Big Data Analytics of Breast Cancer Using Twitter

Rishi Shah[1], Sheetal Pandrekar[2], Fusheng Wang[2], Xinyu Dong[2]

Columbia University[1], Department of Biomedical Informatics, Stony Brook University[2]

Stony Brook University

INTRODUCTION

BACKGROUND

❑ Leverage large scale Twitter data to investigate discourse regarding breast cancer

❑ Analysis performed through the use of natural language processing (NLP) techniques and a machine learning (ML) based approach

CHALLENGES

❑ Developing an effective method for feature extraction, textual tokenization, and analysis

❑ Obtaining robust training data and minimizing time required to train machine learning models on large volumes of data

OBJECTIVES

❑ To develop an effective approach to large scale Twitter data analysis grounded in natural language processing and machine learning

❑ To uncover meaningful insights and trends in public discourse regarding breast cancer

TWITTER DATA

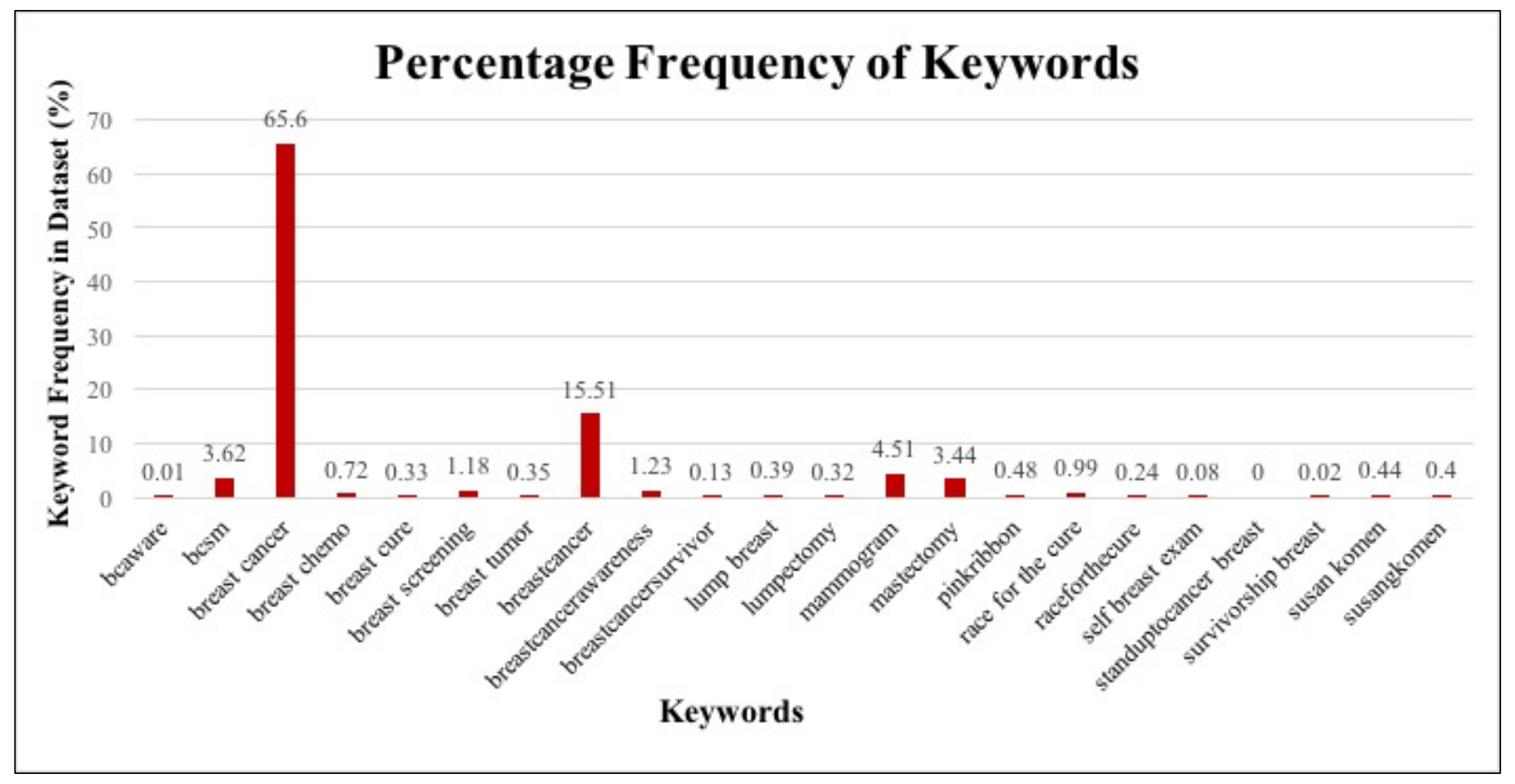
❑491, 172 tweets & 164, 384 unique users (01/01/17 - 06/19/17)

