Machine "Gaydar": Using Facebook Profiles to Predict Sexual Orientation

Nikhil Bhattasali[1], Esha Maiti[2]
Mentored by Sam Corbett-Davies
Stanford University, Stanford, California 94305, USA

Abstract

The rise of social media and the large, rich datasets they make accessible have allowed us to learn about people not only through information shared explicitly, but also through information shared implicitly in the form of trends and patterns. Here, we emphasize the value of implicit data by creating a machine-learning algorithm that uses basic information, photos, and published text on Facebook profiles to predict sexual orientation in males. We constructed a model with Naïve Bayes classifiers and a Support Vector Machine, performing on different types of data. We used 10-fold stratified cross-validation on our dataset as a measure of generalization error. Currently, we have created a model with an accuracy of 91.02%.

I. Introduction

In today's world, our digital lives have merged seamlessly with our personal and professional lives. We use social media to share life events, pictures, interests, and opinions. We cultivate an online identity, both explicitly and implicitly, through what we share and how we choose to communicate on certain social media platforms [1].

This phenomenon has not gone unnoticed by the people and organizations that are interested in studying other people. Companies make predictions about which products different individuals would likely be interested in [2]. Politicians target key demographics to mobilize campaign support. Researchers study patterns to understand how humans interact and communicate.

Similarly, we want to know if there is enough implicit information on people's Facebook profiles to make a prediction of their sexual orientation. On one level, a learning algorithm to distinguish non-heterosexuals is a lighthearted way to study the power of implicit information and to allow interested parties to connect with similar others. On a deeper level, creating a machine learning algorithm is a very organic way to study a population, as it minimizes prior assumptions, being initialized as a "blank slate" and developing patterns based solely on the data [3].

II. Related Work

There is not an extensive amount of existing literature describing research concentrating on sexual orientation alone based on social media profiles; however, there have been multiple machine learning algorithms developed to use very few basic features of social media websites to determine a number of highly sensitive personal attributes. One of these algorithms was developed by Kosinksi, Stillwell, and Graepel [1]: based solely on Facebook likes, detailed demographic profiles, and psychometric tests, Kosinksi et al. were able to use dimensionality reduction and logistic regression to determine the sexuality (homosexual vs. heterosexual males), race (African American vs. Caucasian), and

[1] nikhilb@stanford.edu

political stance (Democrat vs. Republican) of each user.

While predicting the sexuality, race, and political stance of a person using their variety of methods is useful, we posit that these attributes could actually be dependent on each other [4]. Then, we could predict one — like sexual orientation — using the others as features and without having to resort to tests and demographic profiles, instead relying on social media information alone.

We use a model that is not only based on Facebook photos and likes, but also on explicitly provided information such as a person's hometown, religious affiliation, relationship status, etc., as well as text communication through timeline posts and photo comments. Taking into account this host of details may prove fruitful for predicting a person's sexuality.

III. Dataset and Features

Our aim in this project was to distinguish heterosexual (non-LGBTQ, $y = 0$) individuals from non-heterosexual (LGBTQ, $y = 1$) individuals. (We made the assumption that sexual orientation is binary, rather than a spectrum, for the sake of classification simplicity.)

Our training set included $N = 167$ Facebook profiles and information included in those profiles in the following categories:

"About" section
- Hometown
- Political affiliation
- Religious affiliation
- Interested in
- Relationship status

"Friends" section
- Number of friends
- Male-to-female friend ratio
- Number of LGBTQ friends
- LGBTQ-to-non-LGBTQ friend ratio

"Profile Photo" section
- Use of rainbow filter (2015)
- Use of equality sign (2012)
- Number of profile photos with one other male
- Number of profile photos with one female
- Number of prom photos with one other male
- Number of prom photos with one female

"Timeline" section
- Status updates
- Photo comments
- Shared links

Our dataset was 62% LGBTQ and 38% non-LGBTQ. In order to restrict our sampling population, we limited our selection pool to only Stanford undergraduate students.

For the LGBTQ portion of our training set, we picked a simple random sample of user profiles from the secret LGBTQ groups for the classes of 2019, 2018, and 2016 at Stanford. For the non-LGBTQ portion, we picked a simple random sample of students from the official undergraduate class groups for the classes of 2019, 2018, and 2016, and we crosschecked with the list of LGBTQ people to ensure no overlap. (We made the assumption that all Facebook users that are not in the LGBTQ groups are not LGBTQ.)

We did not use Facebook's Graph API for the sake of flexibility, and did not use robotic data collection methods due to Facebook's terms of service. Rather, we gathered all our manually and used custom cleaning/parsing scripts to process the data.

We would also like to note that all of our data example were anonymized; the names associated with each datapoint was erased after the collection process in order to protect privacy.

IV. Methods

We decided to integrate two Naïve Bayes classifier components and a Support Vector Machine in our model.

