Gender Differences in Sentiment from reddit Comments

Fadi Eid Katrina
Maastricht University
i6 154 222

Olivia Kirk
Maastricht University
i6 153 141

1 Introduction

Identification of an author's gender has been done before with successful classification rates spanning 60% to 80% [8][4], but can it be extended to text that varies widely in form such as a reddit comment? There have been prior successes such as identifying gender from tweets using only the body of text [4]. Because of the differences in personality men and women exhibit [11], it is expected that the sentiments of reddit comments are likely to differ by gender - specifically, women have more positive sentiments than men [6]- and thus, it should be possible to classify if a commenter is male or female. This would be useful in finding the pervasiveness of gender-differences online, an environment where one can act freely through the anonymity of the internet. There can also be commercial gain in text-based identification through targeted ads.

Using a data set of gender-labelled reddit comments, two binary classifiers and two sentiment analysis models are made by implementing naive Bayes and logistic regression [9][2][1][5][7][3][10]. How the size of the training data affects the classification rate will also be investigated. Implementation will be aided by the scikit-learn library.

2 Related Literature

Gender differences in language have been studied by looking at things such as terms used, emoticons and even the stylisation of the written text [8][12].

A lot of research tends towards traditional machine learning algorithms such as naive Bayes classifiers and SVM, where it is found that "methods perform comparably when the categories are sufficiently common (over 300 instances)" [10].

3 Data set

The data set is over a terabyte in size uncompressed and contains over 1.7 billion reddit comments, claiming to be all reddit comments up to that point (end of 2015). Its large size is desirable. The "Ask Men" and "Ask Women" subreddits will be sampled since they indicate a user's self-reported gender in their comment flair (a tag that is shown next to their username). Due to the enormous size of the data, only comments from 2011 to 2013 will be used where the extracted data totals to 1GB in size. This is 1,572,904 lines of JSON objects, 809,290 of which are from "Ask Men" and thus 763,614 are from "Ask Women".

No preprocessing was performed due all text features being potentially desirable, from spelling errrors to writing in capitals. This is enabled by the large size of the data set as it acts as a sort of buffer.

For sentiment analysis, different pretrained embeddings are used with Bing Liu's sentiment lexicon. The model will be unused on the unlabelled reddit comments to determine the positivity and negativity of the average comment from men, women, "Ask Men", "Ask Women" and all of the above.

The data set consists of JSON objects that will be dealt with using Python's json library. This gives proficiency for reading the data files. The raw files are parsed using an original parse and the parsed file is extracted into lists.

4 Models

Standard implementations of logistic regression and multinomial naive Bayes have been implemented. A bag of words matrix is used to represent the semantic meaning for the binary classifier models generated from the train set.

For the sentiment analysis, ConceptNet Num-

Classifier and metric	Value	Citation
NB accuracy	[67%, 81.3%]	[4][2]
NB precision	[77%, 80%)	[9][1]
NB recall	68%	[9]
NB F-score	[63%, 72%]	[1][9]
LR accuracy	83%	[7]
LR precision	72%	[5]
LR recall	55%	[5]
LR F-score	[48%, 62%]	[5]

Table 1: Benchmarks.

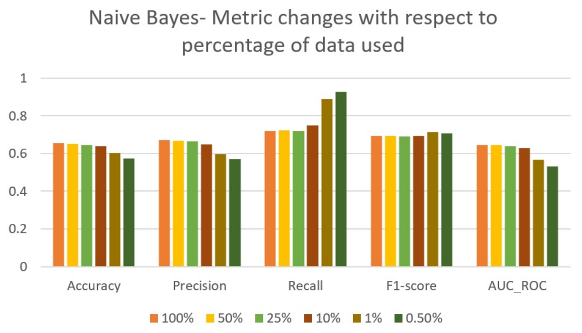

Figure 1: Changes in Naive Bayes' metrics with respect to date input size. For scale, 100% data used is 1572904 data points.

berbatch was originally going to be used, but performed poorly on a sample of 100,000, so GloVe was used instead. Logistic regression and naive Bayes were both used as a classifier.

5 Empirical Results

5.1 Benchmarks

Accuracy, precision, recall and F-score will be used to determine the quality of the models (see Table 1 for some state-of-the-art benchmarks).

To ensure that the models do not over fitting to specific data points, k-fold cross validation will be performed where k = 10 (see Figures 6 and 7 showing different results).

5.2 Experimentation

5.2.1 Effect of the Amount of data on the Classifiers Accuracy

The classifiers were trained and tested based on the percentage of the original total data. The tests were run at 100%, 50%, 25%, 10%, 1% and 0.5%. See Figures 1 and 2 for results.

