in [27]. To this extent, a Facebook application known as myPersonality was setup in 2007. The application would predict a users personality traits and collect the users personal data. This large chunck of data that was collected was used to improve predictions of such algorithms [45]. This shows that algorithms can in fact extract personality features which are otherwise not very evident and make conclusions based on them. The Psychomertrics Center at University of Cambridge has setup an application which allows users to connect their Facebook and Twitter profiles and receive detailed personality reports about themselves. [1]

5 Challenges in Mining Social Networks

There are multiple challenges in the domain of data mining social networks. In this paper, we will talk about three major challenges. First is the challenge of acess control. In order to formulate our training set, we would need to acess social media platforms to gather data. Various platforms put a set of limitations to acess their databases. The most common way to acess data is with the help of an Application Programming Interfaces(APIs). The most common APIs to acess Facebook and Twitter are OpenGraph[2] and Tweepy[3]. However, both of these APIs are rate limited. This means that applications would need to authenticate themselves everytime to acess data. In addition to this, after a certain number of requests in a time frame, applications will not be provided any data. This hinders developers to formulate a broad corpus in a feasible amount of time.

The enviornment of Twitter is much more open than that of Facebook. This is an observation that researchers use to their benefit to scrape data from Twitter. Even though APIs are rate limited, researchers take advantge of Twitter Advanced Search[4] and headless browsers like Selenium[5] to scrape large amounts of data in a feasible amount of time. Facebook on the other hand, requires various forms of authentication to acess any form of search queries. This is a major reason why most research efforts have Twitter data as their main corpus. Platforms like WhatsApp and Tinder contain rich information which has potential to yield better prediction results if research communities have acess to the data on these platforms.

The second issue that we observe is noise. Noise can be in the form of incorrect data provided by users as an error. It can also be erroneous data provided on purpose or even spam. If our learning model is trained on noisy datasets, then our prediction would be erroneous as well. Efforts should be made to preprocess datasets to get rid of noise. Noisy datasets also include deceptive content.

Third, is the issue of bias. We must note that our data mining models are probablistic in nature. If our training corpus is biased to favor a certain section of users, our predictions would be biased as well. Care must be taken to ensure there is no bias in the dataset by eliminating and balancing datapoints.

While, these are some major challenges in the field of data mining, the list is not limited to this. Due to the dynamic and versatile nature of this field, prediction of each personality trait has its own set of challenges.

6 Short History of Predicting Personality Traits

Human beings have always been keen on predicting personality traits of individuals based on a set of observations. Sherlock Holmes in the works of Arthur Conan Doyle's novels is famously known for the science of deduction. Holmes used to observe, theorise and make predictions based on these observations [25]. This rule of three is similar to our approach in data mining social networks. The most distinctive difference is in these two approaches is the fact that Holmes or any other individual is already equipped with a prior knowledge which they can tap into. Data mining methods need to learn from a set of data to accurately make predictions.

Mathematical methods to make personality trait predictions have also been used in the past. In [5], authors try to predict personality characteristics such as gender, age and native language from textual content of unknown authors using machine learning. This is

[1] https://applymagicsauce.com/demo.html

[2] https://developers.facebook.com/docs/graph-api

[3] https://developer.twitter.com/

[4] https://twitter.com/search-advanced?lang=en-gb

[5] https://www.seleniumhq.org/

defined as the authorship profiling problem where we try to infer the identity of an individual based on a piece of text written by them. Such methods are used extensively in the field of foreincis to reveal latent information.

Telephone conversations is another domain where latent information can be detected. In [22], authors use features such as phonetics, accents and meta data from telephone numbers to predict age, gender, native language and even location of the call. In [8], the authors make use of Gaussian Mixture Models(GMM) and Support Vector Machines(SVM) to determine age of children in primary school based on a reading test.

Social Media, provides us with advantages such as network structures, connectivity between users, timestamps and even geo located data to make the task of predicting user personalities easier. Even though we have so many advantages, it is sometimes diffcult to acess all these features. In these circumstances, we just have textual content generated by users to make predictions. This makes our problem definition very similar to the likes of the authorship profiling problem only in a digital context.

7 Mining Methods to Predict Personality Traits

In this Section we discuss data mining methods that are used to predict personality traits on social media. In section 7.1 we will discuss Bayesian Methods and how this method can be used to infer gender and ethinicity of users. In section 7.2 we will discuss Support Vector Machines and predict gender, location and political affilaition of users. We will conclude this Section in section 7.3 where we will discuss Label Regularisation and how this method can be applied to determine political affiliation of users.

