Natural Language Processing & Ideology

A mechanical inference on the pathology of words

I have 9.5 minutes of your time.

This talk will be:

10% of Modern Philosophy

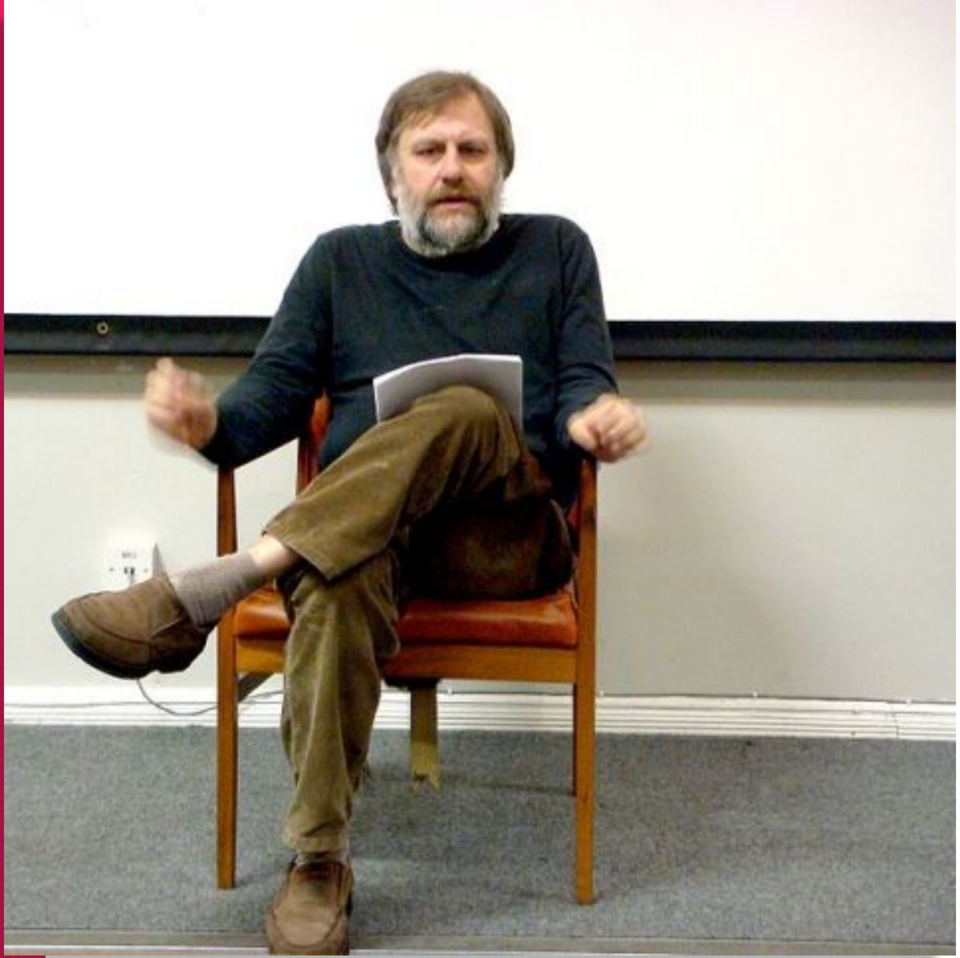
30% Problem statement and EDA strategies and challenges

40% break down of Modeling and Challenges

20% to connect my conclusions

Meet Slavoj Zizek

Slavoj Žižek is a Slovenian philosopher, cultural critic, and Lacanian psychoanalyst.
And his magnum opus is...



THE *SUBLIME* **OBJECT** OF *IDEOLOGY*.....

The Sublime Object of Ideology

The Sublime - from Kant

the sublime as horrible; filling the mind with agitation, until it abandons sensibility altogether

The Object - from Hegel

One's self-consciousness, in fact, will be dependent on one's recognition of those others as similarly recognizing oneself as a self-conscious subject to the broader object.

Ideology - from Marx

Ideology functions as the superstructure of a civilization: the conventions and culture that make up the dominant ideas of a society

WHAT DOES THAT MEAN TO ME AMIR?

does written text inherently hold a relationship to something intrinsically “ideological” or ideology and very much so do we get trained on data very much like machines?