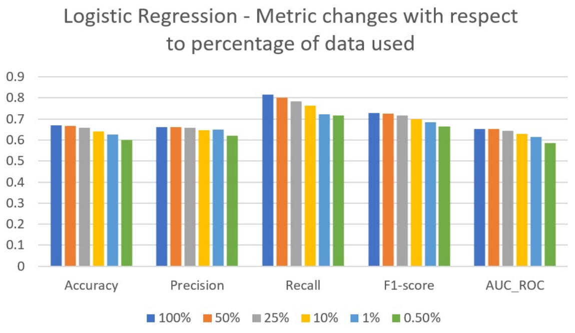

Figure 2: Same as Figure 1 except for logistic regression

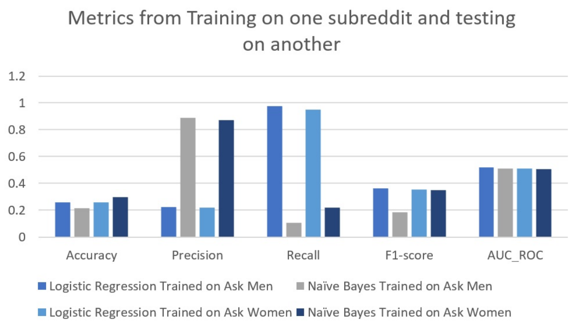

Figure 3: Metrics of both binary classifiers from training on either "Ask Men" or "Ask Women" and then testing on the other subreddit.

5.2.2 Semantic Differences between Men and Women in different subreddits

The classifiers were trained and validated using the data from one subreddit and tested on the other using the full data set. See Figure 3 for results.

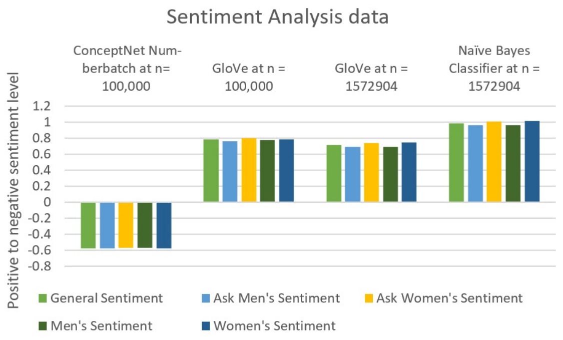

Figure 4: All sentiment analysis data

5.2.3 Sentiment Differences

Due to resource constraints, only the GloVe 42B embedding was used on the full data set.

The average sentiment of the data, men, women, commenters of ”Ask Men” and ”Ask Women” are identified with results shown in Figure 4.

6 Analysis

6.1 The Effect of Data Size

The effect of input data quantity has been well established: more is better. Its also been established that classifying methods tend to reach a plateau after given a certain quantity [10]. Figures 1 and 2 reaffirm this as even when reducing the used data to a tenth of the total size, the classifiers' performance remains largely unaffected.

More interesting trends can be seen in the naive Bayes classifier than in logistic regression. This contrasts with the results from Figure 3 where it is the logistic regression classifier that has great recall where as naive Bayes has superior precision suggesting that it is not caused solely by the classifier model but also by the trends in data.

6.2 Semantic Differences between Men and Women

The results from Figure 3 are interesting because it supports the hypothesis of gender differences being visible in text. As Figure 5 shows, the distribution of men and women is even across the subreddits but the majority gender corresponds to the subreddit name. Because the classifiers were trained on such skewed data sets, it stands to reason that they would fail to become generalisable. Figures 6 and 7 show the validation results run on different data sets for both logistic regression and naive Bayes. For this experiment, the validation set was higher than the ”baseline” models as was the accuracy on unseen data lower.

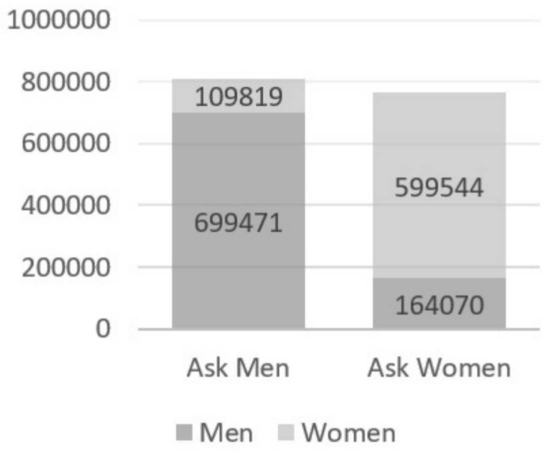

Figure 5: Gender proportions of the data, stratified by subreddit. The number labels denote the number of comments belonging to each gender for that subreddit.

This would suggest overfitting especially as the accuracy is very low. Moreover, looking at the area under the ROC curve so approximately 0.5 as seen in Figure **??**, which emphasises its lack of class-discriminating ability.

These pieces of data imply that there is a clear different in the comments of men and women, as overfitting to one causes results worse than randomly guessing.