7.1 Bayesian Methods

The Naive Bayesian Method evolves from the Baye's Theorem. Baye's Theorem can be stated as

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1}$$

X and Y can be defined as two events. The probability of event X occurring, given event Y, is defined by equation 1.

In simpler terms,

$$PosteriorProbability = \frac{Likelihood * Prior}{Evidence} \tag{2}$$

It must be noted that all equations, terminologies and definitions in this section are based on [19].

In equation 1, we can observe that the Baye's Theorem is dependant on two factors. First, the prior probablity $P(X)$ and second, the conditional problity $P(Y|X)$. This hints that we need to have great prior knowledge about the personality traits that we are trying to predict. Since this is a supervised method, all data points in the dataset must have a class label attached to it. We will use the Bayesian method to predict gender and ethinicity of users.

A. Gender

Prediction of gender has always been defined as a “binary classification problem”. The Bayesian method tries to learn feature characteristics and form a singluar decision boundary to solve this two class classification problem. In order to algorithmically compute the decision boundary, we will first need to identify the features which can clearly disntinguish content between male and female authors.

The blog ecosystem has been studied in detail in terms of the gender classification problem. In [41], the authors note that there are broadly two features that distinguish between male and female generetated content. First, the style in which text is written by male and female users seem to differ. The authors noted that men tend to be more informative while women tend to be more articulate about themselves in terms of writing style. These features which relate to the style of writing are termed as Stylistic Features. Second, the authors noted that topical content between male and female users are distinguishable. The authors state that men are more likely to be authors of textual content related to topics such as "linux, microsoft and gaming". Women on the other hand are likely to be authors of topics such as "shopping, mom, cried and pink".

These observations were further improved in [32]. The authors made note of two additional features, F-Measure and Factor Analysis and Word Classes. Using these four features, the authors proposed an algorithm known as Ensemble Feature Selection (EFS). The goal of the EFS algorithm was to selected the best feature vectors out of these four which would distinguish the two classes clearly. Once the feature vectors were picked by EFS, Bayesian methods could accurately predict gender with an accuracy of 70%.

Upto now, we have only discussed prediction of gender using the Bayesian method in the blog ecosystem. The distinguishing features that we have discovered in the blog ecosystem can be extended to modern day social networks as well.

In [36], the authors propose a hierarchial Bayesian model to predict gender in a semi-supervised fashion. The authors used comments on Facebook and the names of the users who made these comments. The authors annotated the names of users with genders by crawling public census data. The authors used the same linguistic observations made in the blog ecosystem to extract features in the form of n-grams from the comments. This provided the authors with two models, a) Bayesian model using names and b) Bayesian model using textual content(comments) generated by users. Using a combination of these two models, a high accuracy of 79.8% was achieved.

One of the crucial disadvantages in the Bayesian method is the problem of missing prior information. In a scenario where the model has no prior information, data points tend to be misclassified leading to low accuracies. The usage of only textual features can lead to low accuracies. The hierarchial Bayesian model provided in [36] can overcome this disadvantage. Even if there is missing information in terms of n-grams or names from census data, the other feature can provide a strong backup and yield high accuracies.

In [9], the authors predict gender in the Twitter ecosystem. The features they take into account are Screen Name, Full Name, Biographic Description and the content of the Tweet. n-grams are extracted for all the features and an accuracy of 67% is achieved. Even though a substanially large dataset was used in this study, missing priors can be crucial in Bayesian methods to report high accuracies. The hierarchial Bayesian method proposed in [36] not only provides a high accuracy but also uses a substantially small dataset. However, one of the major drawbacks of using census data is that acess is not always easily obtained.

Bayesian methods is not the most accurate of methods to predict gender for users. Various other methods such as Support Vector Machines (SVM), Balanced Winnow2, k-Nearest Neighbours(kNN) provide higher accuracies. However, such methods are not as straightforward to implement such as the Bayesian method and come with its own set of constraints. In SVMs, data needs to be linearly separable. In kNN, the value of k is very crucial and can have huge impacts on accuracy. In [35], the authors show that Bayesian methods outperform SVM and kNN as data could not be linearly separated and an optimal value of k could not be determined. Bayesian methods provided an accuracy of 63%, outperforming SVM at 61% and kNN at 60%.