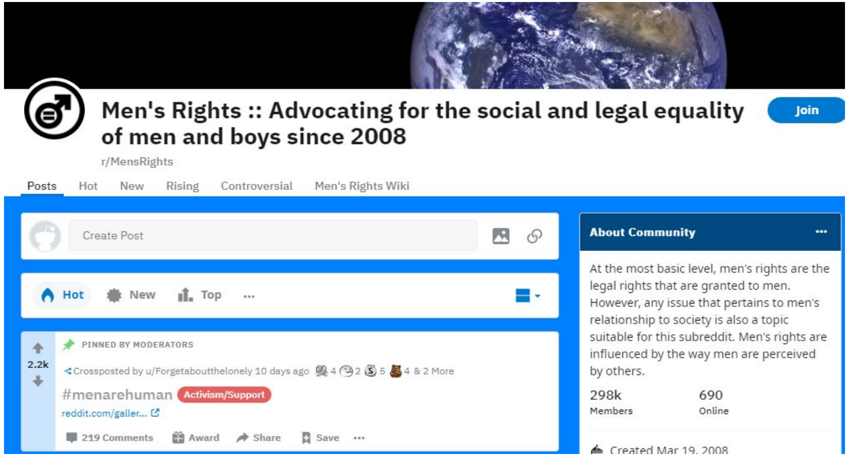
Problem Statement:

Can Natural Language Processing with help of SKLearn distinguish between two ideologically oppositional subreddits?

Two ideologically opposed, adversarial and extremely parallel subreddits

r/MensRights

r/feminism



Men's Rights :: Advocating for the social and legal equality of men and boys since 2008

r/MensRights

PINNED BY MODERATORS

Crossposted by u/Forgetaboutthelonely 10 days ago 4 2 5 4 & 2 More

#menarehuman Activism/Support

reddit.com/galler...

219 Comments

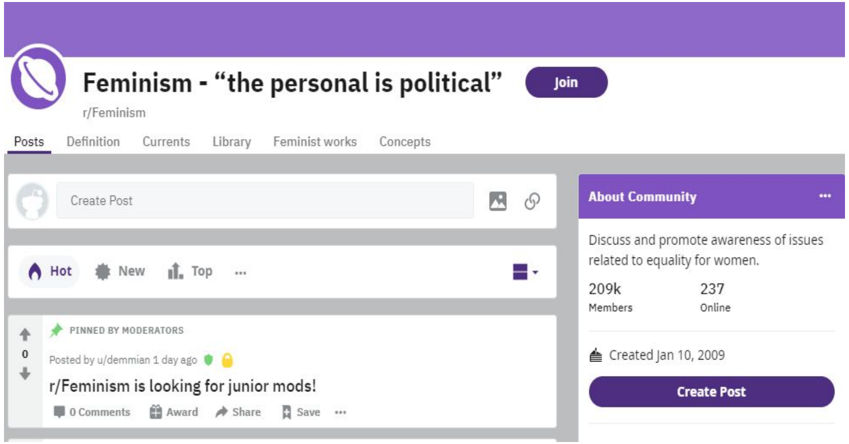
About Community

At the most basic level, men's rights are the legal rights that are granted to men. However, any issue that pertains to men's relationship to society is also a topic suitable for this subreddit. Men's rights are influenced by the way men are perceived by others.

298k Members

690 Online

Created Mar 19, 2008



Feminism - “the personal is political”

r/Feminism

PINNED BY MODERATORS

Posted by u/demmian 1 day ago

r/Feminism is looking for junior mods!

0 Comments

About Community

Discuss and promote awareness of issues related to equality for women.

209k Members

237 Online

Created Jan 10, 2009

We don't care what they think as long as they express it.

We only want to know can our models accurately distinguish them. I have no dog in this fight and my purposes are purely scientific and ontological.

So they will be known as GROUP A= Feminism and GROUP B= Mensrights

The Data

We combined **19,891** rows from Group A and **6,885** Group B that is composed of both submissions and also comments from the Pushshift Reddit API.

