Social Media Mining for Health Monitoring

Team Stellar

Introduction

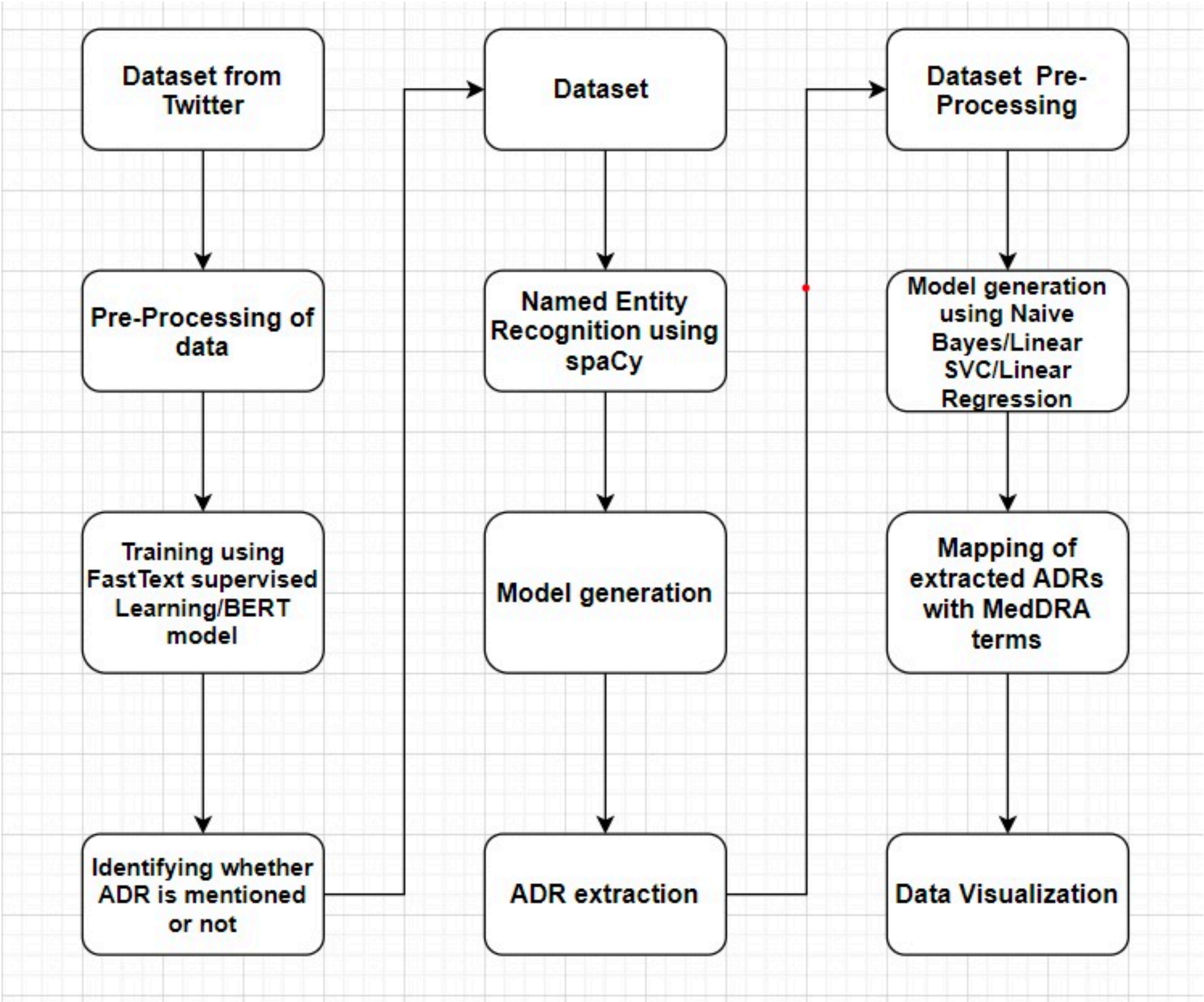
- Problem Statement : Use NLP techniques for mining raw and imbalanced tweets to retrieve and visualize the adverse affects along with MedDRA IDs for various drugs used by the tweeters.

