SENTIMENT ANALYSIS OF ONLINE PRODUCT REVIEWS AND COMPARISON OF SUPERVISED MACHINE LEARNING ALGORITHMS

PROJECT BY: JASMEET SINGH SASAN

AGENDA

- Objective and Motivation
- Sentiment Analysis and Supervised Machine Learning Algorithms
- Hypothesis
- Data
- Research Framework
  - Data Pre-Processing
  - Feature Engineering
  - Model Building
  - Model Evaluation and Comparison
- Experiment Results and Comparison
- Conclusion and Challenges
- Questions

TOOLS AND TECHNOLOGIES USED

- R (Statistical Programming Language)
- Amazon Web Services – Elastic Compute Cloud (AWS EC2)
- Tableau (Data Visualization)

OBJECTIVE AND MOTIVATION

Goal:

- Perform Sentiment Analysis on text data of Online Product Reviews.
- Compare the performance of key supervised machine learning algorithms on big datasets

Motivation

- Ocean of Data is available
- Understand the Efficiency of existing Literature and Algorithms
- Learn core functions of NLP.

SENTIMENT ANALYSIS

- Sentiment Analysis (SA) is a field of Natural Language Processing (NLP) with a primary objective of extracting subjective information from the text data by developing algorithmic models and theories.
- .The text data with subjective information often contains expressions of opinions and viewpoints. SA touches every aspect of NLP yet confined in several ways
- Classification of subjective text data into positive, neutral and negative is the primary focus of SA.
- Sentiment Classification is done at the following levels:
  - Sentence Level
  - Document Level
  - Aspect Level

SENTIMENT CLASSIFICATION APPROACHES

Sentiment Classification can be done using one or more of the following learning methods:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Lexicon Based Methods

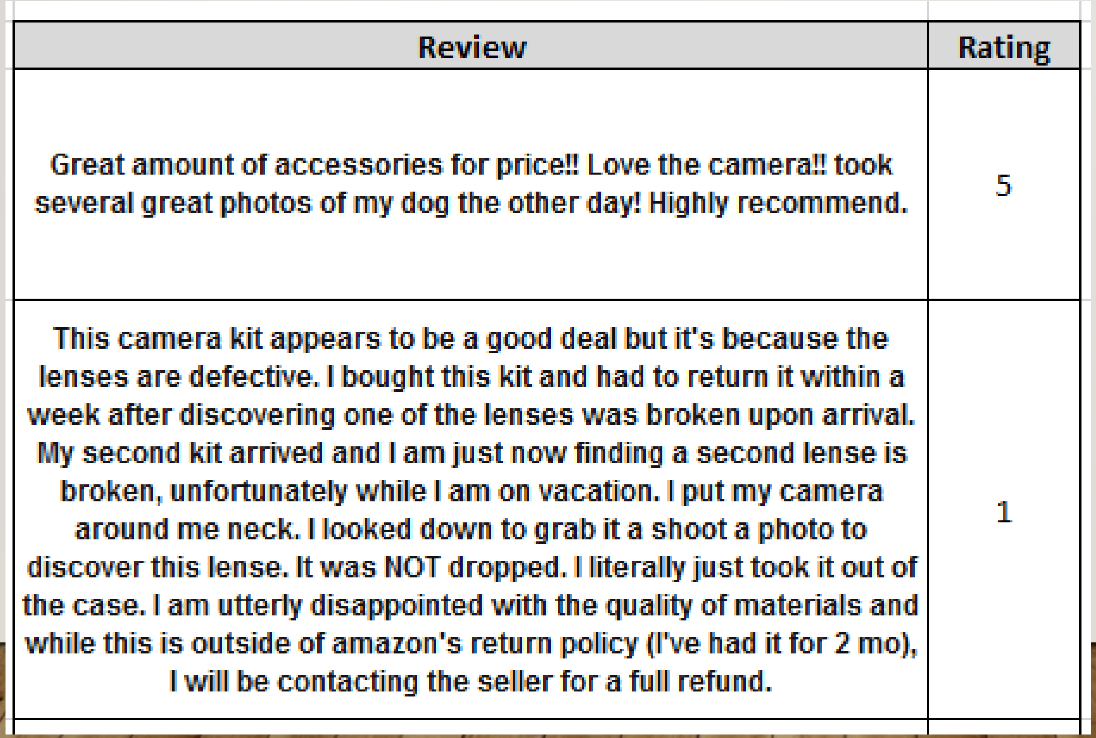
SUPERVISED MACHINE LEARNING

- Supervised Learning is a prediction method where we have output available corresponding to each input and a learning a function is derived using an algorithm which can map inputs to the correct outputs.
- Output could be a continuous or categorical value.
- Input is a set of useful variables which holds a relationship with the output.
- The learning function is used to predict the output of unseen data.
- Following algorithms have been used in this study:

