EXPLORING AMAZON FOOD REVIEW

Priyank Pandya

Big Data

CPT_S_415

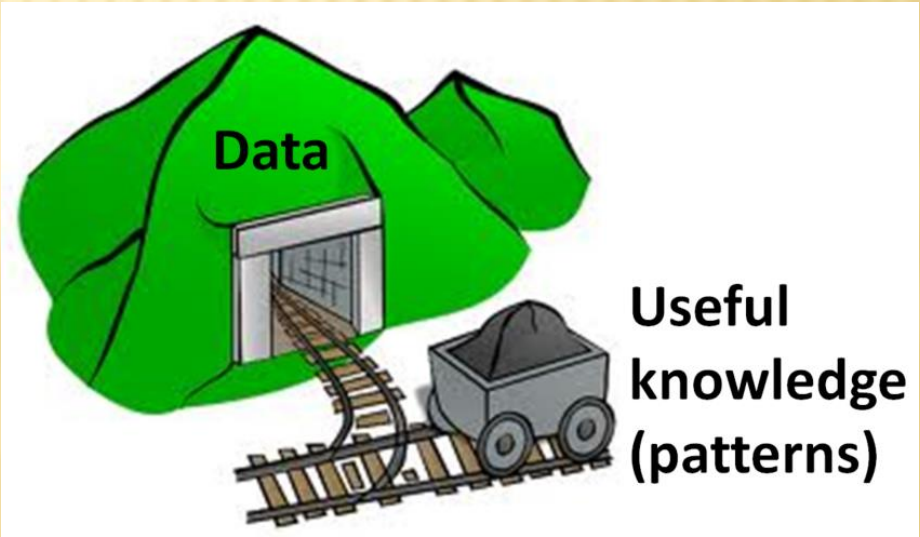
10 December 2018

ABOUT

- Founded on 1994
- Vision is to be earth's most customer-centric company; to build place where people can find and discover product they want

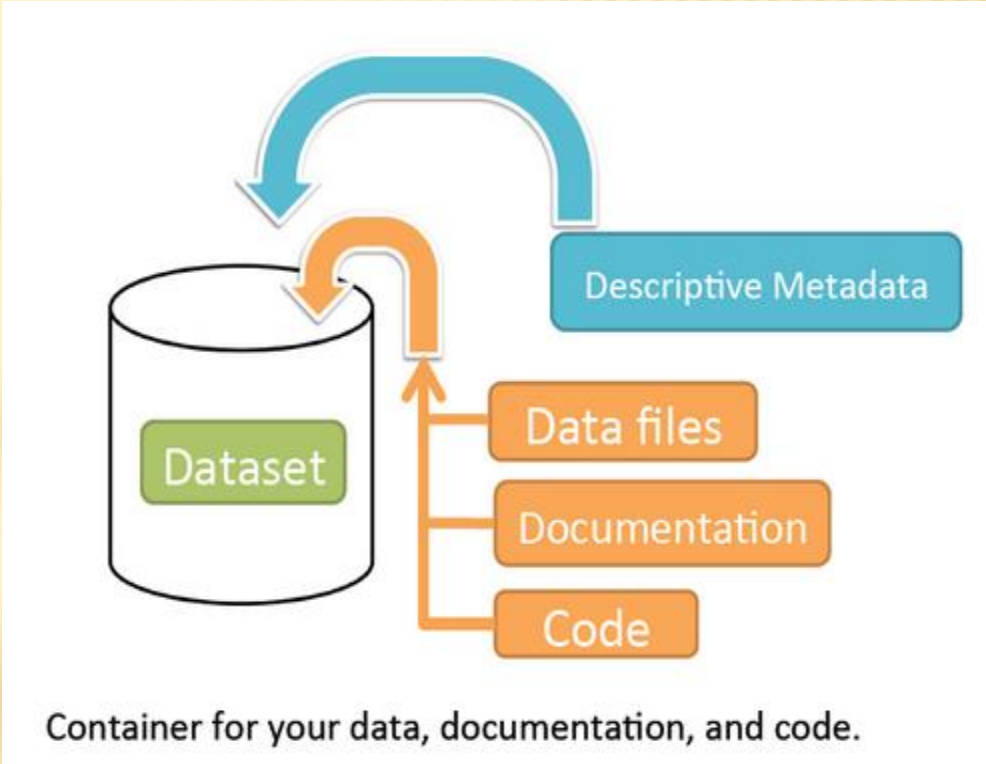
PROBLEM STATEMENT

- To turn food review of product into productive and efficient use for the business improvement
- Distinguish the review into positive and negative feedback
- Emotion
- Technical



DATASET [3]

- SNAP
- Kaggle
- Includes
  - Date and Time
  - User
  - Review
  - Product



METHODOLOGY

Data Manipulation

Text Mining
- ✓ Tokenization
- ✓ Stop Word
- ✓ Stemming

Modeling
- ✓ Classification

Deployment
- ✓ Spark
- ✓ AWS

DATA MANIPULATION

- Filtering
  - Remove irrelevant review
  - Scoring
    - Positive (=1)
    - Negative (=0)

DATA PRE-PROCESSING

"Seems like he can't talk to you without getting extremely rude."