- Architecture :

  - Automatic classification of adverse effects mentions in tweets.

  - Extraction of Adverse Effect mentions.

  - Mapping of extracted ADR's with MedDRA terms.


Dataset from Twitter
Pre-Processing of data
Training using FastText supervised Learning/BERT model
Identifying whether ADR is mentioned or not
Dataset
Named Entity Recognition using spaCy
Model generation
ADR extraction
Dataset Pre-Processing
Model generation using Naive Bayes/Linear SVC/Linear Regression
Mapping of extracted ADRs with MedDRA terms
Data Visualization

Task 1 : Automatic classification of adverse effects mentions in tweets.

Dataset : 55,420 Tweets. 146 Positive, 55274 Negative. The Dataset is SKEWED.
Format of the data: TWEETID <-tab-> USERID <-tab-> TWEET<-tab->CREATED_AT <-tab-> CLASS (1=ADR; 0=NON-ADR)

FASTTEXT Model

- Supervised Model by Facebook.
- Improvement over Word2Vec Model.
- Relied on CBOW Model.
- Data Cleaning and Fasttext prerequisites.
- Performed Under Sampling.
- **F1 Score - 0.83.**

BERT

- "Bidirectional" Encoder Representations from Transformer.
- Tokenization & Input Formatting.
- BertForSequenceClassification.
- Used Mathews Correlation Coefficient as performance measure (-1 to +1).
- MCC - 0.771.
- **F1 Score - 0.88**

Task 2 : Extraction of Adverse Effect mentions

Dataset : 2247 Tweets.

Format of the data: TWEET <-tab-> begin <-tab-> end <-tab-> extraction

SpaCy - en_core_web_sm

- Pre-process.
- ADR Tagging.
- Use en_core_web_sm and train model.
- Evaluation using gold parse.
- **F1 Score - 48.32.**

Task 3 : Mapping of extracted ADR’s with MedDRA terms.

Dataset : Format of the data: TWEETID <-tab-> BEGINNING OF ADR<-tab-> ENDING OF ADR IN TWEET <-tab-> ADR (OR NOT) <-tab-> ADR EXTRACTION <-tab-> DRUG <-tab-> TWEET <-tab-> MEDDRA CODE<-tab-> MEDDRA TERM

Objective :

- To Detect tweets mentioning an ADR.
- To map the extracted colloquial mentions of ADRs in the tweets to standard concept IDs in the MedDRA vocabulary (lower level terms).

- dropna() function
- Text documents to a matrix of token counts (CountVectorizer)
- Then transform a count matrix to a normalized tf-idf representation (tf-idf transformer)