B. Ethinicity

Unlike gender, prediction of ethinicity is very accurate using Bayesian mehods. In [10], the authors state that a generalised data set which covers the entrire spectrum of ethinicities as that in the physical world is not feasible to obtain. This hinders discriminative probabilistic methods from making accurate predictions. Generative probabilistic, like the Bayesian method, computes a conditional probablity before making the final prediction. This step provides with the chance to annotate the data set with external data sources such as census data or data from gazettes to yield higher accuracies.

thinicity is a very important attribute as we can derive a lot of other attributes from the ethinicity of a person such as gender, location and even political affiliation. In addition, the analysis of ethinicities in various sectors provides us with a picture of inclusivity.

In [36], the authors use the same hierarchial Bayesian model along with census data to predict ethnicities solely based on Nigerian names and comments on Facebook. n-gram features were used over names and comments. Using a very small dataset, a very high accuracy of 81.1% was obtained for predicting ethinicities. The authors also proove the merit of the hierarchial bayesian model by showing that models with only names or only comments had much lower accuracies.

A similar approach of using census data is used in [10]. The authors take advantage of census data recorded in the United States of America (USA). There are some very interesting observations made in this paper. First, relationship between individuals are formed in an associative manner. The authors state that Person X with ethinicity Y is highly likely to be in a relationship with another person of the same ethinicity. Second, religious sentiments and political affiliations tend to be similar in ethinic communities. These observations can be used to predict other personality traits.

The prediction of ethinicity is a fairly simple using the Bayesian method along with census data. One of the major merits of this method is that it is weakly supervised. However, there are drawbacks as well. First, census data is not easily available. Second, census data maybe noisy or even outdated.

In Section 7.1, we have seen how effective the Bayesian method is to predict gender and ethinicity of users on social media. Feature extraction as well as setting up the model is a straightforward process. The method is succeptible to errors in the scenario of missing prior knowledge. However, when the data set is annoted with external sources of information, the Bayesian method yields high prediction accuracies.

7.2 Support Vector Machines

Support Vector Machines(SVMs) is a supervised data mining approach. SVMs try to create an optimal decision boundary between two linearly separable classes. The optimal decision boundary is is a hyperplane which clearly distinguishes the two classes. SVMs have a constraint that the datasets should be linearly separable. However, that is not always possible in real life scenarios. In order to tackle this problem, a kernel function is often used. Kernel functions project non-linearly separable data into a higher dimension where the data can be linearly separable. This is known as the kernel trick.

SVMs can only make distinctions in between two classes. In a real world setting however, there are multiple classes. In order to deploy a multi-class problem, SVMs using additional algorithmic procedures such as one against one, one against all and decision trees to name a few. SVMs can be easily implemented as they are availble in ready made toolkits such as WEKA[6] and SVM^{light}[7]. It must noted that all terminologies and definations in this section is based on [21], [16] and [15].

In this paper, we will predict gender, location and political affiliation of users on social media using SVMs.

A. Gender

One of the early works to predict user attributes on social networks is [37]. The authors use only textual content from Twitter and SVMs to predict gender. Fraternities and hygiene products were looked up and the corresponding user content was scraped. To sanitise the dataset, network structure such follower, following and follower-following ratios were analysed. A threshold was set for these ratios and the corresponding users who did not meet the constraints were eliminated from the dataset. This process got rid of inactive users, spam, celebrity and business accounts. This is a necessary step to exclude bias and noise. The authors employed human annotators to

[6] https://www.cs.waikato.ac.nz/ml/weka/

[7] http://svmlight.joachims.org/

assign class labels of Male or Female to each user. Three SVM models were designed. The first model was based only on stylistic features by splitting the tweets into tokens using TF-IDF. The second model was based on n-gram features extracted from textual content of users. The third and the most interesting model is a stacked SVM model. The output of the first two models were fed into the third SVM to make accurate predictions. One can draw parallels of this approach with the hiearchial bayesian model setup in [36] using Bayesian methods. The stacked SVM model provided an accuray of 72.33%.

In [9], the authors take forward the work presented in [37]. The authors eliminate the manual annotation step by crawling blogs of users which are listed in their Twitter description. The assumption is that users who list a blog on their Twitter description and maintain it are likely to not be spam. This step not only reduces cost of annotation but overcomes challenges of noise intrinsically while annotating. In addition to the tweet content, the authors have also used Twitter meta data as features. A high 71.8% accuracy was obtained which improves the benchmark in previous works using SVMs. The authors however note that it took about fifteen hours to train the SVM using a linear kernel.