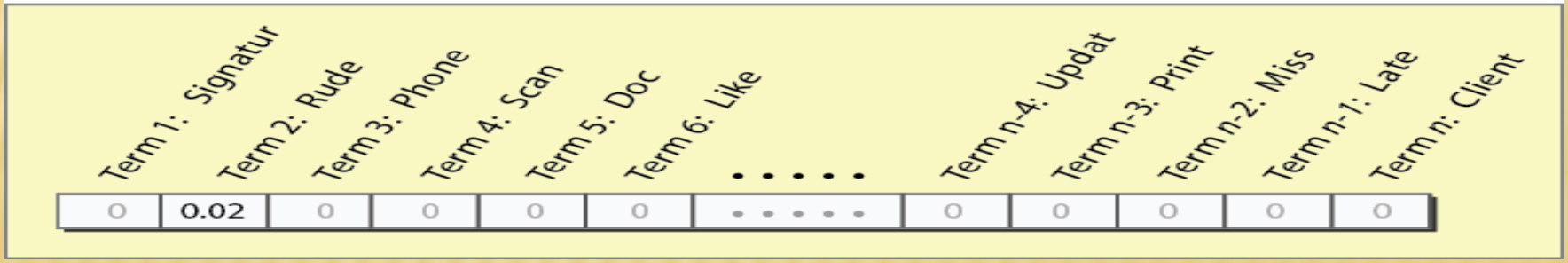
Remove Non-Alphanumeric Characters and Tokenize

[seems, like, he, can, t, talk, to, you, without, getting, extremely, rude]

Remove Stop Words and Stem

[seem, like, talk, without, get, extrem, rude]

