NLP- reddit subset prediction

Data Scientist: Roohullah Mansoor

April 4, 2019

Why I chose these subsets

Ask Reddit	Love
2499	2476
539 [removed]	446 [removed]

Our new DataFrame 4975 columns with 985 [removed]

Steps & Why ?

- Using API, get the data
- Clear all [removed]
- Check NaN values
- Removed all useless columns for this project
- Combined title and subset columns
- Dummified y column

The hypothesis (or prediction)

What do you think will happen?

Exploring Data

Love:

2476 rows, 9 columns

446 [removed]

Askreddit:

2499 rows, 9 columns

539 [removed]

Asklove: 4975 rows, 9 columns

985 [removed]

Tokenize | CV & TFIDF | Model | Lr & nb

Overfit,
CV Lr Train Score :0.983
CV Lr Test Score :0.938

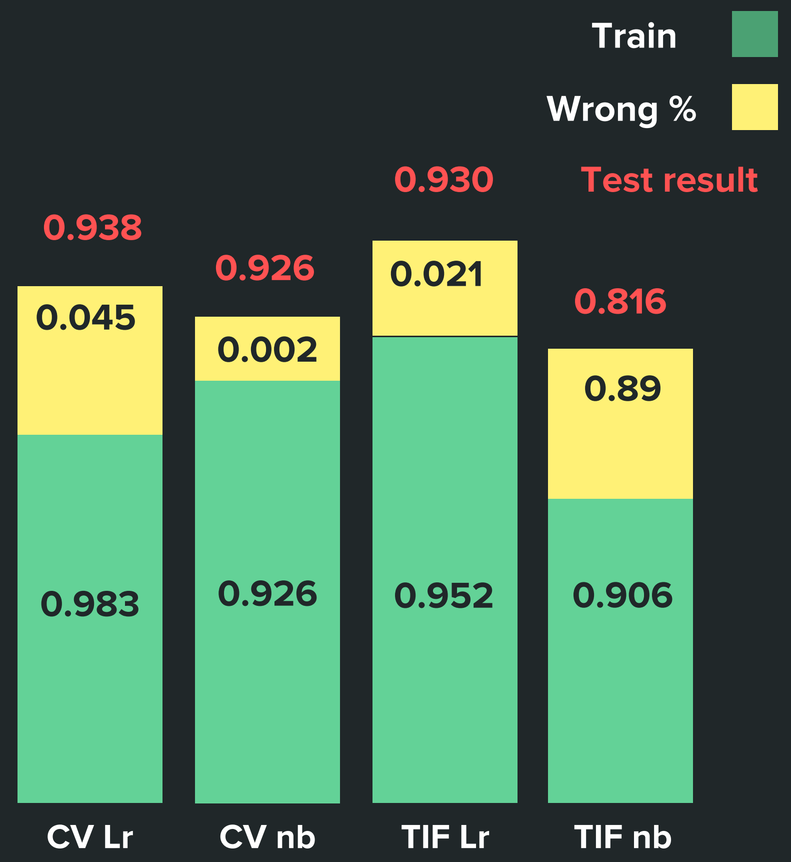
Good fit,
CV nb Train Score :0.926
CV nb Test Score :0.924

Overfit,
TIFDF Lr Train Score :0.952
TIFDF Lr Test Score :0.930

Overfit,
TIFDF nb Train Score :0.906
TIFDF nb Test Score :0.816

Results

Bar chart shows the result for all models and performance of tokenizers

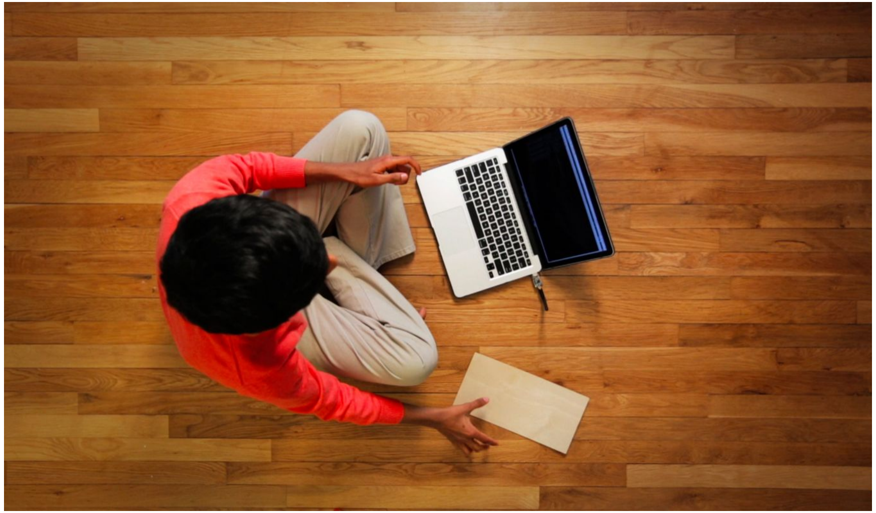


Aha!

My discoveries

What did you learn after testing?

1. CV nb did good job very less error
2. TIFDF nb did the worse job
3. Lr did similar job on both TIFDF and CV



Questions?