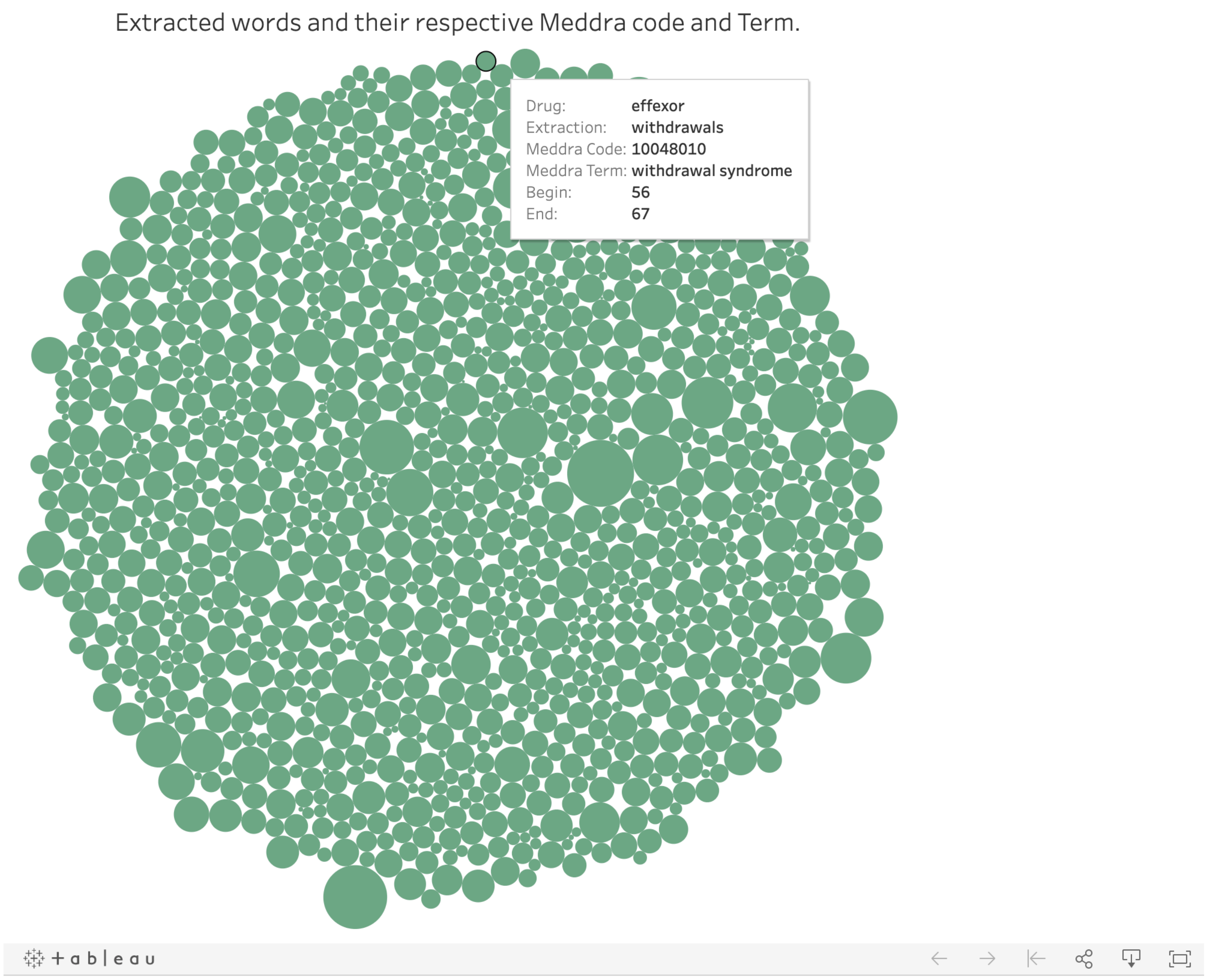
- Train with multiple models.
- Models
  - - Naive Bayes (Baseline)
  - - Linear Support Vector Machine
  - - Logistic Regression
- Use pipeline class
- Fit.

Naive Bayes (F-Measure - 0.16)
Linear SVM (F-Measure - 0.35)
Logistic Regression (F-Measure - 0.36)

	BASELINE MODEL	FINAL
TASK 1	FastText (F-Measure - 0.73)	FastText (F-Measure - 0.83) BERT (F-Measure - 0.88)
TASK 2	SpaCy (F-Measure - 0.21)	SpaCy (F-Measure - 0.48)
TASK 3	Naive Bayes (F-Measure - 0.16)	Naive Bayes (F-Measure - 0.16) Linear SVM (F-Measure - 0.35) Logistic Regression (F-Measure - 0.36)

