Tweets Analysis of "The walking dead" Using Machine Learning

Yixuan Zhao
Alaa Ayach

Problems

Sentiment Analysis

Gender prediction

Combine together

Assumption

- People watch tv show “The Walking Dead”
- Post what they are truly feeling, not fake

Data

collect:

1.tweets from user “The walking fans”

2.tweets have “the walking dead damn” and ”the walking dead sucks”

3.tweets have “#TheWalkingDead”2 weeks

processing:

1.only language is English

2.only the first name of user in name list(census list)

3.only use part of the data to predict the sentiments

Sentiment Analysis

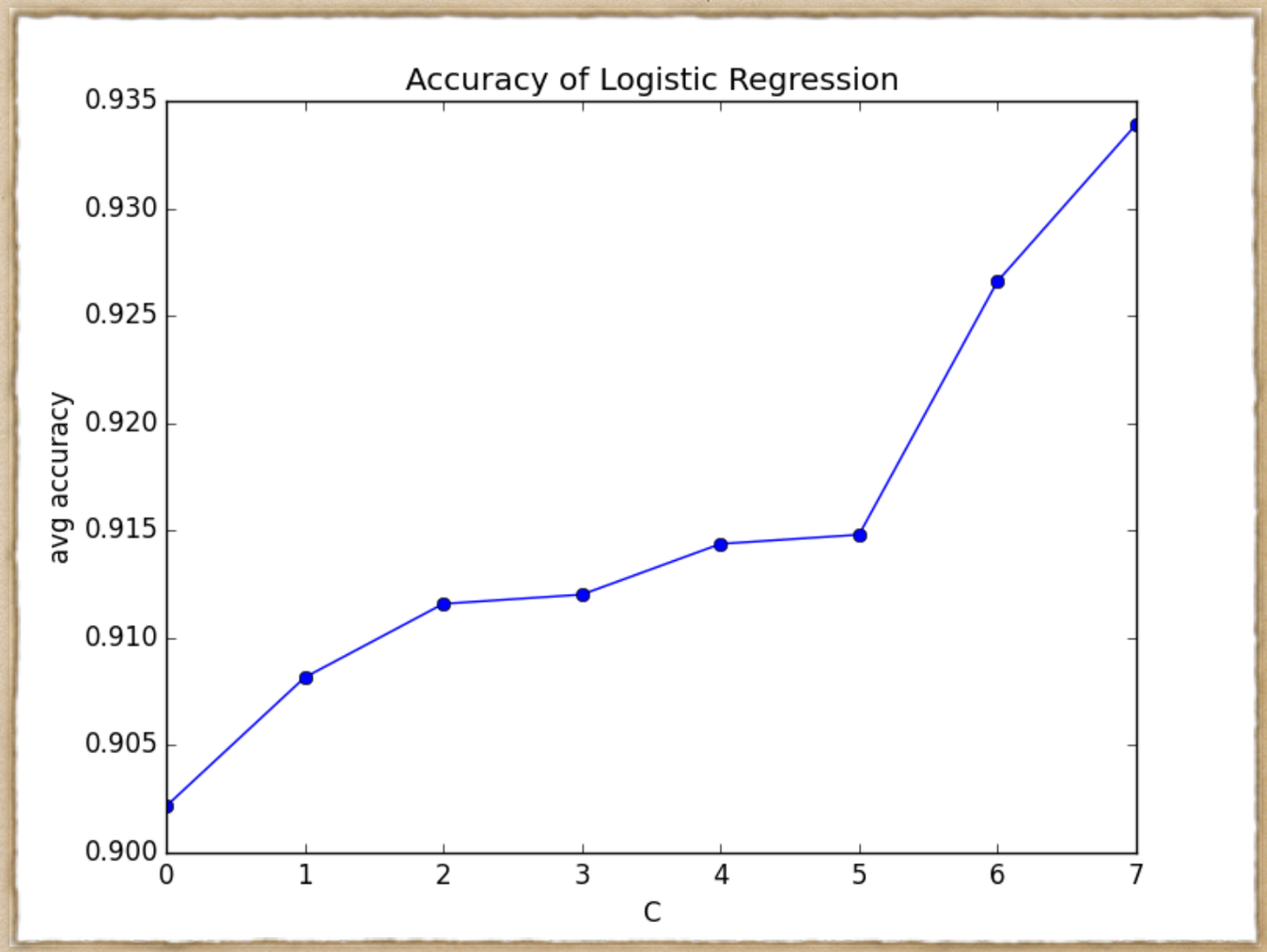
- distribution of positive and negative tweets
- accuracy