They were uniformly organized with the origin, text and timestamp.

The biggest challenge was having a large enough sample. Mostly due to the fact that there were some issues within the data we acquired that would be counterproductive to both NLP and statistical modeling.

Why did you pull so much stuff?

This is due to the fact that reddit post have embedded markup tags that are very difficult to get out of the text. What comes through the API payload is not “raw text”.

[removed]

Controversial topics violate either the moderator’s guidelines or they violate Reddits guidelines when they are removed this particular tag is left. It was however a pain to get

ANSWER:

.[RegEx]*

Vast majority of EDA was done mostly with Regular expressions

1. Terminated Markup url title and links
2. Removed all special chars, punctuations and numbers.
3. Removed [Removed]
4. Removed [Deleted]
5. Removed /n, /t, /r in the text

Upon effectively cleaning the entire data model performance improved. Clean data clean results.

Lexicon analysis

Let’s look at a comparison of the 5 most popular words for Group A and Group B.

Group A

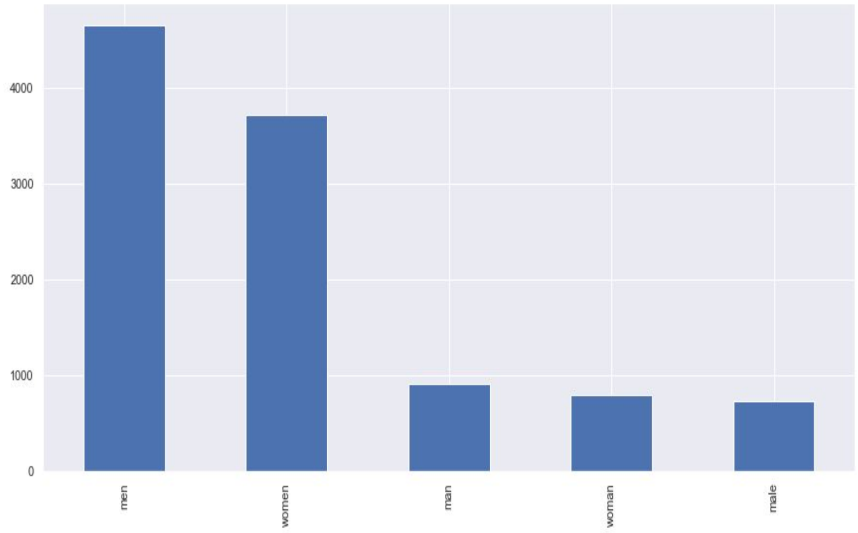
women	5915
men	5873
amp	2231
woman	1204
feminism	1107

Group B

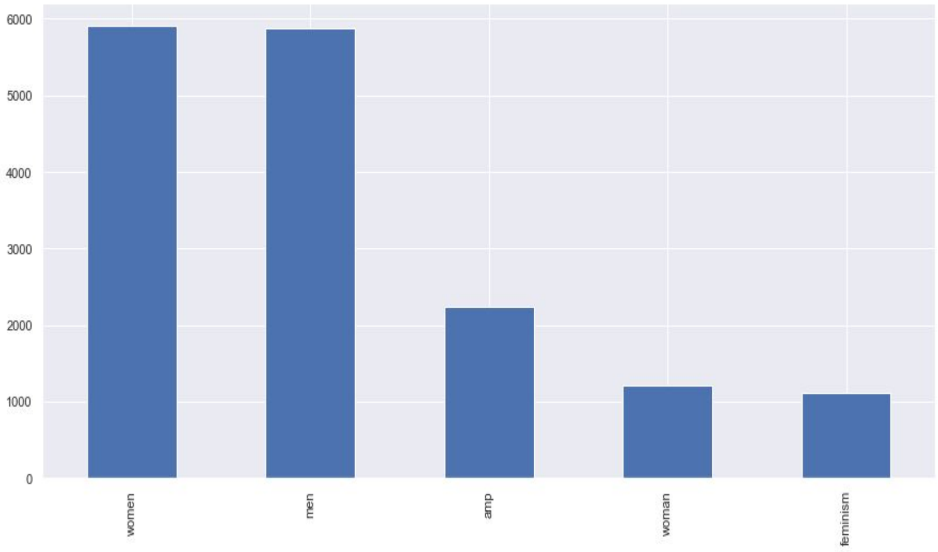
men	4651
women	3724
man	793
woman	793
male	721

