A horizontal brushstroke in the colors of the rainbow (red, orange, yellow, green, blue, purple) serves as the background for the title text.

Correlations and Differences Between Lesbians and Gay Men on Reddit



Problem:




Lesbian divorce rates
are twice as high as
gay divorce rates






Can We Find Significant
Differences in Language
Used in Gay/Lesbian
Subreddit?

 reddit 

USE APP   



An inclusive place for gay men to share their lives and experiences.

r/GayMen

An inclusive place for gay men to share information and discuss issues that relate to their lives & ...
[More](#)

11.7K members • 20 online



Join Community




[View this community's rules](#)


Purpose and Scope

This is an inclusive place for gay men to share information and discuss issues that relate to their lives & experiences of being a gay man.

Whether you're a bro, gent, teen, elder, butch, sissy, or

 reddit 

USE APP   



Actual Lesbians!

r/actuallesbians

A place for discussions for and by cis and trans lesbians, bisexual girls, chicks who like chicks, b...
[More](#)

220K members • 790 online

Join Community

[View this community's rules](#)

1. [Join Our Discord Server!](#)
2. [How to Handle Trolls](#)
3. [Catfish Tracker](#)

Welcome to the sub, please read our [rules](#).




Methods

- Pushshift API
- Two sets of 500 submissions per group for a total of 2000
- Combine to one dataframe
- GayMen: 1, actuallesbians: 0
- Combine title, self text to one column



Methods


- Stopwords adjusted to include pronouns, other common words in community
- Train/Test/Split (.33)
- CVEC/LogReg
- TVEC/LogReg
- TVEC/ Naive Bayes Multinomial



Results

CVEC/Log Regression

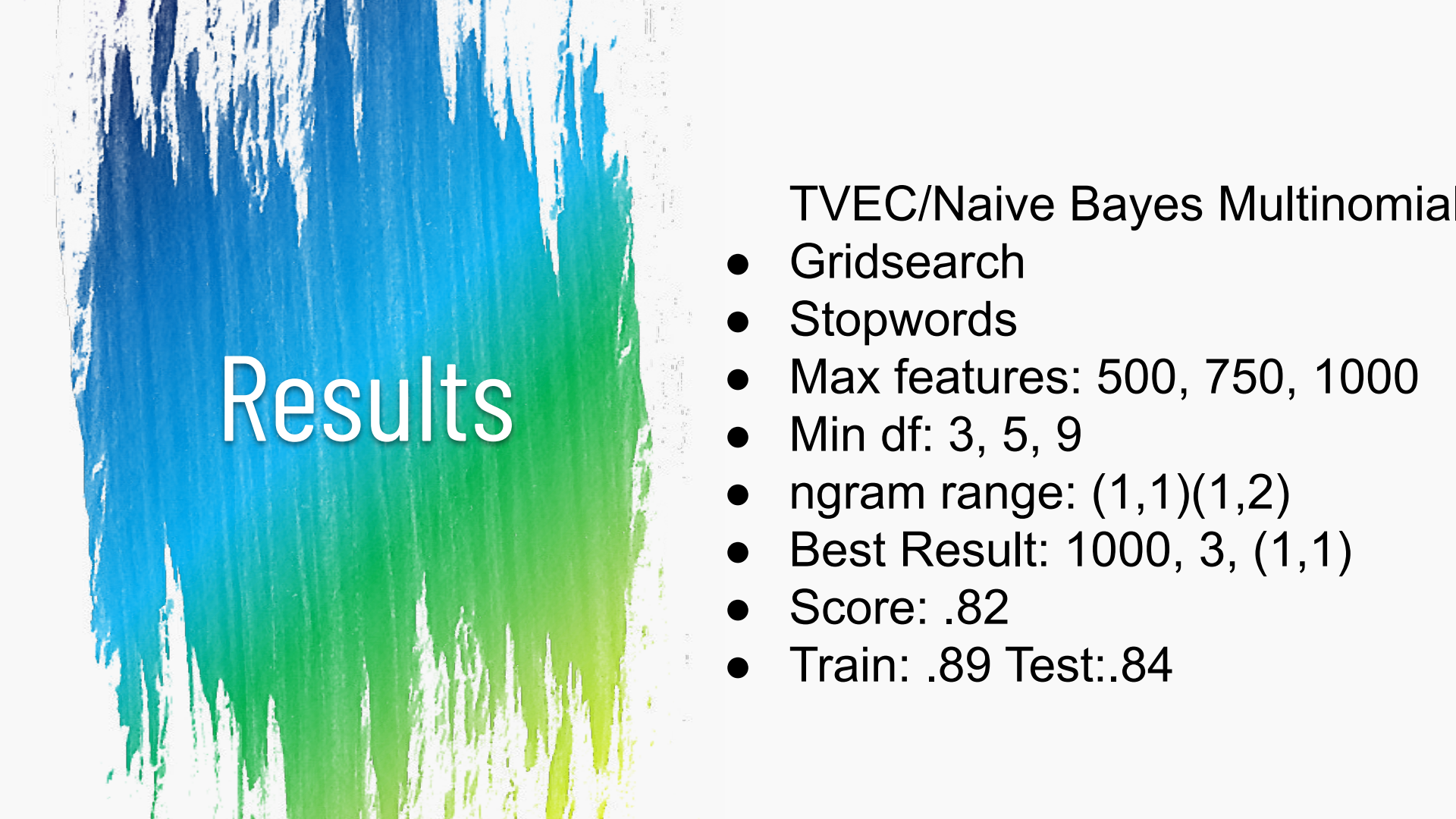
- Gridsearch
- Stopwords
- Max features: 500, 750, 1000
- Min df: 3, 5, 9
- ngram range: (1,1)(1,2)
- Best Result: 1000, 3, (1,1)
- Score: .82
- Train: .94 Test:.82



Results

TVEC/Log Regression

- Gridsearch
- Stopwords
- Max features: 500, 750, 1000
- Min df: 3, 5, 9
- ngram range: (1,1)(1,2)
- Best Result: 1000, 3, (1,1)
- Score: .82
- Train: .92 Test:.85



Results

TVEC/Naive Bayes Multinomial

- Gridsearch
- Stopwords
- Max features: 500, 750, 1000
- Min df: 3, 5, 9
- ngram range: (1,1)(1,2)
- Best Result: 1000, 3, (1,1)
- Score: .82
- Train: .89 Test:.84





Conclusion

- We can find significant differences in gay/lesbian subreddit text
- Comparing gay/lesbian Reddit language is not the best way to diagnose problems within the lesbian community
- Better method: conduct surveys with lesbians in marriages that have ended and marriages that have lasted over x years and find trends.
- Don't get married?