- 1. training data:
    - “the walking fans”——label:4 positive
    - “damn” + ”sucks”——label:2 negative

- Base Line:62.30%
- 2. 3 machine learning methods
  - 1) Logistic Regression
  - 2) SVC
  - 3) Gaussian Naive Bayes

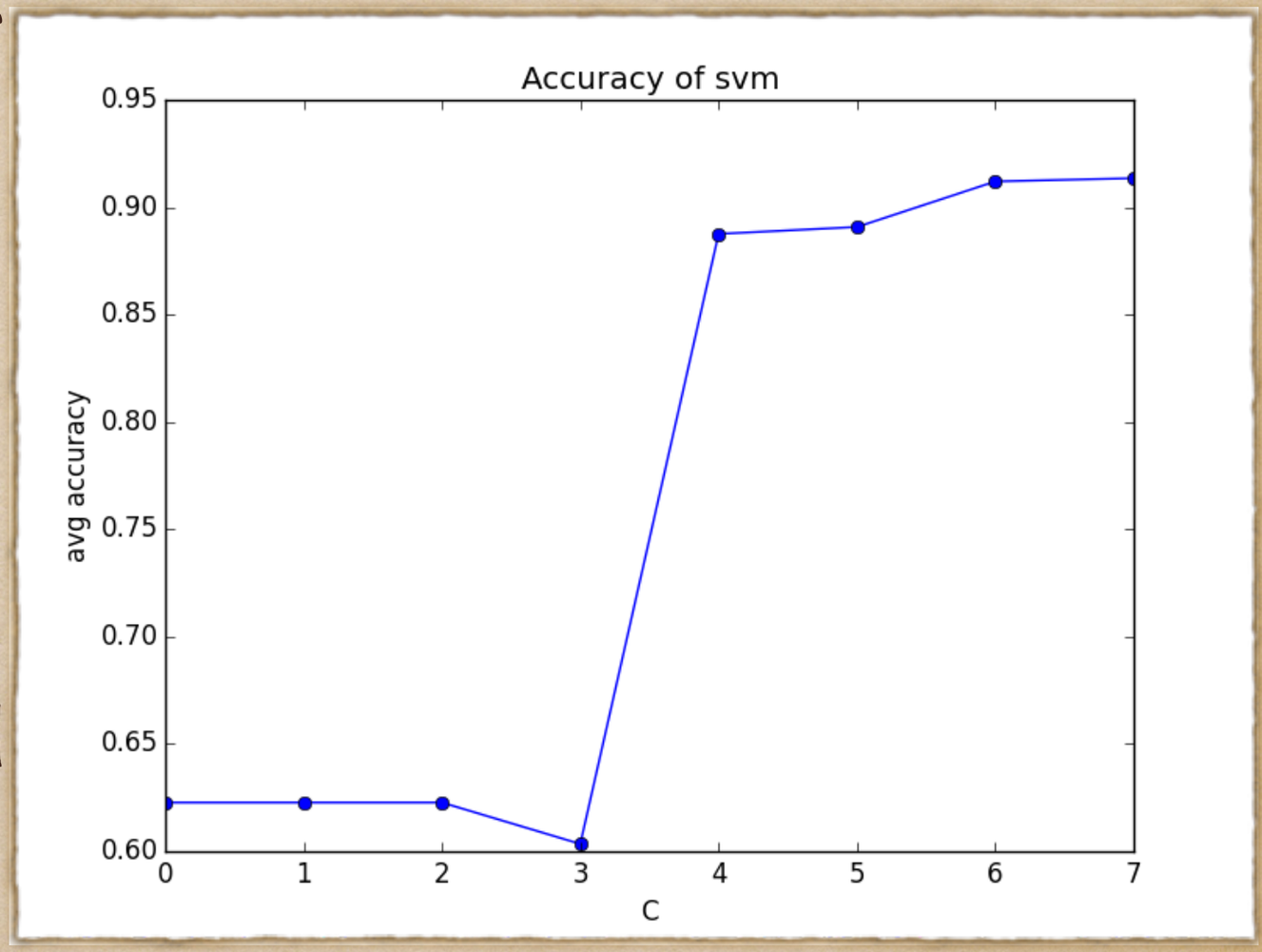
Logistic Regression

- C_var=[0.01,0.1,0.5,1.0,5.0,10.0,100.0,1000.0]
- Logistic regression is a linear model for classification rather than regression.Logistic regression is also known in the literature as logit regression, maximum-entropy classification or the log-linear classifier.