MAN, WOMAN where present in both only one actually mentioned itself by a ideological name that was group A

Group A



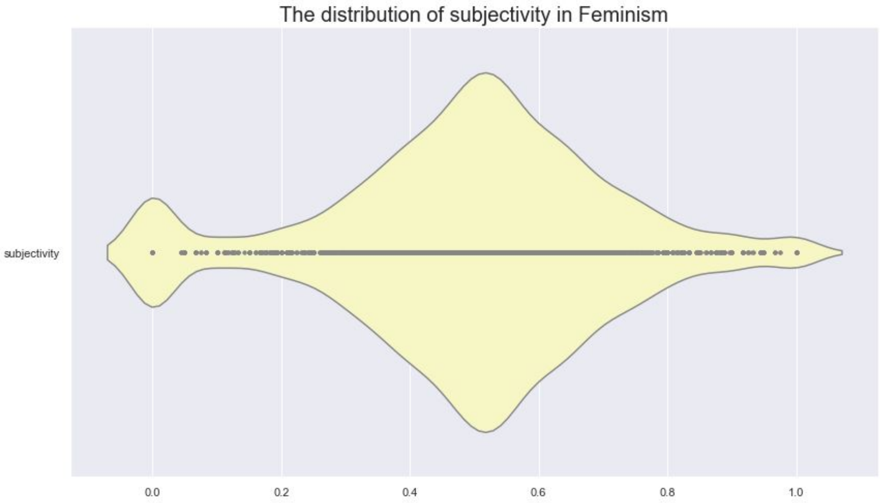
Group B

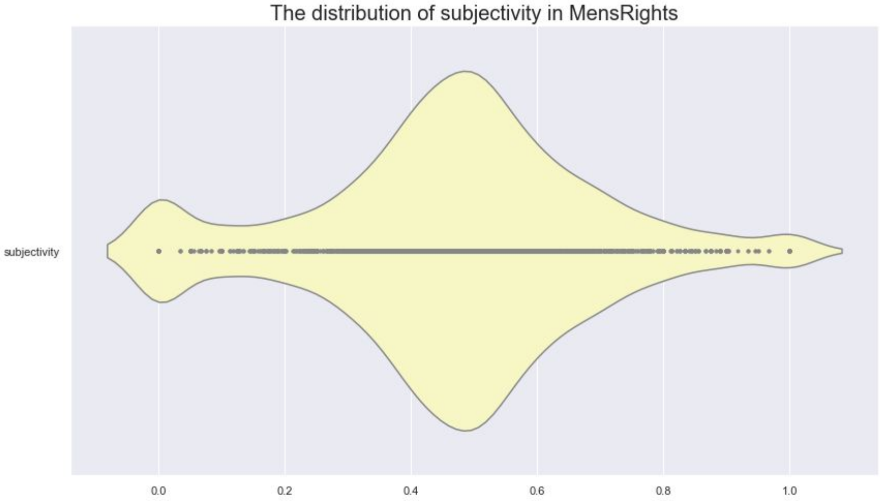


Sentiment Analysis