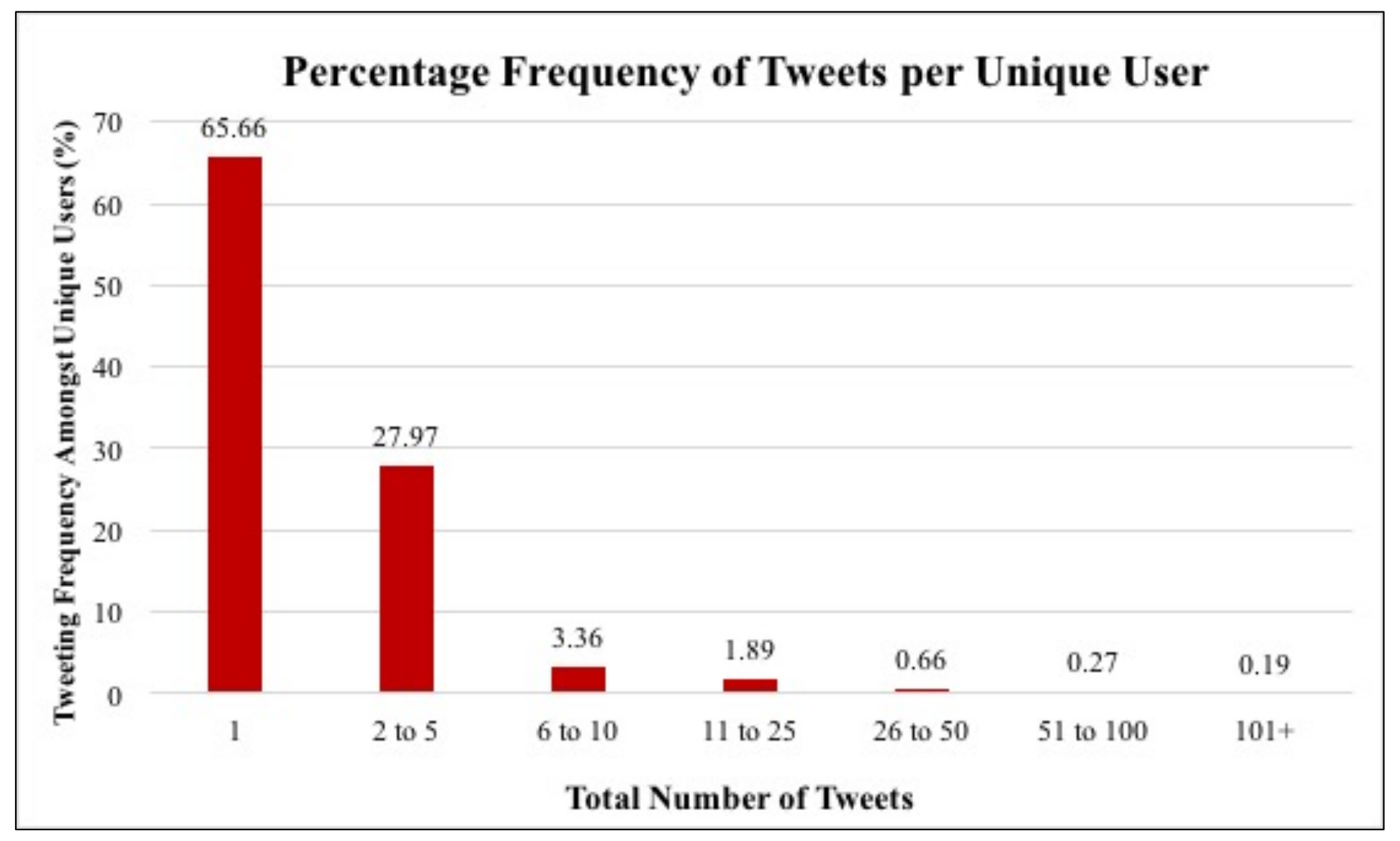
❑Twitter Search Query:

```
[“bcaware”, “bcsm”, “breast cancer”, “breast chemo”, “breast cure”,
“breast screening”, “breast tumor”, “breastcancer”,
“breastcancerawareness”, “breastcancersurvivor”, “lump breast”,
“lumpectomy”, “mammogram”, “mastectomy”, “pinkribbon”, “race for the
cure”,  “raceforthecure”, “self breast exam”, “standuptocancer
breast”, “survivorship breast”, “susan komen”, “susangkomen”]
```