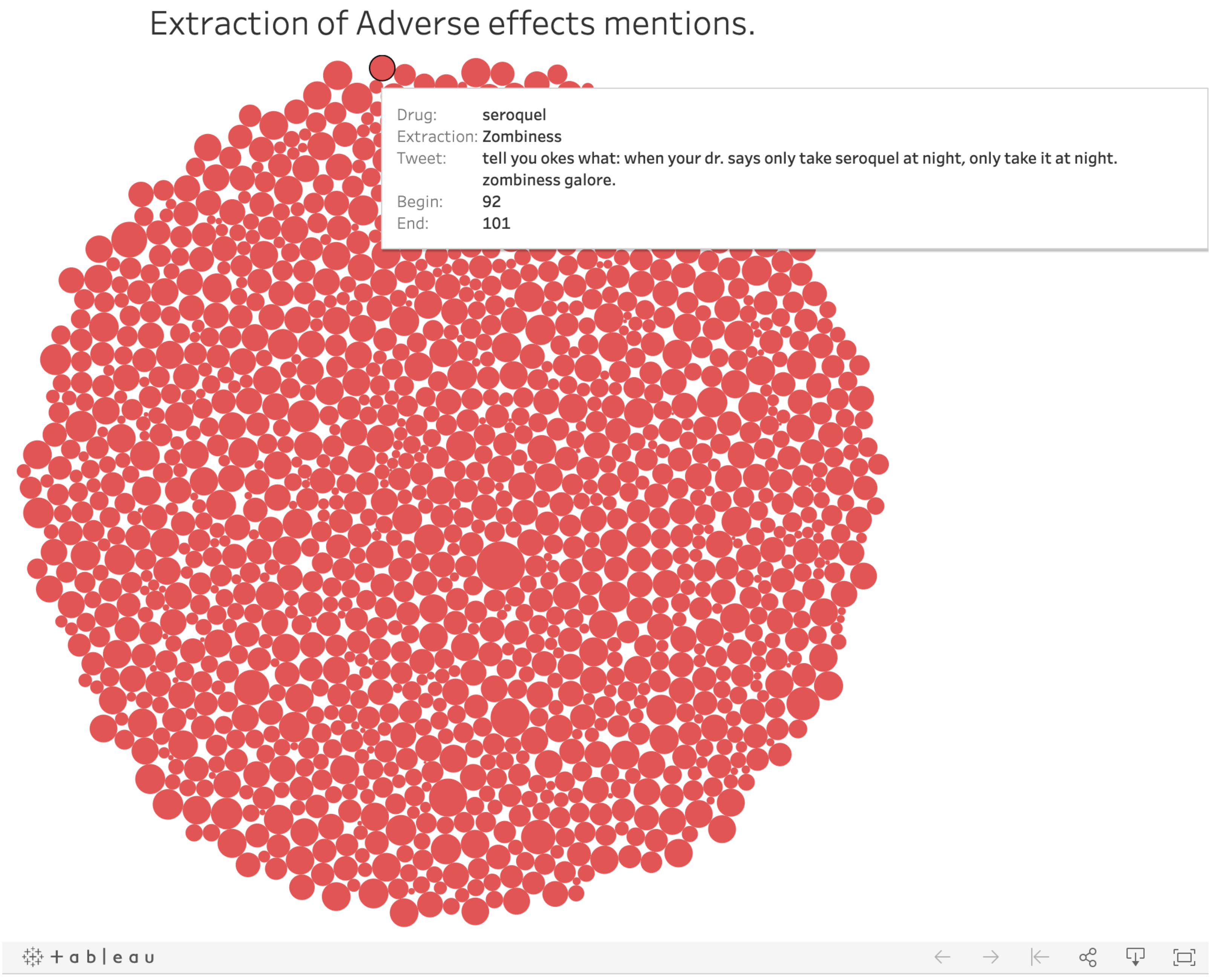
Data Visualization using TABLEAU

Website Link: http://ec2-18-216-171-204.us-east-2.compute.amazonaws.com/



Extracted words and their respective Meddra code and Term.

Drug: effexor
Extraction: withdrawals
Meddra Code: 10048010
Meddra Term: withdrawal syndrome
Begin: 56
End: 67



Extraction of Adverse effects mentions.

Drug: seroquel
Extraction: Zombiness
Tweet: tell you okes what: when your dr. says only take seroquel at night, only take it at night. zombiness galore.
Begin: 92
End: 101

Thank You