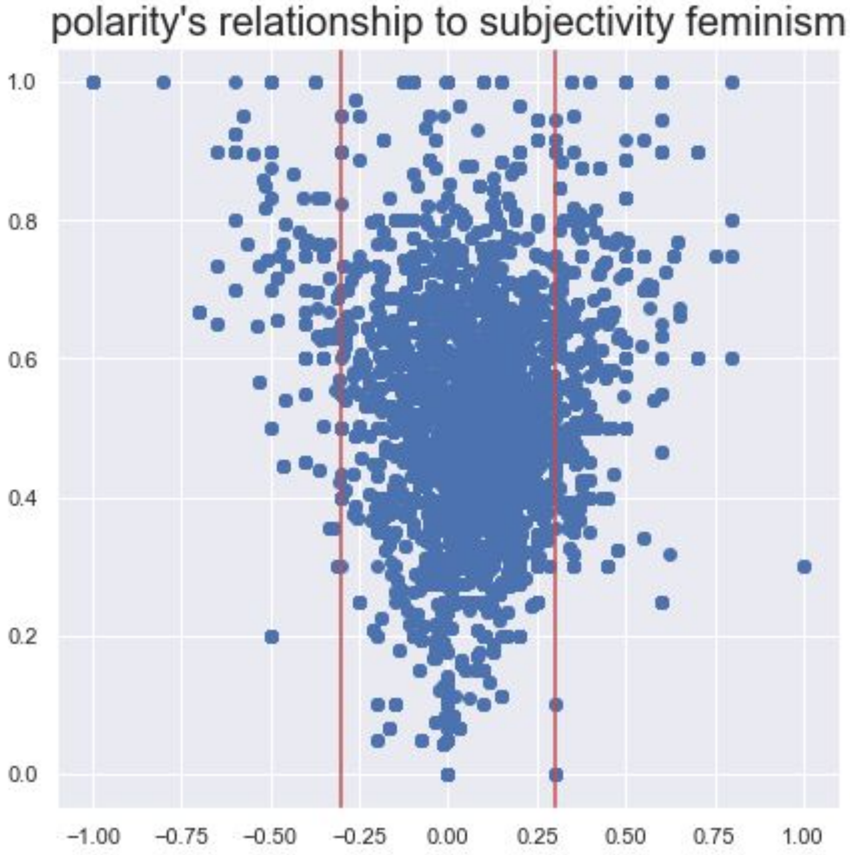
TextBlob

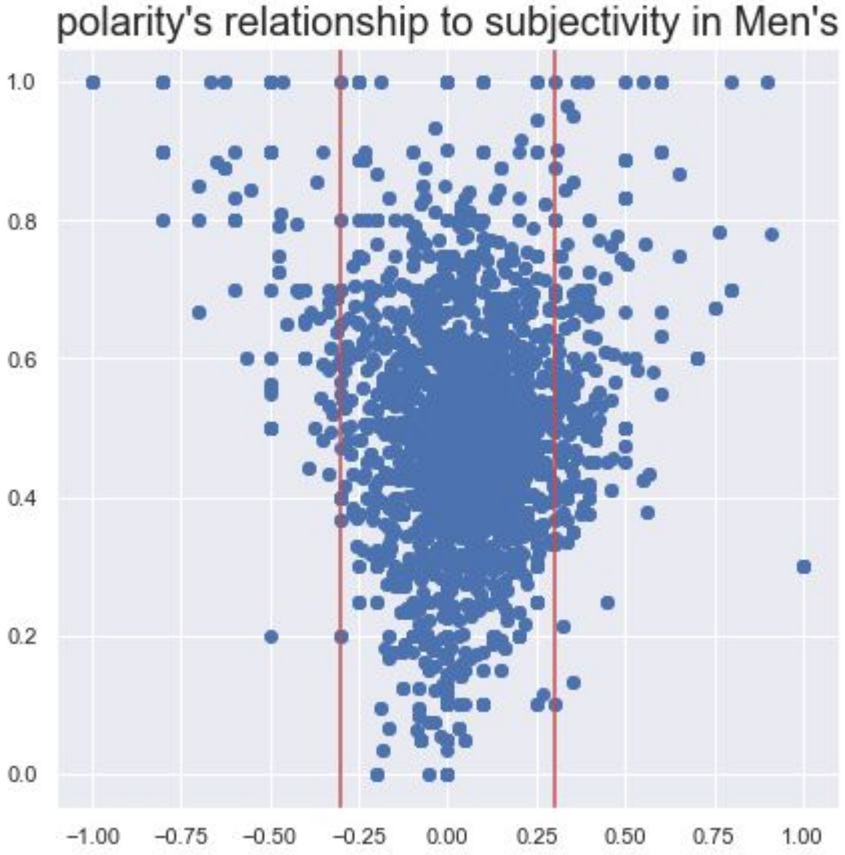
Textblob subjectivity analysis for both shows an ALMOST identical violin plot.The numerical average is 0.489847 for Group A and 0.463823 for Group B.