As of now, all our analysis has been limited to English text. In [12], the authors propose to extend our findings in this domain to other languages. On analysing four languages, the authors conclude that while in some languages such as French a little fine tuning can yield high accurate results. In some languages such as Japanese, current findings fail due to the extremely different syntax of the languge. The authors also note that other languages yield features which have not been discovered in the English languge. One can argue that features discovered in other languages can be present latently in English. This argument rises from the fact that languages evolve from certain parent languges.

B. Location

The problem of predicting the location of an user can be defined in two ways. First, predicting the home location of an user. The user maybe currently living in any city or even traveling. The goal of the problem is to determine the home city of the user. Second, to predict the current location of an user.

In [37], the authors try to determine the home location of users based on the dialects of users in north and south India. The socliolinguistic model yieled an accuracy of 77.08%. One may argue that the dataset used is too well refined and represents extreme cases which made the classification problem trivial. If the same model is subjected to new data from urban areas, it may not fare well.

In this paper, we will look at classifying current location of an user in city level. This attribute is particularly beneficial as it has the potential to boost local business who can promote their products at very cheap costs on social media.

In [11], the authors identify words which serve as strong indicators of location from just tweet content. In addition, they apply a smoothing model which refines the location estimate. The idea is to identify words which can only be used in a very local context. In [6], the authors put forward a premise that the location of an user will be related to their social circles. Using this approach, the authors even outperformed IP Based algorithms by using just tweet content and network structures of users.

In [39], the authors defined the problem of predicting a users location as a classification problem. They used a SVM classifier trained on publicly available datasets. Network structures and places of interest were extracted as features from user content. 62.08% users were located with a high confidence within 50m of their location. This is an improvement from previous benchmark results which only placed 43.61% of users within 50m of their location. The shift of idea to formulate the problem as a classification problem can be said to be responsible for the increase in accuracy.

The premise of locating words in a local context and analysing the social network structure of user were extended in [28]. The authors augmented their dataset with location check-ins on Foursquare. The authors provide a three step appraoch to the problem. First, the dataset is filtered and tweets which do not reveal substaintial information about city level keywords are eliminated. Second, a probablistic value is attached to tweets whose locations are identified. This produces a list of confidence with which prediction can be made. In the third step, the authors develop a SVM classifer with the aid of tweets which are already geo-tagged. On passing the probablistic ranked list of locations developed in step 2 through the classifer, highly accurate predictions can be made. This approach is similar to [30] where the authors use a hierarchial filtering process on tweet content and hashtags.

SVMs are very efficient in differentiating between classes. Hence, training SVMs using already geo-located tweets and gazettes formulate very effective classifiers. In the future, location tags from digital maps(such as Google Maps) can be used to crawl their corresponding social media page to obtain rich information. The use of a hierarchial filtering process is responsible for placing users withing small radii of location spans with high confidence.

C. Political Affiliation

The political affiliation of an user is a personal opinion which is not explicitly stated on any social media platform. The opinion is also subject to change over time. Our efforts would be to predict the current political affiliation of users on social media. We can accurately predict this opinion because users tend to generate content aligned to their belief systems. The basis of political affiliation in turn is based on one's belief system.

In [20], the authors use blogs to predict political sentiments. The authors approach the problem by locating topics in blogs which could hint at the blog's political alignment. A similar approach is made in [37] on Twitter data. The authors scrape users and the content generated by them using topical keywords and hastags aligned to Democrats and Republicans in USA. n-grams extracted from user generated content was used to train SVMs. The SVM model trained on these n-grams yielded an accuracy of 82.84%.

One can argue that the dataset used in [37] is very generalised. Users affiliated to any political party might also speak about a range of other topics. Using only user tweets to train a classifier might make the distinction between topics fuzzy. In addition, one must note that this dataset was annotated manually and thereby may only span a very small range of topics.

In [13], the authors suggest a differnt approach. Instead of training the classifier by using only tweets of users, they train the classifer based data from hashtags. The authors discover political hashtags using a discovery algorithm from two seed hashtags. In this dataset, they identified users with the largest network reach. These users were manually annotated to be either democratic, liberal or ambiguous. The authors use TF-IDF to extract unigrams. For the sake of comparision, they build two SVMs. One SVM is trained using only content from a select number of users. In the second SVM they only use features extracted from hashtags. While the SVM trained on textual content yielded an accuracy of 79%, the one trained on hastag data yielded an accuracy of 90%.