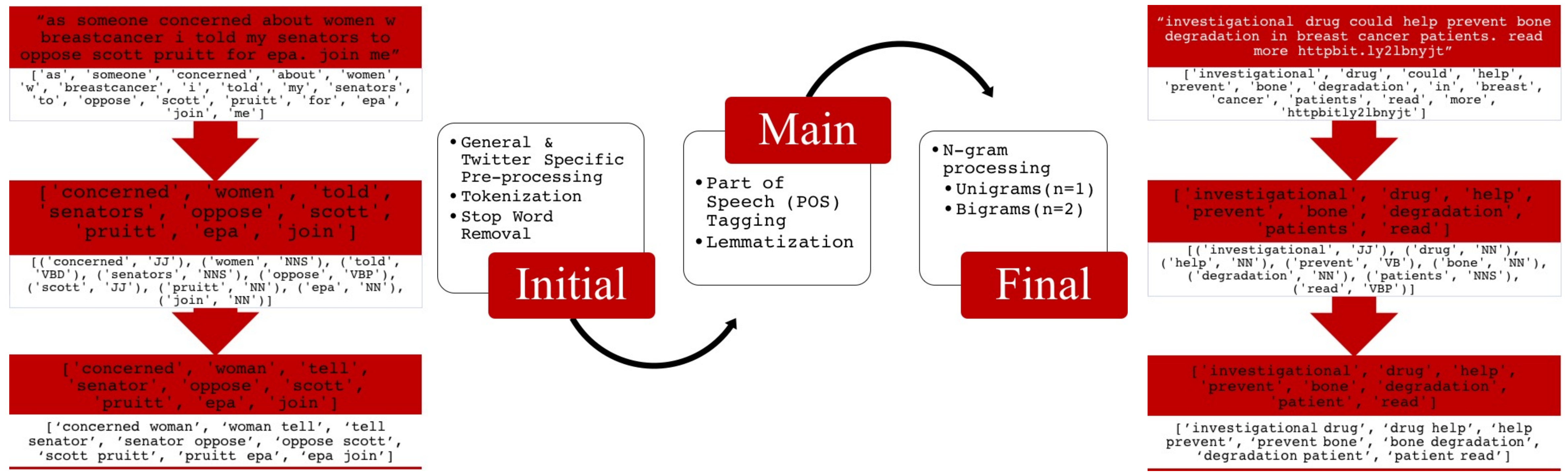
❑ Tweet Feature Extraction:

```
[Sr_No, Tweet_Id,  Created_At,  Timestamp_ms, Text, Keyword, User_Id,
Screen_name, Description, Retweet_cnt, Favorite_count, Friend_cnt,
Location, Geo]
```





NLP MODEL



“as someone concerned about women w breastcancer i told my senators to oppose scott pruitt for epa. join me”

['as', 'someone', 'concerned', 'about', 'women', 'w', 'breastcancer', 'i', 'told', 'my', 'senators', 'to', 'oppose', 'scott', 'pruitt', 'for', 'epa', 'join', 'me']

['concerned', 'women', 'told', 'senators', 'oppose', 'scott', 'pruitt', 'epa', 'join']

[('concerned', 'JJ'), ('women', 'NNS'), ('told', 'VBD'), ('senators', 'NNS'), ('oppose', 'VBP'), ('scott', 'JJ'), ('pruitt', 'NN'), ('epa', 'NN'), ('join', 'NN')]

