Predicting Sexual Orientation based on Facebook Status

Aaron Loh, Kenneth Soo, and Huize Xing

Stanford University, Palo Alto, California 94305

***Abstract*—Do people with different sexual orientations express themselves differently on social media? This paper presents an exploratory study for predicting the sexual orientation of a Facebook user using text features extracted from their status updates. We examine which types of features are most informative for a reliable prediction of sexual orientation in the social media setting, and we explore the performance of different models such as SVM, Naive Bayes, Logistic Regression and Random Forest. Our result shows that there are distinctions in the ways people with different sexual orientations express themselves, and that male homosexuals are slightly more likely to talk in a effeminate manner.**

Keywords—Sexual Orientation, Facebook, Status updates, Logistic Regression, Multinomial Naive Bayes, Random Forest, SVM, ROC

I. Introduction

A. Task Definition

Our task is to predict a user's sexual orientation from his/her Facebook status updates. The input will be a Facebook status update, and we want our algorithm to output a prediction of the sexual orientation for the user that wrote the update.

B. Motivation

Social media websites such as Facebook have become an inseparable part of many people's lives. However, with every post we Like, every photo we upload and every status update, we leave behind a set of digital footprints that may be used to uncover our traits and personality. Our project specifically aims to analyze indicators of sexual orientation in Facebook status updates. We wish to explore the subtle differences in how people of different sexual orientations express themselves on social media.

Additionally, in our CS 221 Project that we worked on in parallel, we explored how males and females express themselves differently in their Facebook updates and constructed models to predict gender based on their Facebook status. Combining the results from our CS 221 project and CS 229 project, we seek to test the stereotype that male homosexuals tend to use more feminine language.

C. Challenges

Our task came with the challenges listed below. In the rest of the paper, we will discuss how we overcame them.

Challenge 1: The dataset was skewed with most of the population being heterosexual. Guessing that everyone was heterosexual would have given us $> 90\%$ accuracy. Therefore, we needed to measure the result using a more meaningful metric than test error.

Challenge 2: It is not immediately obvious what text features from Facebook status updates would distinguish homosexuals from heterosexuals, and it is difficult to extract all the important semantics from natural language.

Challenge 3: We expect there to be much noise in the data, including misspellings, useless features, unoriginal facebook updates (something not written by the person itself) and non-standard language variation. Such noise could hinder our learning.

Challenge 4: Many people do not post long status updates on Facebook. Using just a few words to predict a person's sexual orientation is very difficult.

II. Literature Review

There has not been an extensive amount of research on predicting sexual orientation from textual evidence. In [1], Kosinski and Stillwell showed that personality (based on the Five Factor Model) can be predicted based on properties of Facebook and Twitter profiles. In another paper, Kosinski researched how sexual orientation can be predicted based on Facebook likes [2]. Furthermore, analyzing friendship associations has shown to be predictive of sexual orientation [3], and the conclusion was that the percentage of a given user's friends who self-identify as LGBT has strong correlation with the user's sexual orientation. Many related papers talk about how gender can be extracted from speech features [4][5], many of which can be used to test for correlation with sexual orientation as well. Lastly, a past CS 229 project used Naive Bayes classifier to predict male sexual orientation based on manually collected data from 167 Facebook profiles [6]. Our project plans to expand on the previous CS 229 project by using a much larger dataset and applying different models and features to get a better result.

III. Dataset and Features

We used data from the myPersonality project (http://mypersonality.org) after obtaining the kind permission of the database owner, Dr. Michal Kosinski (Stanford GSB).

The project database contains more than 22mn status updates of 154k Facebook users. The label for the sexual orientation of a user was generated by comparing the user's gender to that of his/her partner. For this project, we focused on the binary classification of ”homosexual” versus ”heterosexual”. Some data processing was needed, which we describe here:

A. *Large Dataset*

Our computers did not have the computational resources to train a model on a dataset this large. As such, we decided to take 560k status updates to work on instead. We kept in mind that we could use additional data if we wished to improve our model accuracy later on. Indeed, we saw that our metric for success, Receiver Operating Characteristic Area Under Curve (ROC AUC), improved as we moved from smaller datasets to larger datasets, suggesting that more data could be helpful.