SVC

- In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

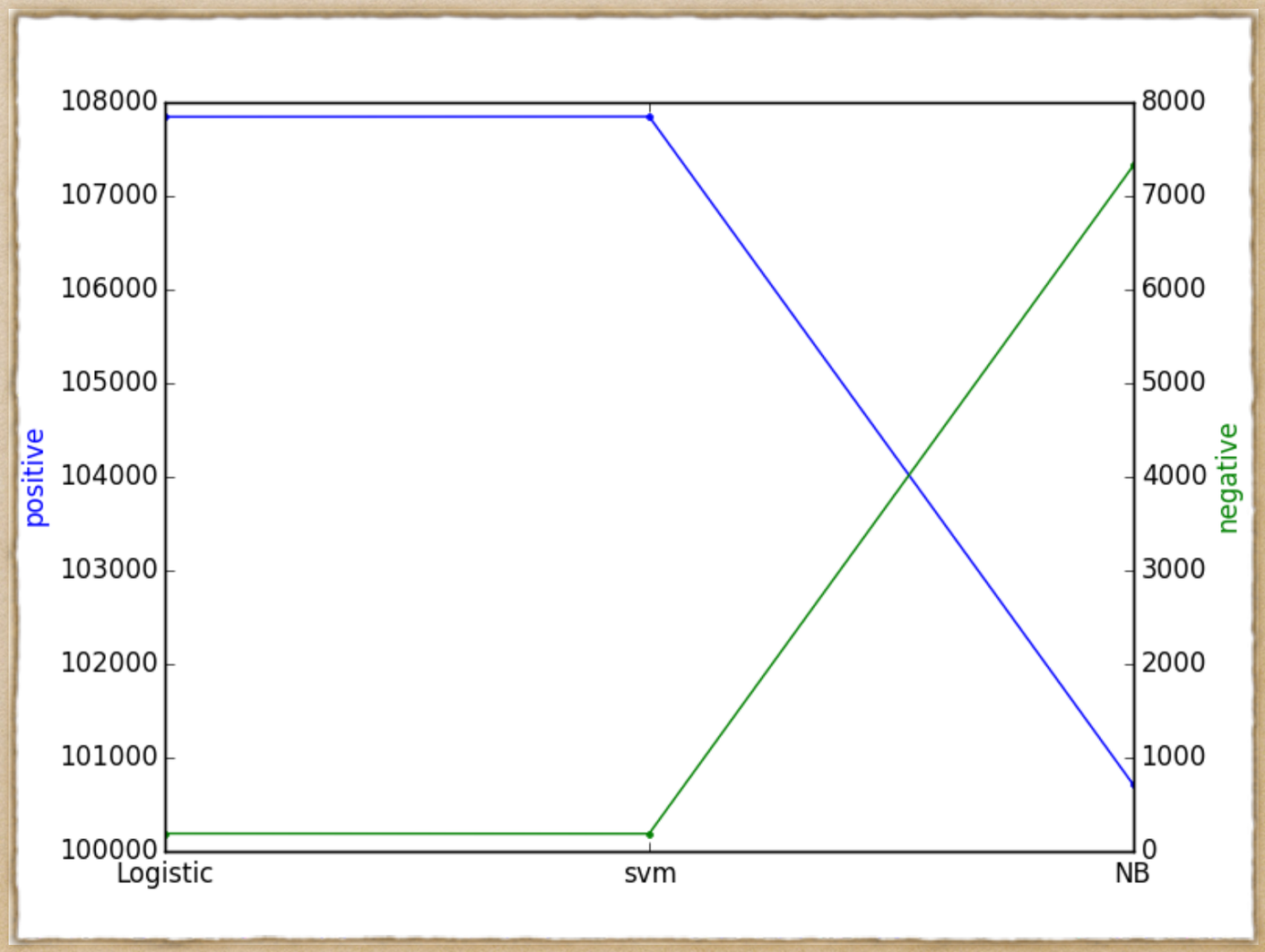


Gaussian Naive Bayes

- GaussianNB implements the Gaussian Naive Bayes algorithm for classification.