['concerned', 'woman', 'tell', 'senator', 'oppose', 'scott', 'pruitt', 'epa', 'join']

[‘concerned woman’, ‘woman tell’, ‘tell senator’, ‘senator oppose’, ‘oppose scott’, ‘scott pruitt’, ‘pruitt epa’, ‘epa join’]

Initial
- General & Twitter Specific Pre-processing
- Tokenization
- Stop Word Removal

Main
- Part of Speech (POS) Tagging
- Lemmatization

Final
- N-gram processing
  - Unigrams(n=1)
  - Bigrams(n=2)

“investigational drug could help prevent bone degradation in breast cancer patients. read more httpbit.ly2lbnyjt”

['investigational', 'drug', 'could', 'help', 'prevent', 'bone', 'degradation', 'in', 'breast', 'cancer', 'patients', 'read', 'more', 'httpbitly2lbnyjt']

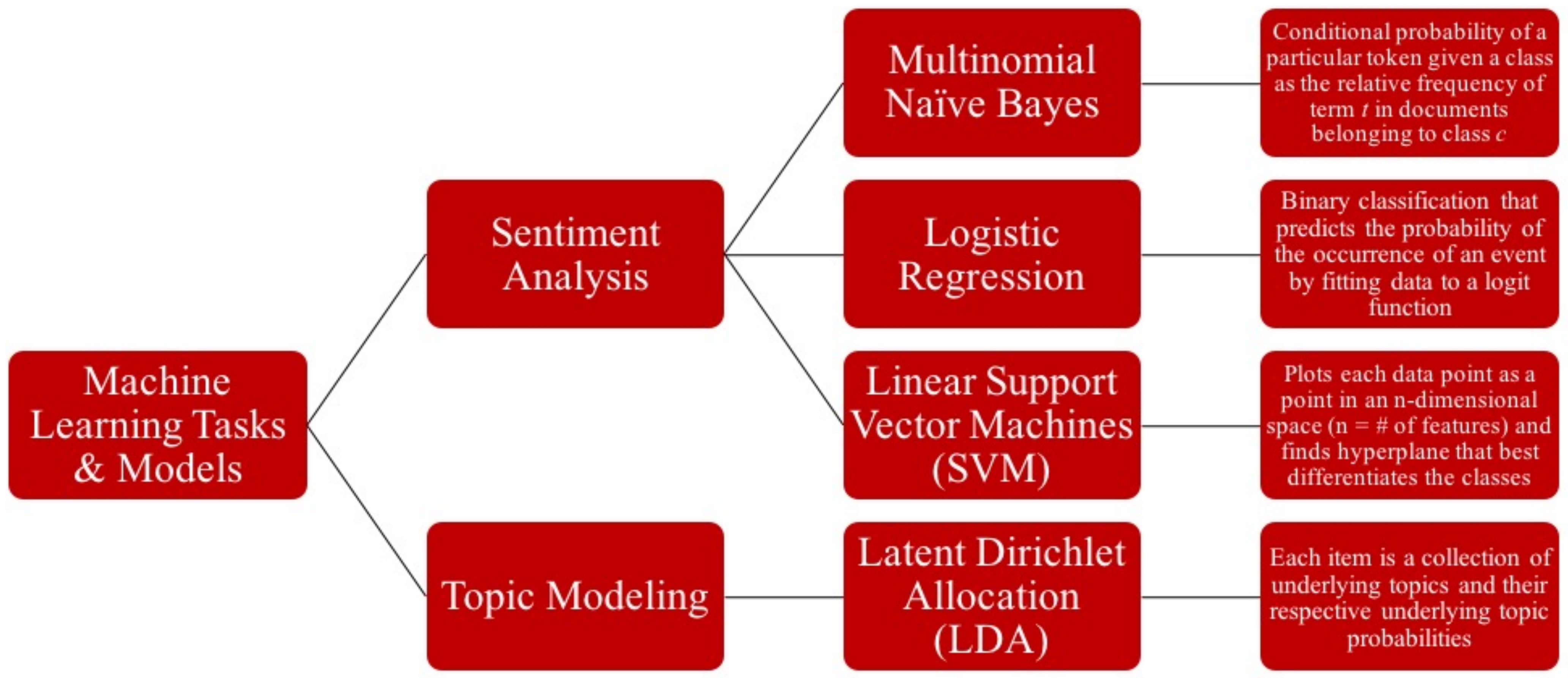
['investigational', 'drug', 'help', 'prevent', 'bone', 'degradation', 'patients', 'read']