B. *Status updates in foreign language*

Some of the updates were not in English. Theoretically, our model should be able to handle such cases, but because the proportion of training examples we had for any particular foreign language was less than 0.2%, this was insufficient for the model to train on. As such, we decided to remove them from the dataset. This was done through filtering the user's language preferences in Facebook, though this was not a complete filter as there were some users using the English version of Facebook who posted in a foreign language.

C. *Word Stemming*

We used word stemming to group words with the same root together. This increases the frequency count of certain words, making them more useful for prediction. For example, the words ”excited”, ”exciting”, ”excitement” will be reduced to the root ”excit”. Here we assume that sexual orientation does not greatly affect the form of the words used.

D. *Separating Males from Females*

The difference between male and female speech had been shown in our CS 221 project. We hypothesised that male homosexuals and female homosexuals talk differently in their Facebook updates. Therefore, we decided to separate our dataset into males and females and run our model separately on each group.

E. *Feature Selection*

We hypothesised that homosexuals may use words and phrases that are different than their counterparts. Therefore, we chose n-grams to be our feature to capture those differences. Additionally, we picked counts of periods, exclamation marks, smileys, and capital letters as additional features. These additional features were selected as we found significant differences in how homosexuals and heterosexuals used them.

IV. Our Approach

We used the rich functionality of the scikit-learn python library [7] to do much of our modeling and learning. In this section, we describe the experiment we performed and the learning models we used. Figure 1 shows a graphic representation of our pipeline process.

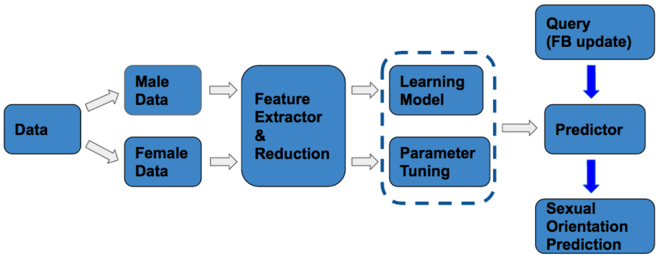

Fig. 1: Graphic Representation of Our Machine Learning Approach

A. *Feature Reduction and Noise Elimination*

To avoid overfitting, we attempted to reduce the number of n-gram features. We eliminated stop words since they just added noise to the data. We also used document frequency threshold to eliminate n-gram features that appear too few or too frequently. For example, if an n-gram feature appeared in over 90% of the Facebook status in our dataset, that feature was not likely contribute to learning and was thus removed. Both maximum document frequency threshold and minimum document frequency threshold were parameters that we tuned to maximize F1-score.

B. *Parameter Tuning and Result Metric*

We used the Grid Search functionality of the scikit-learn library to search a set of parameters that optimized our prediction result. Some of our parameters include n-gram ranges, max document frequency, min document frequency and kernel type (SVM only). Our dataset was first divided into 70% training data and 30% testing data. For the 70% training data, we used Grid Search with 3 fold cross validation to maximize F1-score. We then tested our model on the 30% test data to measure our generalization accuracy.

To obtain a meaningful metric, we resorted to using F1-score and Receiver Operating Characteristic (ROC) as our score function for Grid Search. These metrics are good indicators of the accuracy of our models despite the skewed nature of our dataset, which had less than 10% of the data points being generated from homosexuals. Additionally, we also generated a confusion matrix for error analysis.

C. *Learning Algorithms*

For our learning algorithms, we experimented with a set of models each having their own strengths and weaknesses.

1) Multinomial Naive Bayes: We applied Multinomial Naive Bayes classification, a probabilistic classifier that calculates and chooses the class with highest posterior probability. The Multinomial Naive Bayes model assumes that whether a person is a homosexual or heterosexual is determined with probability $p(y)$. Then as that person is writing his Facebook status, each word k is generated from the probability distribution $p(x_i = k|y)$ for any word position i in the Facebook status.

The posterior probability is then calculated using the equation:

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

and the class with the higher posterior probability is chosen.

Just like other Naive Bayes models, the Multinomial Naive Bayes model makes the strong assumption that features are conditionally independent given the class label.

2) Logistic Regression: We also used logistic regression, a discriminative classifier, on our data. Logistic regression does not make the strong probabilistic assumption that Naive Bayes makes, which makes it more flexible and less prone to errors in our probabilistic assumptions.

Our prediction is made using the equation:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

withe θ being chosen the maximize the log likelihood function:

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$

The function is maximized using gradient ascent.