Comparing the relationships between sentiments





So can my models find mensrights(group b)?

Model Overview

SVM

Black box model with two main parameters. Right from the start took more than hour to run gridsearch and about 45 minutes with stringent EDA I have it down to 25-30.

Logistic Regression

Logistic regression was easier to implement, interpret, and very efficient time wise.

Data overview

Group A:

8551 Rows after eda

Initial Concern:

that having more data would skew the results somehow in a way where the group. Also having so many removed and deleted submissions and comments that somehow that would not give an accurate training data

Group B:

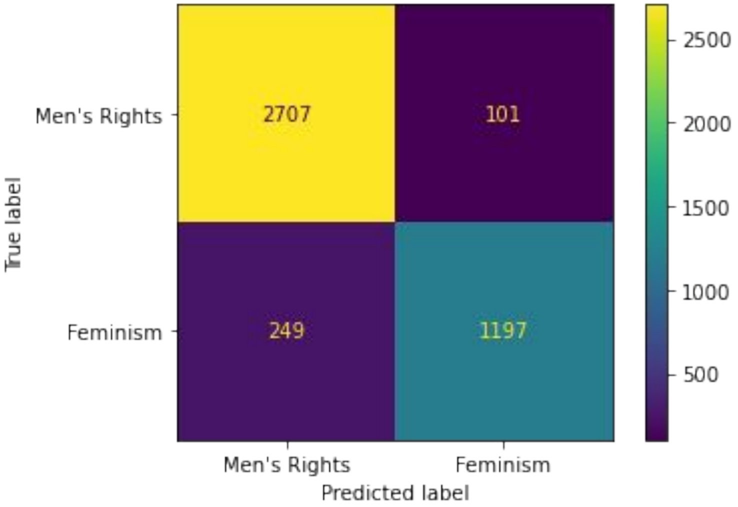
4331 Rows after eda

Initial Concern:

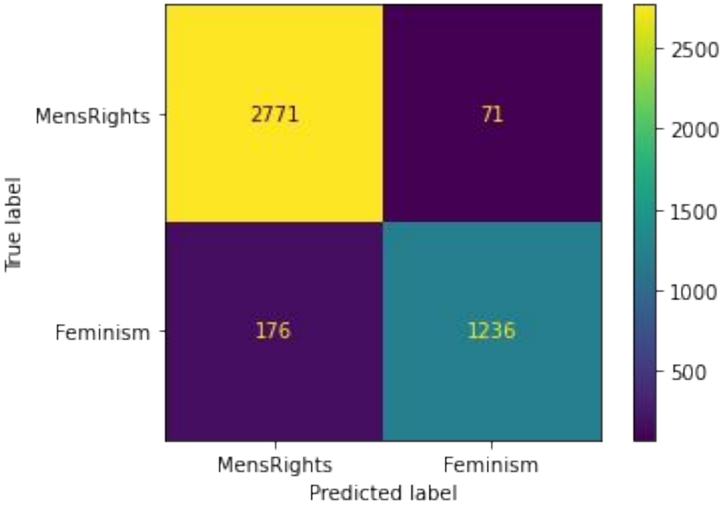
Somehow that the eda robbed the text of its ability to somehow be distinguishable.

Comparing Confusion Matrices

SVM



Logistic Regression

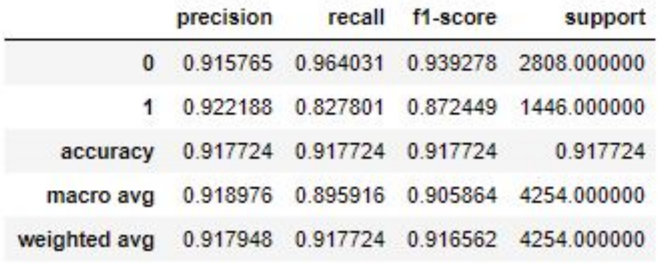


As it relates to the question asked two slides back we can effectively pick Group B which would be the TP(True positive).