Convert to TF-IDF Vector

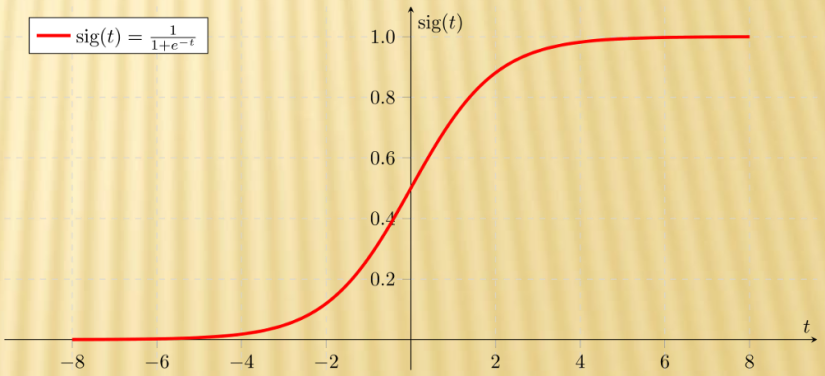


MODEL

- Used Spark

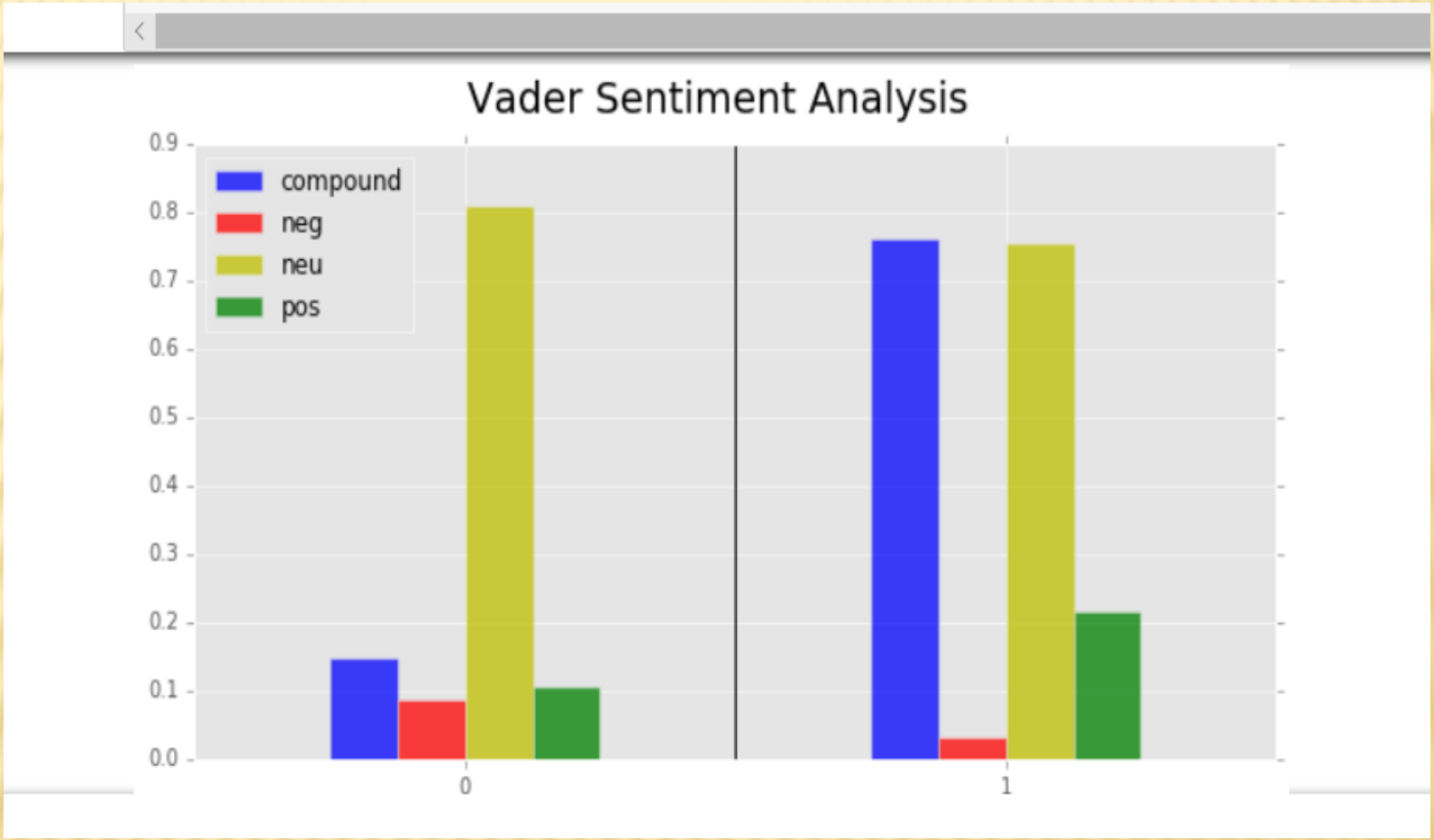
Method:

- Logistic Regression
- Gradient Boosting Tree
- Naïve Bayes



DEPLOYMENT

RESULT



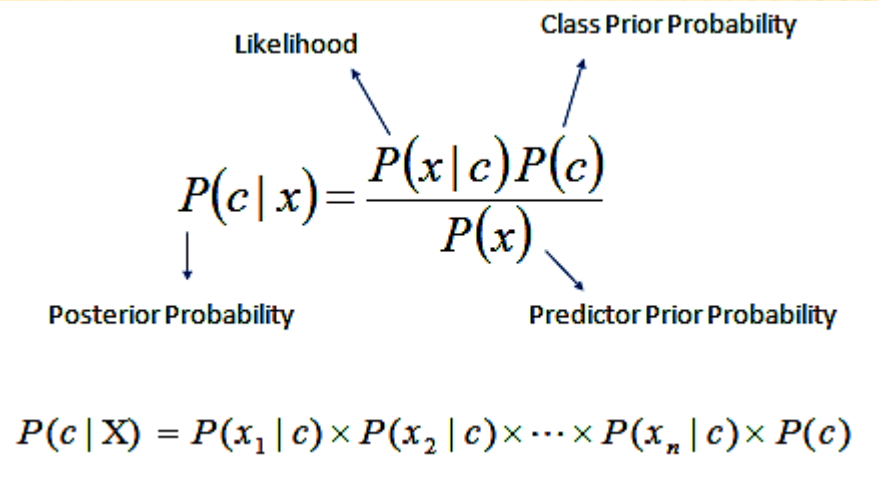
SENTIMENT ANALYSIS USING VADER SENTIMENT LIBRARY

RESULT (CONT.)


```
Out[69]: [0.9561484487899199,
          0.9471773117678483,
          0.9401087396786587,
          0.9519581876966348,
          0.9062738138953563,
          0.901637725429577]
```

LOGISTIC REGRESSION WITH 5-FOLD CROSS VALIDATION

RESULT (CONT.)




```
Out[11]: [0.5750640256303141, 0.57489283866258]
```

NAÏVE BAYES

ADVANTAGES

- More intelligent
- Accurate

DISADVANTAGE

- The time increases compare to linear as the number of failure attempt had increase in finding the new cities close by.
- As number of tree increases there is memory issue in GBT

OTHER APPLICATIONS

- Different Domain like social network, Search Engine, etc.

CONCLUSION

- In Logistic Regression, From the regularization parameter of 0.1 we have receive accuracy are 95.61%
- Similar, Using Naïve Bayes We received less than 58% accuracy
- Gradient Boost Tree suffered from memory problem as number of tree increased

REFERENCE

- [1]https://www.slideshare.net/gabrielspmoreira/discovering-users-topics-of-interest-in-recommender-systems-tdc-sp-2016
- [2] https://www.amazon.com/
- [3]http://guides.dataverse.org/en/latest/user/dataset-management.html

QUESTIONS