The reason of this improvement can be from the fact that hastags contain very rich topical information which describe the corresponding class labels. Therefore, the features so obtained are very distinguishing from each other. Similar approaches and findings have been made in the field of biomedicine. Researchers trained classifiers based on a description of topics and not the full textual information itself [29].

SVMs provide with highly accurate results when predicting political affiliations with respect to a specific country(in this paper, USA). However, if we try to generalise political affiliation prediction for multiple countries, SVMs may have problem in scaling the scope. In such situations, unsupervised clustering algorithms maybe better suited. In recent works of analysing political enviornments, clustering algorithms have been used in conjuction with hashtags and sets of websites whose political affiliations are known [13]. Researchers have also used Gradient Boosted Decision Trees(GDBT) to predict political affiliations [34].

From this section we can conclude that the advantage of SVM lies in the fact that it can be easily implemented with the aid of readily available toolkits. The kernel trick can make data linearly separable very easily. In addition, SVMs are fairly generalised and can be extended to various geographies and languages with a certain amount of fine tuning.

7.3 Label Regularisation

Up to now, we have seen two data mining methods which has been championed over the years. In this section, we will look at a fairly recent method known as Label Regularisation. Using this method, we will try to predict political affiliation of users.

In [31], the authors propose a new approach for prediction models with unlabelled data known as expectation regularization(XR). A type of XR in which some data is labelled in the entire dataset while the majority is not came to be known as Label Regularisation. This is semi-supervised approach. One may draw parallels of this approach with the methods we have used earlier in this paper to predict location and political affiliation using census and gazette data. However, the approach seen earlier in the paper uses the data in the form of a dictionary. In case of label regularisation, our goal would be to deploy learning mechanisms so that the unlabelled data can learn from the labelled data and make predictions accurately.

The first step of label regularisation is to define a set of constraints over features of labelled data in the dataset. Having defined the constraints, we draw a probablity distribution over these constraints. The next step is to define an probabilistic objective function which tries to place unlabelled data points on this distribution to infer a label. This deduction of a label using an objective function is esentially the prediction of the new data point.

In [2], the authors collect labelled data from Twitter from verified accounts of politicians. Constraints are defined over county, first name of user and follower(of other users or hashtags). To enumerate on these constraints, the authors report that democrats are likely to follow accounts or hastags such as “legedems, dennis kucinich, sensanders, repjohnlewis, keithelli- son, #p2”. Republicans are likely to follow “gop, nrsc, the rga, repronpaul, sen- randpaul, senmikelee, repjustinamash, gopleader, #tcot”. The task of the constraint function is to place users who follow a certain set of constraints on an appropriate position on the probablity distribution. When a new data point is subjected to the objective function, it approximates the position of the data point on this very probability distribution. The class label of this approximate region is the label to be predicted. The authors also propose algorithmic search routines to find the appropriate position of data points on the probablity distribution. The improved-greedy search procedure proposed by the authors produces an accuracy as high as 79.5%.

This approach is further developed in [3]. The authors assign weights to labels and propose to arrange them in a hierarchial fashion. This in turn optimises the search routines which locates an appropriate position on the probablity distribution. In addition, it also make the previously proposed algorithm scalable.

One may argue about the existence of noisy and biased labels in this method. In [4], the authors propose an approach to use these erroneous labels to our advantage. In fact, in [2], the authors show that even if we ignore noisy labels, the method yields accuracies competitive to standard supervised approaches.

Label Regularisation overcomes a huge step in the process of data mining - annotation of data. This step is not only expensive in terms of human labour but also time consuming. It may also be subject to biases leading to erroneous predictions. In addition, this method yields highly accurate results. The merit of this method lies in the fact that it uses a very small proportion of labelled data to learn and predict labels of unlabelled data in the dataset.

8 Sociological Aspects of Mining Personality Traits

Mining social networks for personal data of users has its advantages. However, there are concerns that must be addressed to ensure that this powerful tool is not abused. Privacy of users is of prime importance.

Recently, there have been reports that personal data of users shared with thrid party organisations has been abused. [26] [38]. The EU General Data Protection Regulation(GDPR), has put forward strict laws to ensure privacy of users [8]. However, the challenge remains in the implemenation of the same.