3) Support Vector Machine: We explored the SVM model that does well when the classes are separated by a good amount of margin given the kernel used. In our experiment, we found that SVM trains much slower than the Logistic and Multinomial models. We therefore used a smaller dataset of 50k facebook updates due to the expensive computational cost. We further explored different kernels, including linear kernel and Radial Basis Function (RBF) kernel.

4) Random Forest: Lastly, we explored the Random Forest Model, which is a bootstrap aggregating approach applied to decision trees. At each split of the decision tree learning process, a random subset of features were available for the split. 500 trees were generated, and the prediction of the Random Forest is given as the mean prediction of the trees.

V. Result and Analysis

A. ROC, F1-scores and Confusion Matrix

Table I and II below shows accuracy result for different learning models. ROC AUC is the area under the curve of the Receiver Operating Characteristic while F1 score is a composite measure of the precision and recall.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The results were based on a list of the best parameters chosen for each learning algorithm. The parameters explored and the best parameters for each algorithm can be found in the Appendix.

Out of the three learning algorithms, SVM gave the best result in term of the F1 score, which suggests that there is a wide enough margin separating the two classes.

Model	ROC AUC[1]	F1 Score
SVM	0.61	$0.94_{(0.98,0.17)}$[2]
Multinomial	0.52	$0.91_{(0.96,0.21)}$
Logistic	0.57	$0.92_{(0.97,0.20)}$
RF	0.58	-

TABLE I: Accuracy Scores for Males

[1]ROC AUC stands for Area under the curve for the Receiver Operating Characteristic curve. [2]For F1 score, the figures in parenthesis indicate F1-scores for heterosexual and homosexuals respectively.

Model	ROC AUC	F1 Score
SVM	0.62	$0.85_{(0.93,0.36)}$
Multinomial	0.58	$0.84_{(0.93,0.30)}$
Logistic	0.62	$0.84_{(0.94,0.24)}$
RF	0.63	-

TABLE II: Accuracy Scores for Females

Figure 2 and 3 shows the corresponding ROC curves for SVM algorithm. The ROC score for both females and males were above 60%, which told us that there were distinctions in how the homosexuals expressed themselves on social media, even if the distinction was not great enough to consistently predict one's sexual orientation.

Our confusion matrix shows that our model performs well with heterosexuals, as given a group of heterosexuals, it is able to label more than 90% of them correctly. However, it does not perform as well with homosexuals, only being able to label 19.5% of male homosexuals accurately, and 28.2% of female homosexuals correctly. Interestingly, we see that the model performs better with females. This is consistent with our CS 221 results, which also found that the model was able to perform better with females as compared to males. This low accuracy suggests that we need to find stronger indicators for when a person is homosexual.