The classification report and looking outside the TP

SVM

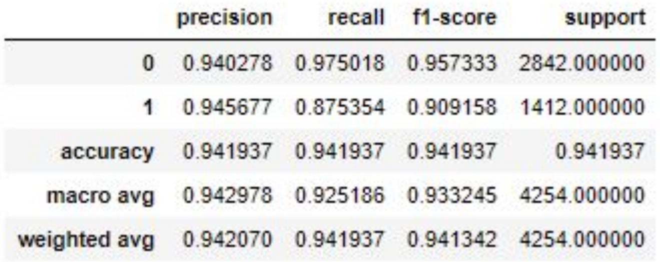
Did less better predicting the MensRights group and the True negative the Feminist subreddit.



	precision	recall	f1-score	support
0	0.915765	0.964031	0.939278	2808.000000
1	0.922188	0.827801	0.872449	1446.000000
accuracy	0.917724	0.917724	0.917724	0.917724
macro avg	0.918976	0.895916	0.905864	4254.000000
weighted avg	0.917948	0.917724	0.916562	4254.000000

Logistic Regression

The winner in this situation.



	precision	recall	f1-score	support
0	0.940278	0.975018	0.957333	2842.000000
1	0.945677	0.875354	0.909158	1412.000000
accuracy	0.941937	0.941937	0.941937	0.941937
macro avg	0.942978	0.925186	0.933245	4254.000000
weighted avg	0.942070	0.941937	0.941342	4254.000000

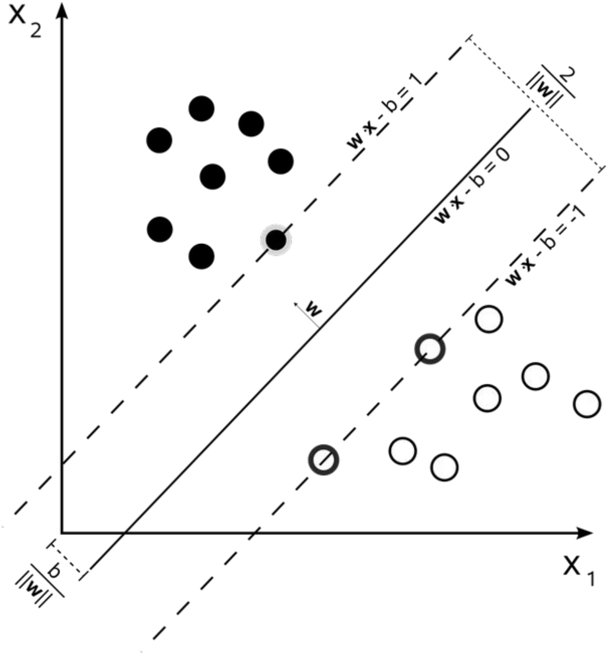
Why? Did SVM not perform better in this case:

Here is what I think might be why:

“postulate a probability distribution and thus does not assume any randomness. It merely tries to draw a simple line(or plane or hyperplane in higher dimensions) to separate the data points into two parts.” Well maybe it had a harder time dealing with the harder to distinguish cases.

Daeyoung Lim, “Self-proclaimed Bayesian”

[What is the difference between Support Vector Machine and Support Vector Regression?]


X_2
X_1
$\mathbf{w}\cdot\mathbf{x}-b=1$
$\mathbf{w}\cdot\mathbf{x}-b=0$
$\mathbf{w}\cdot\mathbf{x}-b=-1$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$
$\mathbf{w}$

Conclusion

Did we gain any insight to text without full context delivering to a machine an ideological paradigm.

1. **So can my models find mensrights(group b)?** Despite similarities in sentiment, similar words or overall focus on the same subject a machine can and is able to distinguish them

2. **Can Natural Language Processing with help of SKLearn distinguish between two ideologically oppositional subreddits?** Yes, they can.

This means that there is to a lesser or greater degree ideology embedded in language.