A Naïve Bayes classifier is a generative learning model that predicts the probability of a particular data vector given past distributions of discrete features [5], given by eq. (1):

$$p(y=1|x) = \frac{\Pi_{i=1}^{n} p(x_i|y=1)p(y=1)}{\Pi_{i=1}^{n} p(x_i|y=1)p(y=1)+\Pi_{i=1}^{n} p(x_i|y=0)p(y=0)} \quad (1)$$

Such an algorithm not only helped us determine a user's sexuality from the text from his Facebook "Timeline" section, but could also provided us with a few of the most commonly used words among users marked as LGBT for further analysis [6].

Another algorithm that proved central was the Support Vector Machine, which is a discriminatory learning algorithm that finds an optimal decision boundary between the two different kinds of data points: by finding a hyperplane (**w**, b) [7] to maximize the quantity

$$\gamma = \min_i y^i\{< \boldsymbol{w}, \phi(\boldsymbol{x}^i) > - b\} \quad (2)$$

as exemplified in Fig. 1 below:

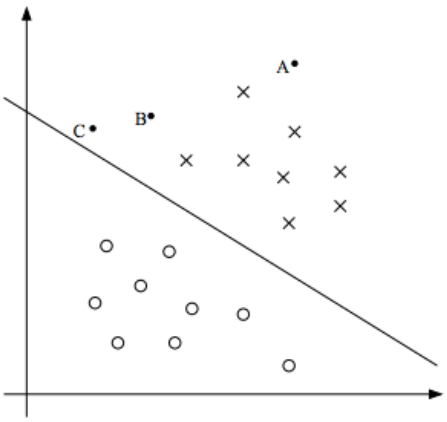

FIG. 1. An example of a decision boundary created by running an SVM on a system

Because our training examples are described by have both binary and continuous features, we created a model with two Naïve Bayes classifiers: one that analyzes the discrete attributes in a profile's "About" and "Photos" sections, and one that analyzes the text from the profile's "Timeline". Each of these computes a score that is fed into the Support Vector Machine and combined with other continuously-valued features to assign a final classification, as illustrated below in Fig. 2:

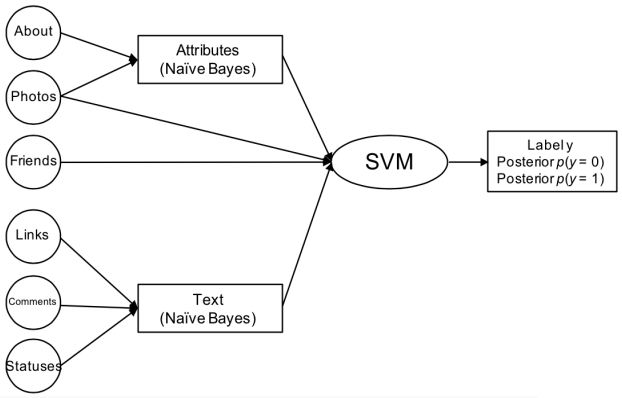

FIG. 2. A diagram of the method involving Naïve Bayes and an SVM that was implemented

In order to create a model that would generalize well, we did not use an empirical prior based on the skewed dataset, but rather computed a more accurate prior probability to be used in the algorithms based on the entire undergraduate population and LGBTQ group.

V. Results and Discussion

As of the latest version, our model runs with a training error of 2.40% and a generalization error of 8.98%. This suggests overfitting, and thus high variance, requiring a fix such as decreasing the number of features or adding more training examples to the training set.

Nevertheless, the error is fairly low, considering that we did not have time to fine-tune the algorithm to minimize error, and works quite proficiently on outside profiles that were not used in the dataset.

Below, we provide a confusion matrix, which shows the predictive performance of the classification algorithm (Fig. 3).

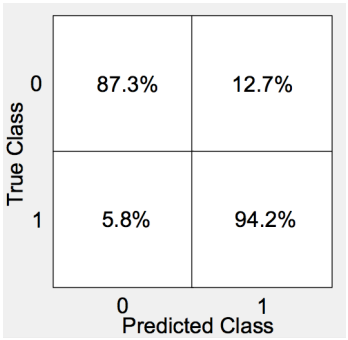

FIG. 3. Contingency table. The top left box shows the true positive rate, and the bottom right box shows the true negative rate. The top right and bottom left show the Type I and Type II errors, respectively.

Here, we include the ROC curve, which summarizes the performance of our classifier over all possible thresholds (Fig. 4).

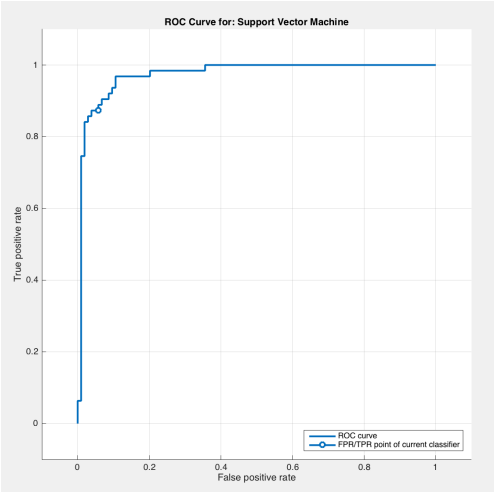

FIG. 4. ROC curve. The optimal curve has an area underneath it equal to 1. This curve has an area underneath it equal to 0.973, which is quite close to 1.