- The accuracy is 69.22%

- 3.predict the label
  - only use about 33% of the tweets “#TheWalkingDead”
  - different label from different method

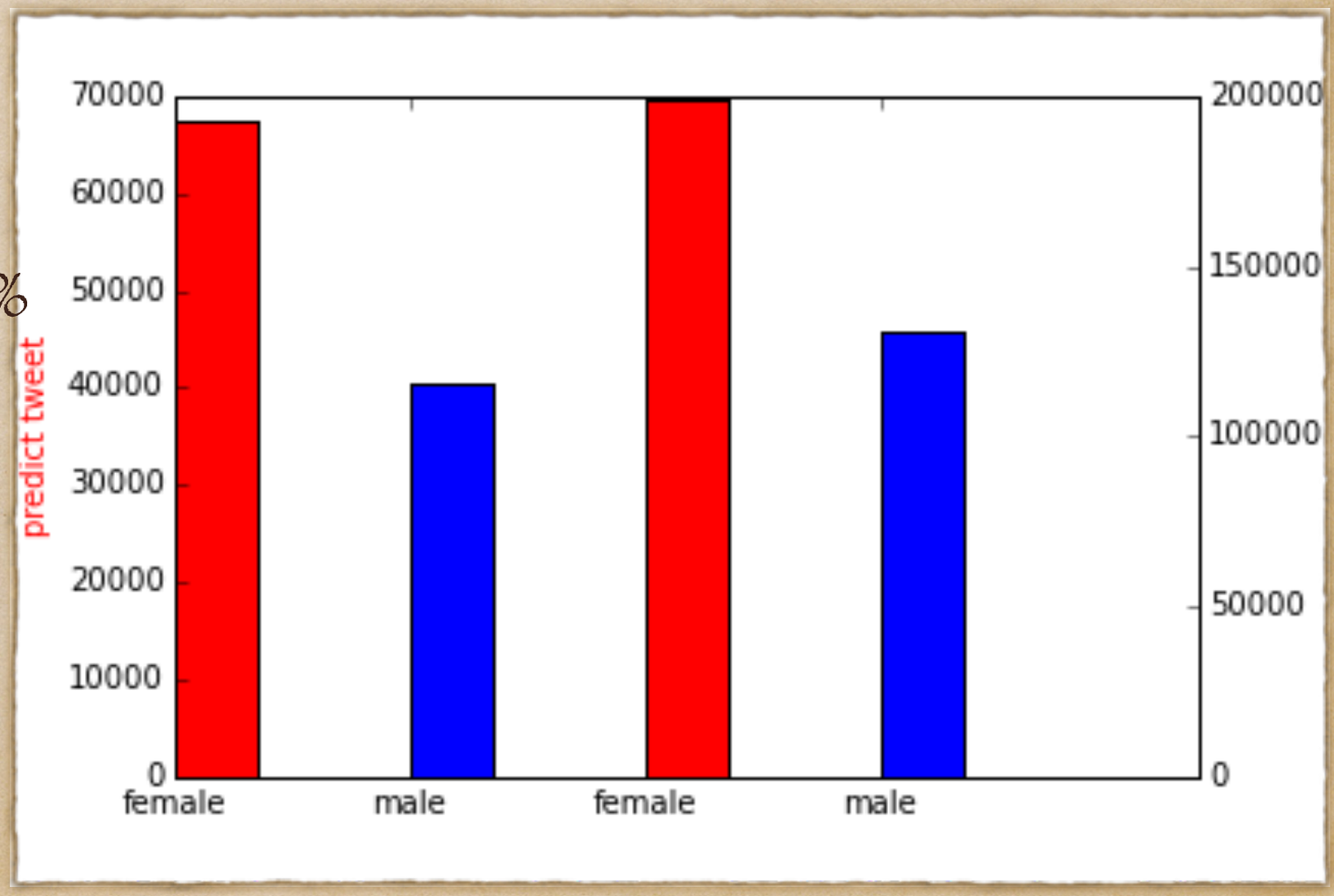


Gender Prediction

- distribution of male and female users
- combine sentiments and gender prediction
- distribution of sentiments with different genders