We have also seen the important census data to accurately predict personality traits. While this is true, census data can be the cause of identity theft and even raise concerns about national security. Aadhaar, a digital census programme in India which contains personal information and biometrics of citizens was recently subjected to secuity concerns [17]. The Aadhaar initiative helps organisation easily access digital records of citizens. Security and ethical concerns lurk over this programme.

In addition to security concerns, researchers have also raised concerns over the ownership of data. It is believed that while only a small section of the society owns data of a larger section of the society, inequality between different socio-economic classes is bound to drive up.

Correct and inclusive digital representation of all sections of the society is important. Only then can the ground truth be easily represented in form of datasets. The classification of gender is always considered as a "binary classification problem". In a real world scenario, gender is far from being a binary variable. It is in fact, a spectrum. Social media paltforms do not allow users to mention any other gender than male or female. In addition, researchers have not annotated datasets manually with a spectrum of genders.

Data mining methods to predict political affiliation of users have been subject to scrutiny. Elections held in the digital era are believed to have been swayed by using data mining tools to convinve individuals to cast their vote for specific parties.

Prediction of personality traits using social media is a very powerful tool. Improper regulation or misuse can lead to a society in disarray. Popular television series Black Mirror [9], theorises of a dystopian society caused by the ill effects of technology.

9 Conclusions and Future Work

In this paper, we have presented three data mining methods which can very efficiently predict personality traits of users. Using these methods, we have predicted gender, ethinicity, location and political affiliation of users.

We have shown how Bayesian methods are efficient in predicting gender and ethinicity. We have also modeled a special case of the Bayesian method known as hierarchial Bayesian method. In the hierarchial Bayesian method, we augment the original dataset with census or gazette data. This produces high accuracy predictions using small datasets. We have also shown that Bayesian methods are not the most optimal methods for predicting gender. However, it is very efficient as it can be easily implemented unlike other methods which have certain constraints in their implementation techniques.

Support Vector Machines(SVMs) have been used to predict gender, location of user and political affiliation. A reverse approach where data from hastags and checkin locations on platforms like Foursquare have been used to predict location and political affiliation with high accuracy. A similar approach to the hierachial Bayesian model known as the Stacked SVM has been used. SVM is a powerful method as multiple tookits are readily availble for implementation. In addition, non-linearly separable data can be linearly separated using the kernel trick very easily.

Bayesian methods and SVMs have issues with regards to scalability. In addition, the cost of building an annotated set of data is very high. To tackle these issues, we discuss a semi-superivsed approach known as Label Regularisation. We use this method to predict political affiliation of users. The noveltly of this method lies in the fact that it uses a small proportion of labelled data to assign labels to unlabelled data in the dataset. Challenges of data mining such as noise and bias can be easily tackled using this method.

There are various issues that remain in the domain of predicting personality traits. Future work in these domains can prove to be beneficial for the field of web mining at large. First, methods discussed in this

[8] https://www.eugdpr.org/

[9] `https://en.wikipedia.org/wiki/Black_Mirror`

paper are very specialised to suit a particular context. Efforts have been made to show that SVMs can be generalised to a certain extent in terms of non-English text for predicting gender. We propose research focus on developing generalised methods in the form of frameworks encompassing large geographics and native languages.

Second, the usage of personality traits in form of features can yield higher prediction rates. For example, ethinicity of a user can help us predict their location. In turn, location can help us predict political affiliation. This transitive property can be beneficial to deduce traits which are not easily inferred using traits which can be easily inferred.

Third, social media offers us a rich repository of multimedia content. Images, videos and gifs are posted in plenty by users. Multimedia content analysis can be beneficial to predict attributes such as location. In [44], the authors determine location of images on Flickr. The same premise can be extended to determine location of users from shared images. The rich information embedded in multimedia content can determine complex traits using semi-supervised and unsupervised methods.

References

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Ehsan Mohammady Ardehaly and Aron Culotta. Inferring latent attributes of twitter users with label regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195, 2015.

[3] Ehsan Mohammady Ardehaly and Aron Culotta. Mining the demographics of political sentiment from twitter using learning from label proportions. *arXiv preprint arXiv:1708.08000*, 2017.

[4] Ehsan Mohammady Ardehaly and Aron Culotta. Learning from noisy label proportions for classifying online social data. *Social Network Analysis and Mining*, 8(1):2, 2018.