As one can see, running an SVM on a dataset using only two of our features can still create an effective decision boundary between LGBTQ and non-LGBTQ data. This is illustrated in Fig. 5 below:

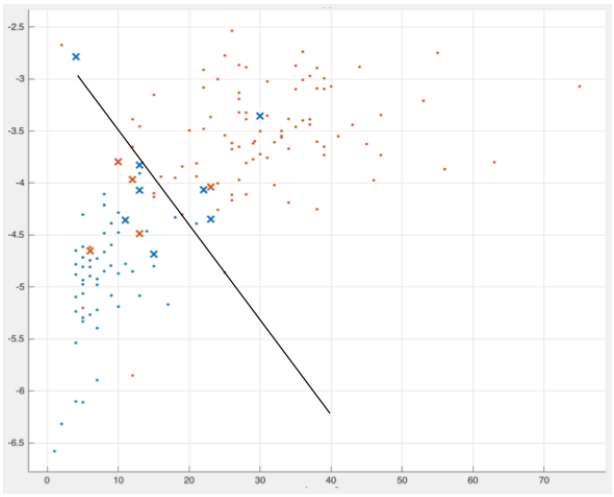

FIG. 4. 2D SVM decision boundary. This plot shows the decision boundary for just two features: the number of LGBTQ friends (x-axis) and the log ratio of non-LGBTQ-to-LGBTQ friends (y-axis). A remarkably clear separation between the two classes can be observed.

We observed some interesting results from the Naïve Bayes text classifier that we ran for the text on each profile's "Timeline". Some of the words with highest predicting power, given by

$$\log\left(\frac{\phi_{y=1}}{\phi_{y=0}}\right) \tag{2}$$

for an LGBTQ user were (Table 1):

Word	*Rank of Predictive Power*
LGBTQ	1
Clinton	2
Gaga	6
Stonewall	7
Horror	8
Transgender	11
Terra	14
LGBT	15
Equality	16
Parenthood	18
Gay	22
Gender	28
Queer	32
Caitlyn Jenner	36, 37

TABLE 1. Ranking of words based on their predictive power scores.

VI. Conclusion and Future Work

Our work is an example of how a powerful classifier can be built with some of the fundamental machine learning techniques. Combining a carefully-curated dataset, flexible machine learning model, and strong error testing metrics, we show both the power of implicit information and social media. Although an individual profile may not share information explicitly, it is possible to use the aggregation of sparse data from many profiles to train a learning algorithm to tell us whether certain present (or missing) data contributes to the probability [1, 2] of a person being LGBTQ or non-LGBTQ. Such a finding is interesting from both a scientific and ethical perspective.

In the future, we would consider using a neural network on a similar but larger dataset, as the way that neural networks extract features may lend itself naturally to describing the complexities of identity and different subsets of discrete features. For instance, does omitting the "Interested in Women" field have more predictive power for someone in Saudi Arabia than someone in California? Such a model would allow these kinds of questions to be studied, and it would perhaps importantly lead to the confirmation or challenging of mainstream stereotypes.

VIII. Acknowledgements

We would like to thank our CS 229 professor Andrew Ng and our project mentor Sam Corbett-Davies for guidance, as well as Stanford's Department of Computer Science for the opportunity to perform and present this research.

VIII. References

[1] M. Kosinski, D. Stillwell, and T. Graepel, *Private Traits and Attributes are Predictable from Digital Records of Human Behavior*, PNAS **110**(15), 5802-5805 (2013).
[2] F. Bodendorf and C. Kaiser, *Detecting Opinion Leaders and Trends in Online Social Networks*, SWSM **18**, 65-68 (2009).
[3] E. Alpaydin, *Introduction to Machine Learning* (The MIT Press, Cambridge, 2014).
[4] A. Stone, *Diversity, Dissent, and Decision Making: The Challenge to LGBT Politics*, GLQ **76**(2), 179-206 (2011).
[5] A. McCallum, K. Nigam, *A Comparison of Event Models for Naïve Bayes Text Classification*, AAAI-98 **752**, 41-48 (1998).
[6] W. Kraaij, M. Spitters, and M. Van Der Heijden, *Combining a Mixture Language Model and Naïve Bayes for Multi-Document Summarisation*, SIGIR2001 (2001).
[7] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, *Support Vector Machine Classification and Validation of Cancer Tissues Using Microarray Expression Data*, Bioinformatics **16**(10), 906-914 (2000).