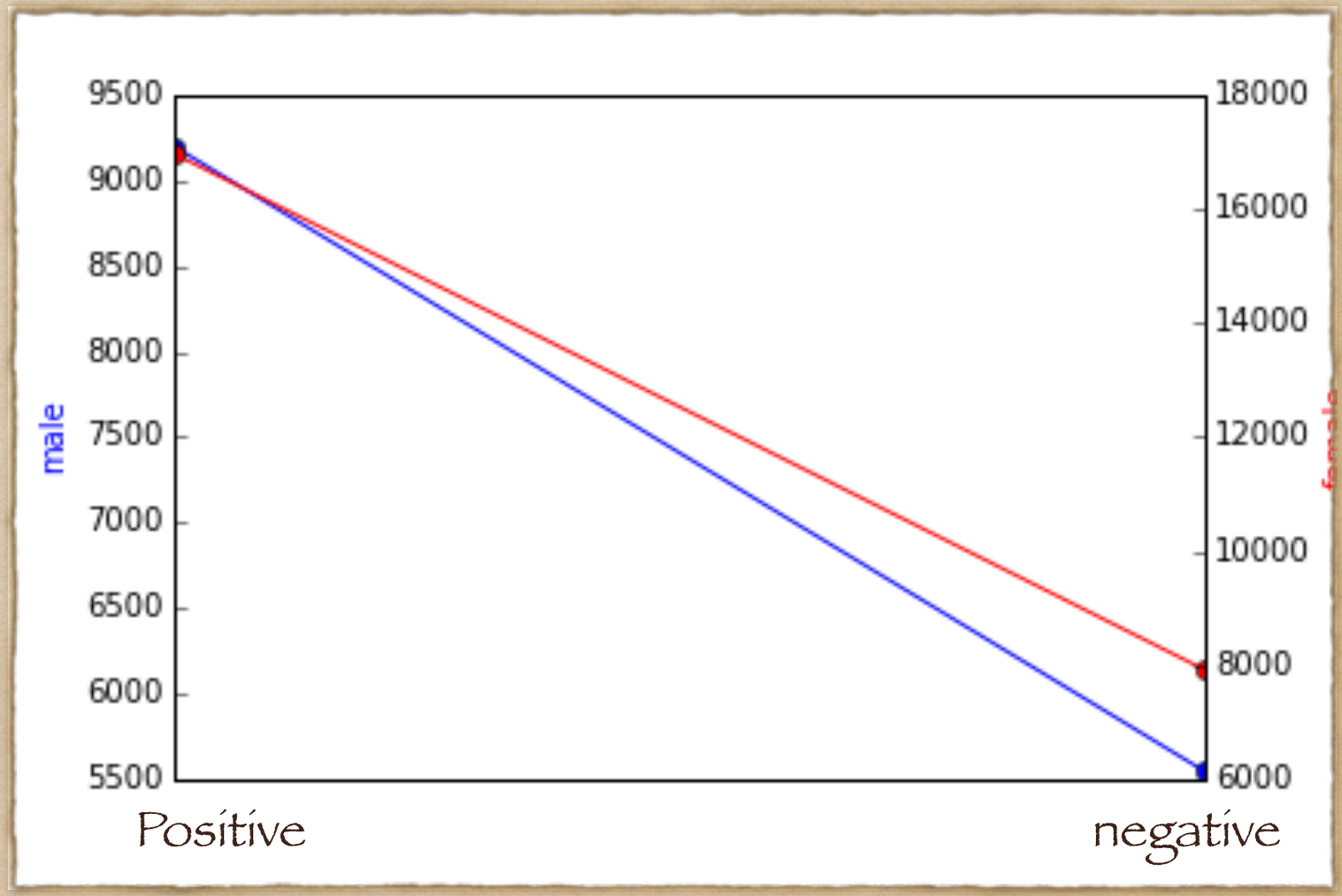
Distribution

- Census list+Afinn
- only predict tweets:
- male/female=60.11%
- All tweets:
- male/female=65.86%



