9th Annual FDA Scientific Computing Day

NLP OF CGMP INSPECTIONAL OBSERVATIONS

Who are we?

- Taylor Henderson
- PhD Student at George Mason University
- Summer 2021 ORISE Fellow
- Mentored by John Wan

Who are we?

- Tingting Wang
- Recent graduate from University of Florida
- 2020-2021 ORISE Fellow
- Mentored by John Wan

Background Information

Disclaimer: Findings and conclusions contained herein have not been formally disseminated by the FDA and should not be construed to represent any agency determination or policy

Background Information

- Compliance with current good manufacturing practice (cGMP) through routine surveillance inspections.
- An FDA form 483 is issued when an investigator observed conditions that may constitute violations.
- FDA 483 contains detailed observed violations with the associated violation citation code.
- Understanding this data could help the FDA-CDER better carry out its regulatory function.

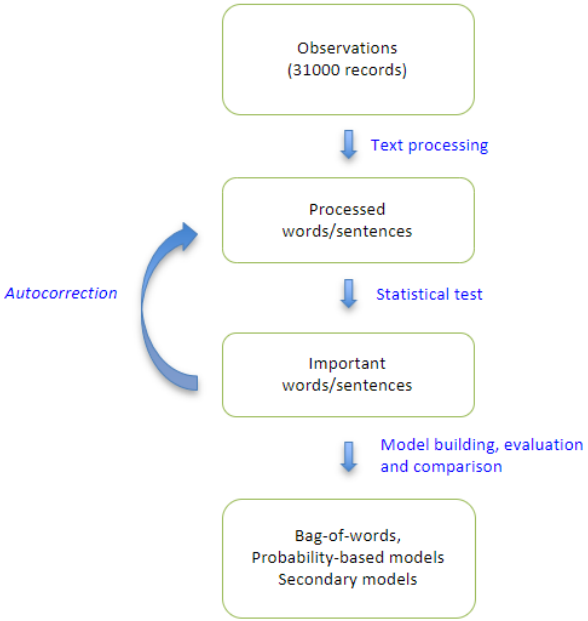
Dataset Challenges

- Unique vocabulary
  - Many pharmaceutical and medical terms
- Small dataset
  - 17,390 unique samples
  - μ = 145 words, σ = 169 words
- Imbalanced
  - 1:23

Feature Engineering/Selection

- Stemming/lemmatization
  - Reduce novel words
- Autocorrect
  - Reduce OCR errors in data
- Chi-squared feature selection and stop word removal
  - Reduce the vocab size to best use available resources
- SMOTE
  - Minimize class imbalance

Processing Pipeline



- Whitespace tokenization
- Stemming
- Chi-squared feature selection
- Bag-of-words
- Frequency-based Naïve bayes

Top Model Results (two class)

Naïve Bayes w/ BOW

Metrics	Value
F1-Score	0.72
Precision	0.63
Recall	0.82
AUC	0.90

Longformer (1024)

Metrics	Value
F1-Score	0.98
Precision	1.0
Recall	0.96
AUC	0.72

Top Model Result (multiclass)

Word2vec w/ 1d CNN

Metrics	Value
F1-Score	0.82
Precision	0.91
Recall	0.77

Current Work

- Continued exploration of deep learning models
  - MLP, word2vec, CNN
- Continued use of transfer learning to address corpus size
  - BERT, Longformer


FDA
U.S. FOOD & DRUG
ADMINISTRATION