[('investigational', 'JJ'), ('drug', 'NN'), ('help', 'NN'), ('prevent', 'VB'), ('bone', 'NN'), ('degradation', 'NN'), ('patients', 'NNS'), ('read', 'VBP')]

['investigational', 'drug', 'help', 'prevent', 'bone', 'degradation', 'patient', 'read']

[‘investigational drug’, ‘drug help’, ‘help prevent’, ‘prevent bone’, ‘bone degradation’, ‘degradation patient’, ‘patient read’]

SENTIMENT ANALYSIS & TOPIC MODELING



Machine Learning Tasks & Models

Sentiment Analysis

Topic Modeling

Multinomial Naïve Bayes: Conditional probability of a particular token given a class as the relative frequency of term *t* in documents belonging to class *c*

Logistic Regression: Binary classification that predicts the probability of the occurrence of an event by fitting data to a logit function

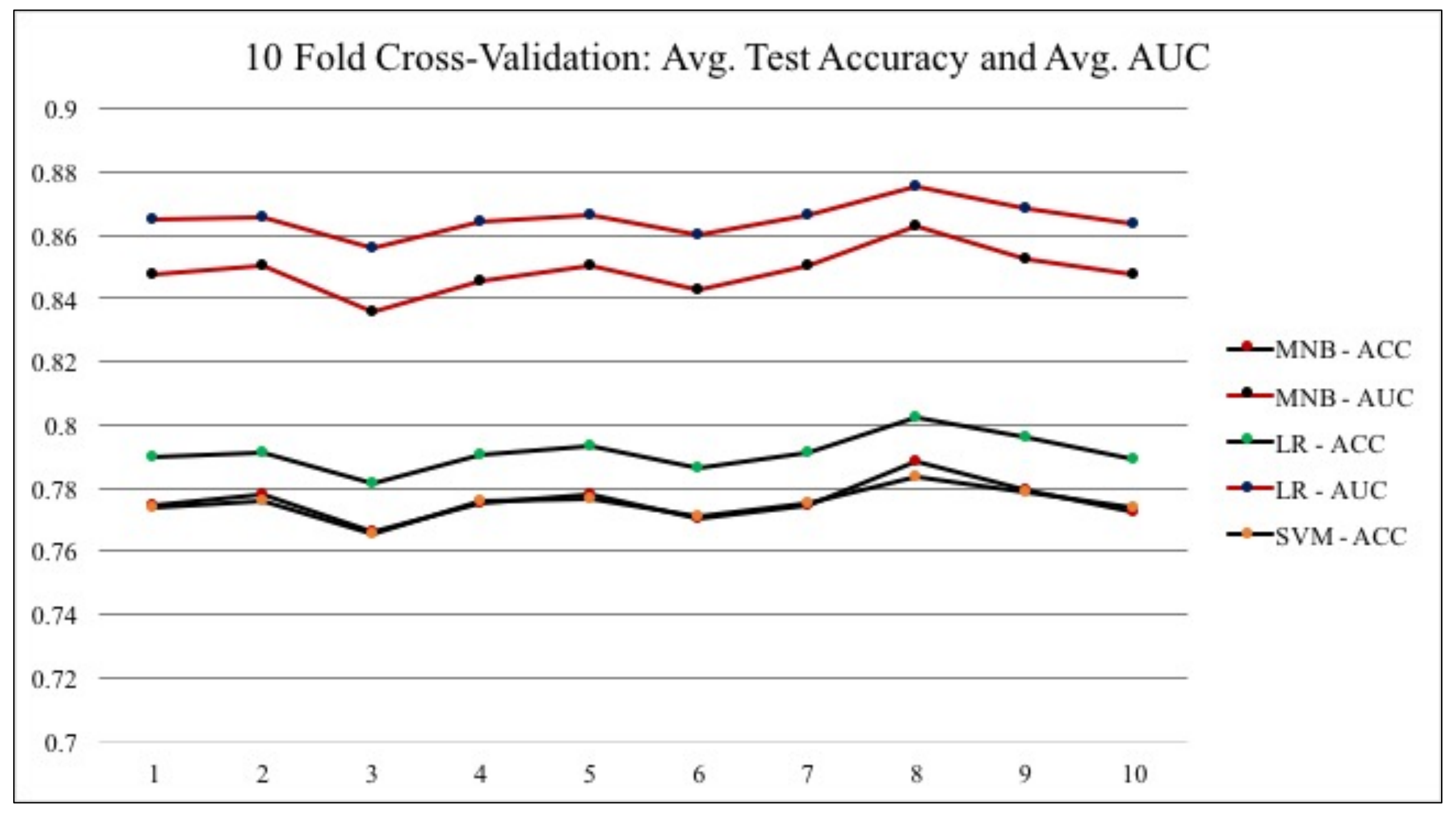
Linear Support Vector Machines (SVM): Plots each data point as a point in an n-dimensional space (n = # of features) and finds hyperplane that best differentiates the classes

Latent Dirichlet Allocation (LDA): Each item is a collection of underlying topics and their respective underlying topic probabilities

TRAINING & EVALUATION

TRAINING DATA

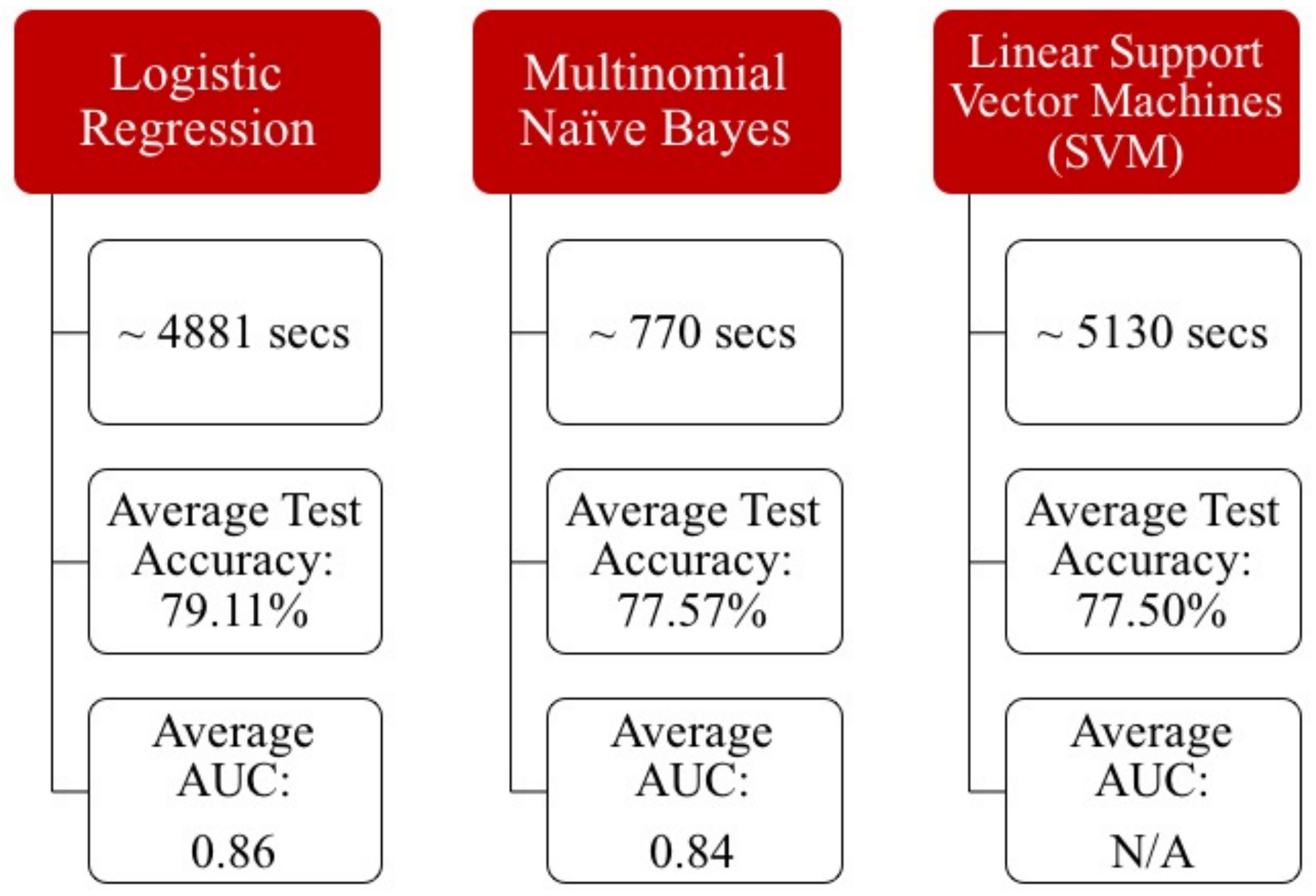
❑Sentiment140 Dataset: 1.6 million classified tweets (50% Positive, 50% Negative)



10 Fold Cross-Validation

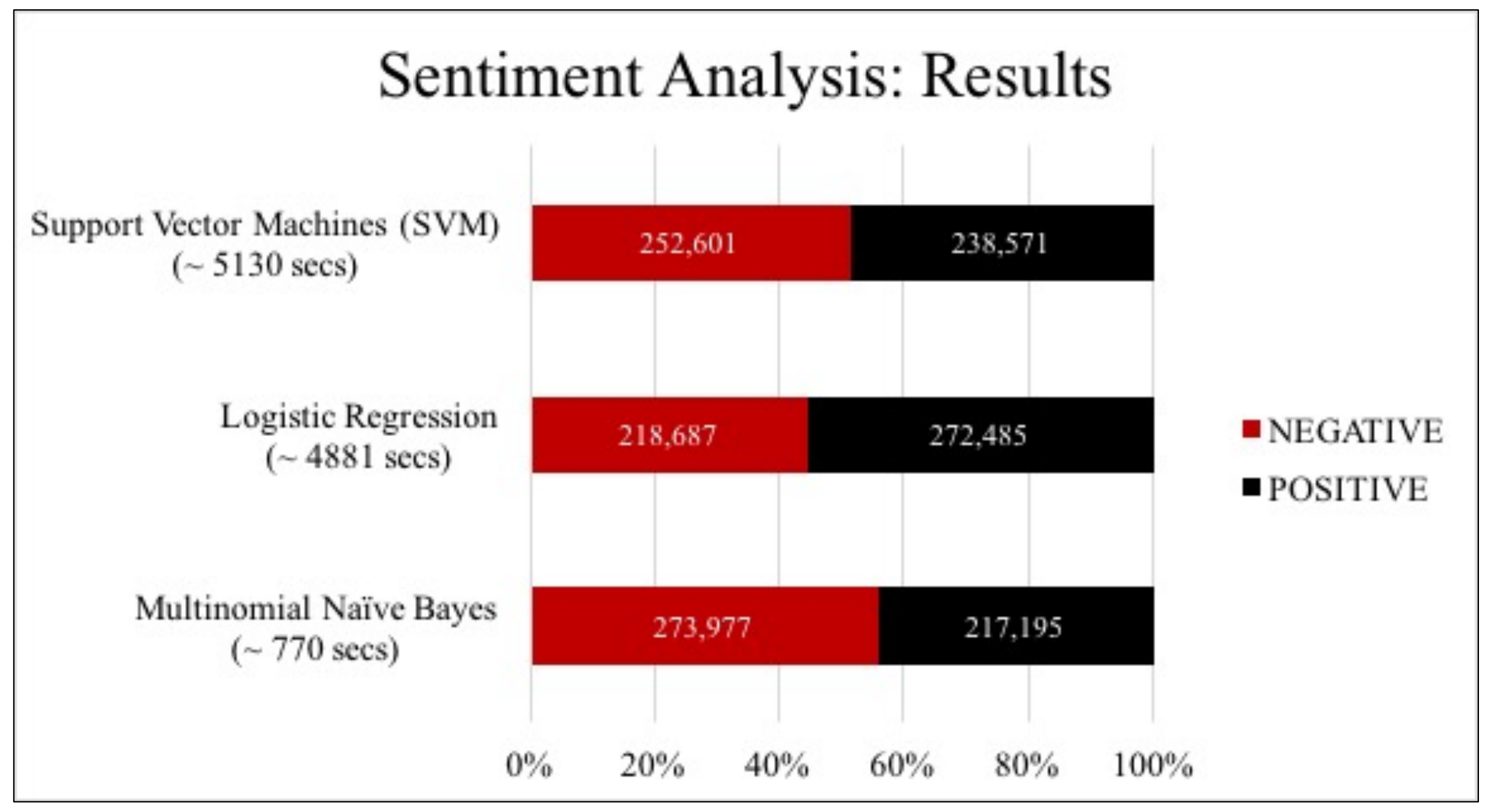
❑Avg. Test Accuracy (Baseline: 50%)

❑Avg. AUC (Area Under the Curve)



Logistic Regression	Multinomial Naïve Bayes	Linear Support Vector Machines (SVM)
~ 4881 secs	~ 770 secs	~ 5130 secs
Average Test Accuracy: 79.11%	Average Test Accuracy: 77.57%	Average Test Accuracy: 77.50%
Average AUC: 0.86	Average AUC: 0.84	Average AUC: N/A

EXPERIMENTAL RESULTS



```
Topic #1 ['via', 'life', 'save', 'metastatic', 'live', 'pay',
'senator', 'million', 'shes', 'bra']

Topic #2 ['love', 'tweet', 'day', 'please', 'joke', 'mri',
'check', 'hope', 'month', 'camilameetshanelle']

Topic #4 ['support', 'fight', 'help', 'join', 'please',
'walk', '2017', 'donate', 'hair', 'charity']

Topic #5 ['pin', 'market', 'boob', 'forecast', '2022',
'diabetes', 'start', 'cant', 'kissward', 'change']

Topic #6 ['drug', 'increase', 'roche', 'call', 'cut',
'trial', 'cell', 'positive', 'nhs', 'growth']

Topic #7 ['amp', 'raise', 'read', 'wifes', 'day', 'fund',
'brca', 'surgerys', 'coverage', 'autistic']

Topic #8 ['woman', 'mom', 'surgery', 'double', 'diagnosis',
'prevent', 'mammogram', 'symptom', 'cause', 'tell']

Topic #9 ['study', 'foundation', 'research', 'woman',
'reduce', 'link', 'read', 'find', 'aspirin', 'advance']

Topic #10 ['woman', 'news', 'health', 'battle', 'sign',
'post', 'mother', 'wife', 'stop', 'amp']

Topic #12 ['patient', 'survivor', 'look', 'care', 'dont',
'mammography', 'woman', 'ultrasound', 'doctor', 'test']

Elapsed Time: ~28,051 secs
```

CONCLUSION & FUTURE WORK

❑ Utilizing an approach grounded in machine learning and natural language processing allows for robust and scalable insights into large-scale, textual datasets

❑ Twitter, as a medium for exploring societal discourse, is an effective means of understanding current trends and discussion topics affecting the public

❑ Future Work: User-Wise Classification, Parameter Modification, Minimization of Training Time

ACKNOWLEDGEMENTS

I would like to thank Dr. Fusheng Wang, Sheetal Pandrekar, Xinyu Dong, and Heather Paquette from the Department of Biomedical Informatics at Stony Brook University for their support and guidance. This work was made possible through the National Science Foundation (NSF) REU program and was funded by the NSF award “CIF21 DIBBs: Middleware and High Performance Analytics Libraries for Scalable Data Science.”

REFERENCES

[1] S. Zhang *et. al.*: Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *Journal of Biomedical Informatics* 69 (2017) 1– 9.

[2] Thackeray *et al.*: Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC Cancer* 2013 **13**:508.

[3] Bird, Steven, Ewan Klein, and Edward Loper (2009), Natural Language Processing with Python, O'Reilly Media.

[4] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[5] A. Go, R. Bhayani, L. Huang: Twitter Sentiment Classification Using Distant Supervision. *Processing* (*2009*).

[6] Manning, Christopher D., et al. *Introduction to Information Retrieval*. Cambridge University Press, 2009.