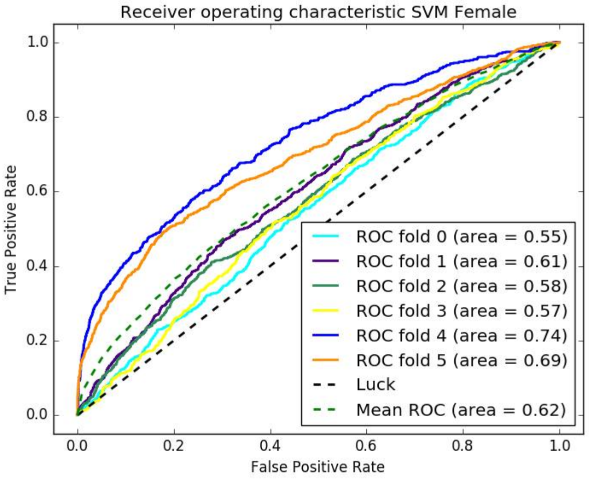

Fig. 2: ROC Curve using SVM for Females

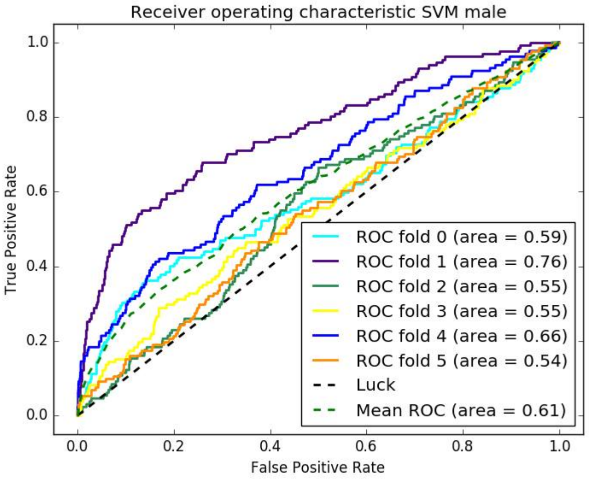

Fig. 3: ROC Curve using SVM for Males

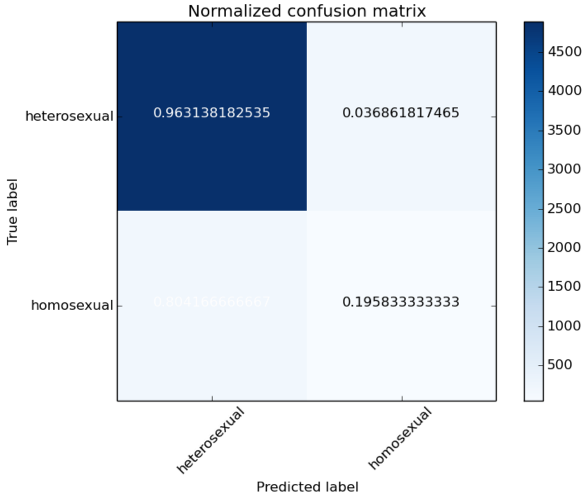

Fig. 4: Confusion Matrix for SVM for Males

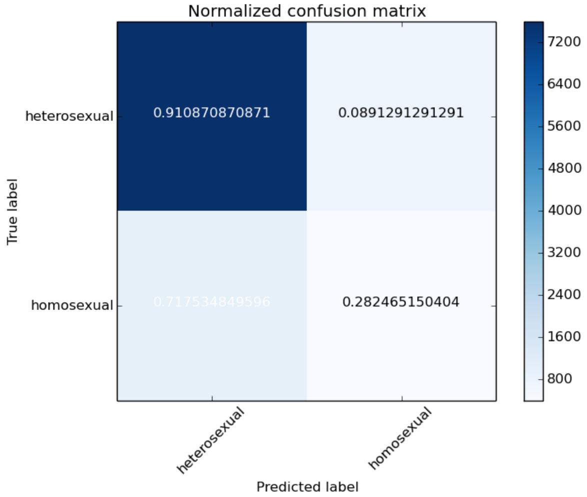

Fig. 5: Confusion Matrix for SVM for Females

B. Top Features and Misclassifications

The top features for straight males are: girlfriend, wat, wife, texas, angel, mile, game, drive, gotta, goal. Top features for homosexual males are: gay, omg, comment, sister, yay, okay, funni-, hair, b***h, sex. Top features for straight females are: boyfriend, husband, famili-, clean, today, bless, hubbi-, pray, work, church. Top features for homosexual females are: aint, wanna, your, hahaha, drunk, text, gunna, random, f**k, mum.

Our top features show that mentioning a partner of the opposite gender (e.g. when males mention ”wife”) is a strong indicator of heterosexuality. Also, using conservative-related words like ”Texas”, ”Church”, and ”Bless” are also good indicators. Our model also showed that a male homosexual was 4 times more likely to use the word ”gay”. In fact, a male who mentions ”gay” has a 1 in 4 chance of being homosexual.

Our top features also show that age could be a confounding factor that is driving the differences between homosexuals and heterosexuals. For example, it may be popular for young girls to declare on Facebook that they are in a ”relationship” with a good female friend even if they are heterosexual. Further evidence for this is that the top word features for female homosexuals are more often associated with young people, whereas the words for heterosexuals are more associated with older people (e.g. husband, church, work).

One limitation we found in our feature engineering is that our features are not able to distinguish between the different meanings of words in different contexts. For example, the word ”gay” can have different meanings depending on the context.

A human is able to discern the difference but the features in our model were not able to discern them as well. For this reason, we found multiple misclassified examples that were thrown off by the use of the word ”gay”, in contexts where it did not refer to sexual orientation.