[5] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[6] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[7] Smriti Bhagat, Moira Burke, Carlos Diuk, Ismail Onur Filiz, and Sergey Edunov. Three and a half degrees of separation. *Facebook Research Blog*, 2016.

[8] Tobias Bocklet, Andreas Maier, and Elmar Nöth. Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression. In *international conference on Text, Speech and Dialogue*, pages 253–260. Springer, 2008.

[9] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[10] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. epluribus: Ethnicity on social networks. *ICWSM*, 10:18–25, 2010.

[11] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[12] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, 2013.

[13] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 192–199. IEEE, 2011.

[14] Josh Constine. Facebook now has 2 billion monthly users and responsibility, 2017. https://techcrunch.com/2017/06/27/facebook-2-billion-users. Accessed: 19/03/2018.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

[17] Pranav Dixit. India's national id database with private information of nearly 1.2 billion people was reportedly breached, 2018. https://www.buzzfeed.com/pranavdixit/indias-national-id-database-with-private-information-of. Accessed: 22/03/2018.

[18] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45, 2013.

[19] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification.* John Wiley & Sons, 2012.

[20] Kathleen T Durant and Michael D Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In *International Workshop on Knowledge Discovery on the Web*, pages 187–206. Springer, 2006.

[21] Tristan Fletcher. Support vector machines explained. *Tutorial paper*, 2009.

[22] Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 710–718. Association for Computational Linguistics, 2009.

[23] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[24] Frigyes Karinthy. Chain-links. *Everything is different*, 1929.

[25] Sheheryar Ahmed Khan. The science of deduction, 2017. https://www.huffingtonpost.co.uk/sheheryar-ahmed-khan/the-science-of-deduction-b-14929694.html. Accessed: 20/03/2018.

[26] Ido Kilovaty. The cambridge analytica debacle is not a facebook data breach. maybe it should be., 2018. https://techcrunch.com/2018/03/17/the-cambridge-analytica-debacle-is-not-a-facebook-data-breach-maybe-it-should-be/. Accessed: 22/03/2018.

[27] Renaud Lambiotte and Michal Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, 2014.

[28] Kisung Lee, Raghu K Ganti, Mudhakar Srivatsa, and Ling Liu. When twitter meets foursquare: tweet location prediction using foursquare. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 198–207.

ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[29] Anália Lourenço, Michael Conover, Andrew Wong, Fengxia Pan, Alaa Abi-Haidar, Azadeh Nematzadeh, Hagit Shatkay, and Luis M Rocha. Testing extensive use of ner tools in article classification and a statistical approach for method interaction extraction in the protein-protein interaction literature. In *BioCreative III Workshop 2010*, pages 1–5. University of Delaware, 2010.

[30] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, 12:511–514, 2012.

[31] Gideon S Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th international conference on Machine learning*, pages 593–600. ACM, 2007.

[32] Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics, 2010.

[33] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm*, 11(122-129):1–2, 2010.

[34] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[35] Bayu Yudha Pratama and Riyanarto Sarno. Personality classification based on twitter text using naive bayes, knn and svm. In *Data and Software Engineering (ICoDSE), 2015 International Conference on*, pages 170–174. IEEE, 2015.

[36] Delip Rao, Michael J Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. Hierarchical bayesian models for latent attribute detection in social media. *ICWSM*, 11:598–601, 2011.

[37] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[38] Siladitya Ray. Facebook ceo mark zuckerberg apologises for cambridge analytica incident, calls it a "major breach of trust", 2018. https://www.medianama.com/2018/03/223-zuckerberg-apologises-for-cambridge-analytica-incident/. Accessed: 22/03/2018.

[39] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.

[40] Douglas Rushkoff. *Program or be programmed: Ten commands for a digital age.* Or Books, 2010.

[41] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[42] Doris Schmallegger and Dean Carson. Blogs in tourism: Changing approaches to information exchange. *Journal of vacation marketing*, 14(2):99–110, 2008.

[43] Elizabeth P Schneider, Emily E McGovern, Colleen L Lynch, and Lisa S Brown. Do food blogs serve as a source of nutritionally balanced recipes? an analysis of 6 popular food blogs. *Journal of nutrition education and behavior*, 45(6):696–700, 2013.

[44] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[45] David J Stillwell and Michal Kosinski. mypersonality project: Example of successful utilization of online social networks for large-scale social research. *American Psychologist*, 59(2):93–104, 2004.

[46] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[47] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59, 2012.

[48] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.