Distrubution

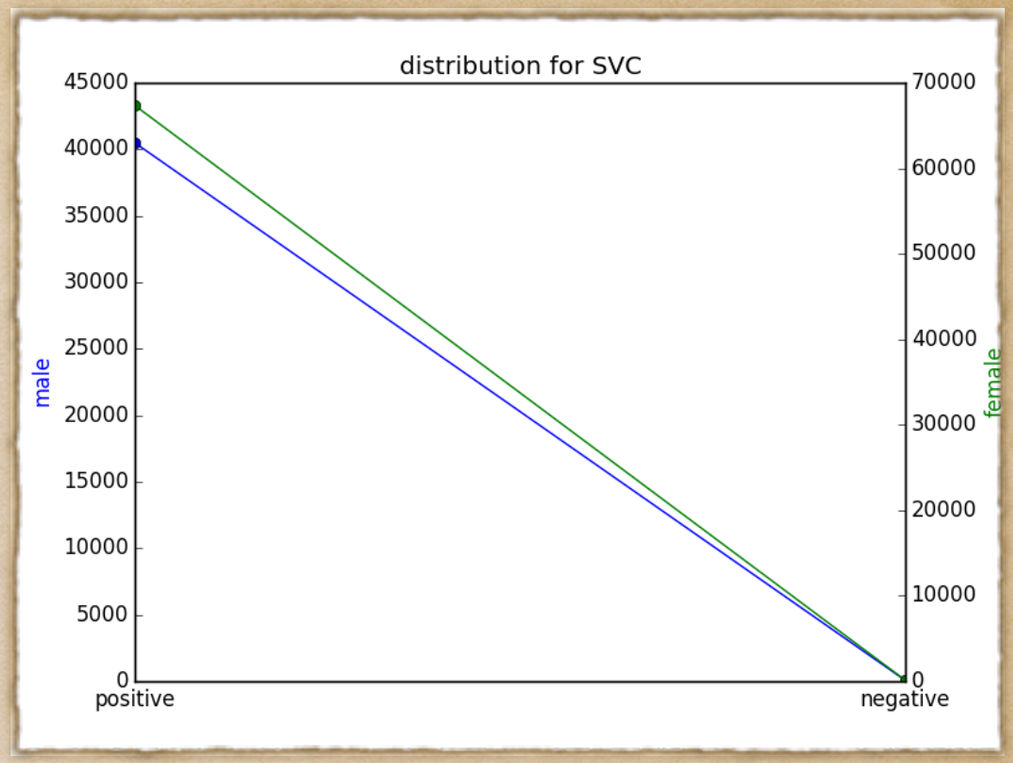
- predict tweets:
- male:
- positive:62.37%
- negative:37.63%
- female:
- positive: 68.16%
- negative: 31.84%