C. Are Homosexual Males More Effeminate in Their Speech?

We had earlier built and trained models to predict gender for our CS 221 project. Using the gender model, we found that our classifier predicted that 40% of all heterosexual males were females. We ran the same model again on a dataset of all male homosexuals, and it predicted 45% of the male homosexuals as female. This meant that a male homosexual was 5 percentage points more likely to be predicted female. Our results suggest that there is slight evidence that male homosexuals express themselves more like females, as compared to male heterosexuals. This claim should be investigated further before any conclusions may be formed.

VI. Conclusion and Future Work

Our results show that there are differences in how homosexuals and heterosexuals express themselves on Facebook, and also that there is evidence that male homosexuals have a feminine edge in their status updates. However, our models are still unable to predict one’s sexuality consistently enough to be used in practice.

To better our models, we would consider using more sophisticated word features such as Word2Vec [8]. The fact that our ROC AUC and F-1 improve with the data size suggests that we can further improve our model by providing more data. We would also look into automatically learning features using neural networks, with the downside that the features would be less intuitive to interpret. However, the use of a Recurrent Neural Network would potentially be suitable for our problem, as it would take into account the interactions between terms.

Acknowledgements

We would like to thank Dr. Michal Kosinski (Stanford GSB) for allowing us to use the data, and for providing valuable input on the direction of our project. We would further like to thank the CS 229 instructors and TAs for their guidance throughout this project.

Appendix A
Parameters used in Grid Search

Parameters	values searched	best value
n-gram	(1, 2), (1, 3), ...	(1, 2)
min doc freq	1, 0.5, 0.1, 0	1
max doc freq	0.9, 0.8, 0.75	0.8

TABLE III: Model Parameters for Logistic Male

Parameters	values searched	best value
n-gram	(1,2), (1,3), ...	(1,2)
min doc freq	0.001, 0.002	0.001
max doc freq	1, 0.95, 0.9	0.95
kernel	linear, poly, rbf	linear

TABLE IV: Model Parameters for SVM Male

Parameters	values searched	best value
n-gram	(1,2), ... (1,7)	(1,6)
min doc freq	0.05,...,0.02	0.01
max doc freq	1, 0.8, 0.9	1

TABLE V: Model Parameters for Multinomial Male

Parameters	values searched	best value
n-gram	(1, 2), (1,4) , (1,7)..	(1, 4)
max doc freq	1, 0.9, 0.95	1
mmin doc freq	0.01,0.02,0.005	0.005

TABLE VI: Model Parameters for Logistic Female

Parameters	values searched	best value
n-gram	(1, 2), ..., (1,7)	(1, 2)
min doc freq	1, 0.001, 0.002	0.001
max doc freq	1, 0.95, 0.9	0.95
kernel	linear, poly, rbf	linear

TABLE VII: Model Parameters for SVM Female

Parameters	values searched	best value
n-gram	(1, 2), ... , (1,4)	(1, 4)
min doc freq	1, 0.001...0.01	1
max doc freq	1,0.8,0.95,0.9, ...	0.9

TABLE VIII: Model Parameters for Multinomial Female

References

[1] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, ”Personality and Patterns of Facebook Usage,” in ACM Digital Library, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2380722.

[2] M. Kosinski, D. Stillwell, and T. Graepel, ”Private Traits and Attributes are Predictable From Digital Records of Human Behavior,” in Proceedings of the national Academy of Sciences of the United States of America, 2012. [Online]. Available: http://www.pnas.org/content/110/15/5802.short.

[3] C. Jernigan and B. mistree, ”Gardar: Faceook Friendship Expose Sexual Orientation,” in First Monday, 2009. [Online]. Available: http://firstmonday.org/article/view/2611/2302.

[4] F. Voegeli, ”Differences in the speech of men and women,” in venusboys.com, 2005. [Online]. Available: http://www.venusboyz.com/PDF/DissertationFVoegeli.pdf.

[5] A. Mukherjee and B. Liu, ”Improving Gender Classification of Blog Authors,” in aclweb.lorg, 2010. [Online]. Available: http://www.aclweb.org/anthology/D10-1021.

[6] N. Bhattasali and E. Maiti, ”Machine Gaydar, Using Facebook Profiles to Predict Sexual Orientation,” 2015. [Online]. Available: http://cs229.stanford.edu/proj2015/019_report.pdf.

[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[8] T. Mikolov, I. Sutskever, K.Chen, G.Corrado and J. Dean, ”Distributed Representations of Words and Phrases and their Compositionality”, NIPS 2013.[Online]. Available: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf