Big Data Analytics of Breast Cancer Using Twitter
Rishi Shah¹, Sheetal Pandrekar², Fusheng Wang², Xinyu Dong²
Columbia University¹, Department of Biomedical Informatics, Stony Brook University²

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:
  - Breast cancer, breast cancer survivorship, breast cancer cure, lumpectomy, mammogram, mastectomy, breast screening, breast tumor
  - Breast cancer awareness, breast cancer screening, breast cancer treatment, breast cancer research

Twitter Feature Extraction:
- Location, Geo

EXPERIMENTAL RESULTS
SUPPORT VOTE MACHINE (SVM)
- Training & Evaluation
  - Training Data: 1.6 million classified tweets (50% Positive, 50% Negative)
  - 10 Fold Cross-Validation
  - Multinomial Naive Bayes
  - Linear Support Vector Machines (SVM)

TRAINING DATA
- Sentiment140 Dataset: 1.6 million classified tweets (50% Positive, 50% Negative)

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 DBHLS: Middleware and High Performance Analytics Libraries for Scalable Data Science.”

REFERENCES