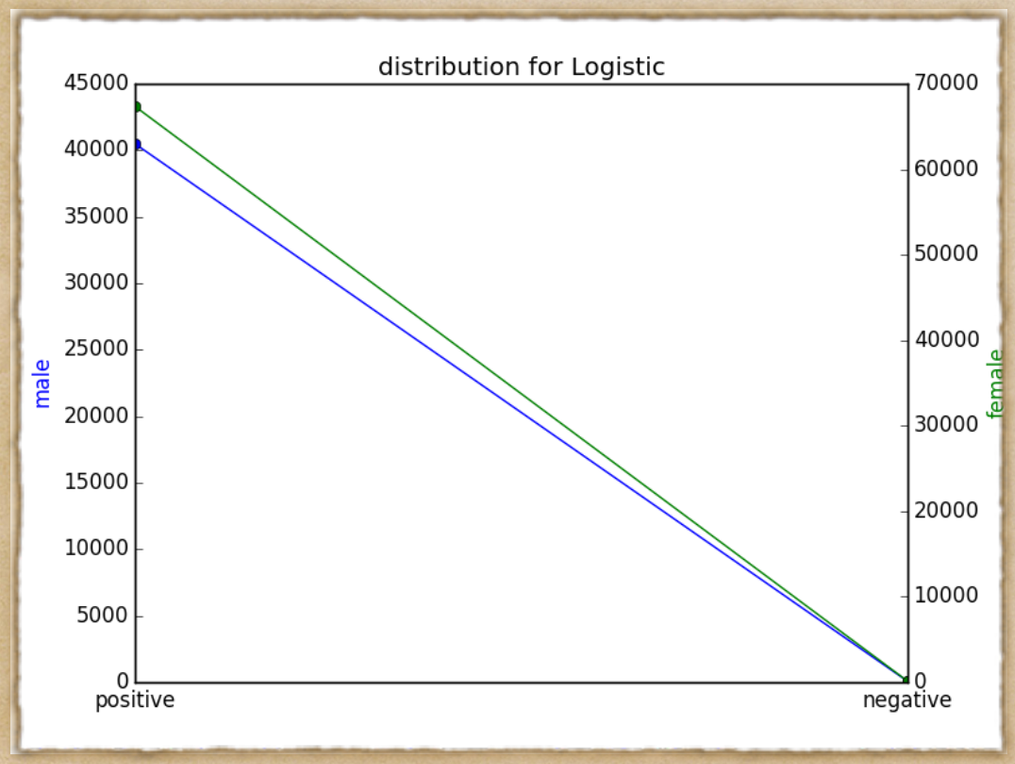


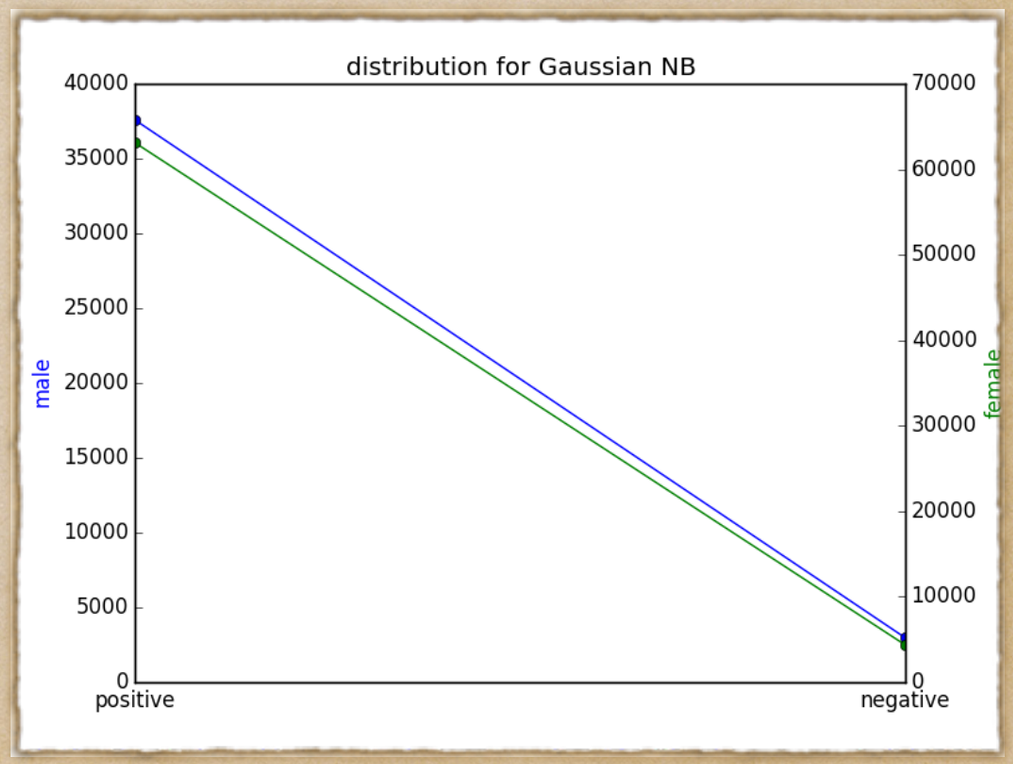
Distribution

- All tweets:
- male:
- positive: 53.20%
- negative: 46.80%
- female:
- positive: 56.74%
- negative: 43.25%

Same Trend!


distribution for SVC
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
male
70000
60000
50000
40000
30000
20000
10000
0
female
positive
negative


distribution for Logistic
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
male
70000
60000
50000
40000
30000
20000
10000
0
female
positive
negative


distribution for Gaussian NB
40000
35000
30000
25000
20000
15000
10000
5000
0
male
70000
60000
50000
40000
30000
20000
10000
0
female
positive
negative

Gender prediction

- Use different tokens to compare accuracies
- Use different parameters to compare accuracies:
  - C
  - min_df
  - maxdf
  - ngram

Future Work

- 1.Sentiment: neutral
- 2.collect more training data
- 3.compare accuracy between machine learning methods and census list+ afinn sentiments
- 4.Use more methods to compare

Question?