6.3 Sentiment Analysis

It was expected that there would be a clear difference in sentiment between men and women, however the data provides little support for this theory. Figure 4 does show some discrepancy between men and ”Ask Men” as well as women and ”Ask Women” (expected to correlate due to Figure 5). Both deviate from the general sentiment by approximately the same amount but the total distance between the two never exceed 0.06, which is not compelling.

It is possible that this lack of significant discrepancy is due to using pretrained embeddings that fail to capture local sentiments and connotations. Another explanation could be that due to the subreddits' question-answer nature, a lot of neutral sentiments skew the data due to occurrences of follow up questions.

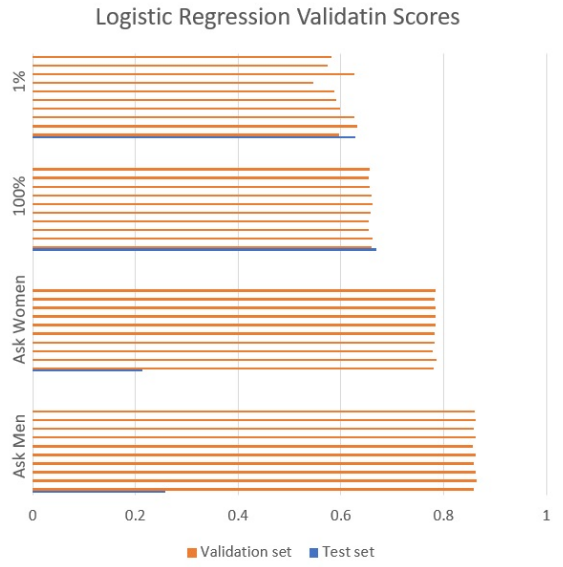

Figure 6: The top set if from using only 1% of the data set, the second from using the full 100%. The third and fourth are from training a logistic regression classifier on ”Ask Women” and ”Ask Men” before testing on the other subreddit, showing abysmal results from the blue test result.

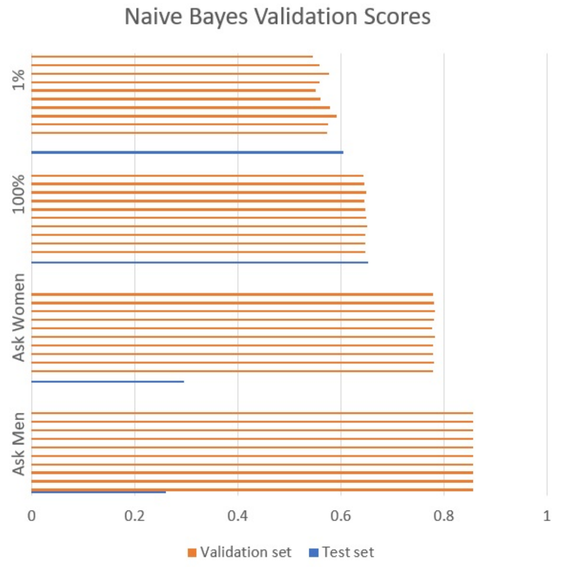

Figure 7: Same as Figure 6 except with the naive Bayes classifier.

7 Conclusion

Gender classification from text has been done before and now is shown to be do able from form varying texts such as a reddit comment. Furthermore, this can be done without special implementations of classifiers nor preprocessing of the data given there is enough data.

Possible avenues for future research could be accounting for sexual orientation or transgenderism and seeing their effects on the results. Another approach would be to create a semantic embedding from multiple subreddits as to capture reddit lingo better rather than just restricting the data sources to two.

References

[1] Citius: A naive-bayes strategy for sentiment analysis on english tweets.

[2] Lei Huang Alec Go, Richa Bhayani. Twitter sentiment classification using distant supervision.

[3] David Madigan Alexander Genkin, David D. Lewis. Large-scale bayesian logistic regression for text categorization.

[4] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter.

[5] Frederic Bechet Hussam Hamdan, Patrice Bellot. Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis.

[6] Sukhvinder Uppal Mike Thelwall, David Wilkinson. Data mining emotion in social network communication: Gender differences in myspace. 2009.

[7] Atoosa Mohammad Rezaei. Author gender identification from text. 2014.

[8] Cevdet Aykanat Tayfun Kucukyilmaz, B. Barla Cambazoglu and Fazli Can. Chat mining for gender prediction.

[9] Christos Troussas and Jaime D. L. Caro. Sentiment analysis of facebook statuses using naive bayes classifier for language learning. 2013.

[10] Liming Yang and Xin Liu. A reexamination of text categorisation methods. 1999.

[11] 2 * Yanna J. Weisberg, 1 Colin G. DeYoung and Jacob B. Hirsh3. Gender differences in personality across the ten aspects of the big five. 2, 178, 2011.

[12] Cathy Zhang and Pengyu Zhangy. Predicting gender from blog post. 2010.