SVM	Bagging
Neural Net	SLDA
Random Forest	Tree
Boosting	Maximum Entropy
Naïve Bayes	

HYPOTHESIS

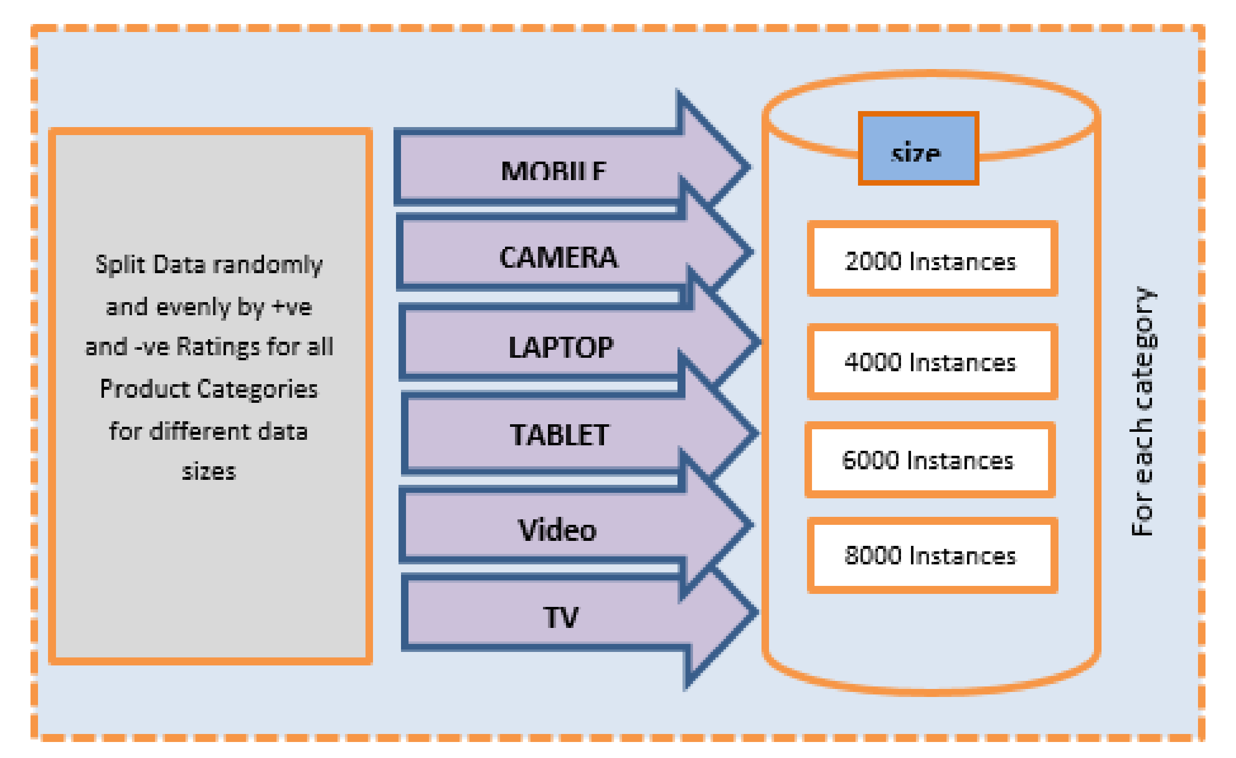
- Product reviews with rating of 1 or 2 are negative, and 4 or 5 are positive.
- Data instances with less than 10 words are not used in the models



Review	Rating
Great amount of accessories for price!! Love the camera!! took several great photos of my dog the other day! Highly recommend.	5
This camera kit appears to be a good deal but it's because the lenses are defective. I bought this kit and had to return it within a week after discovering one of the lenses was broken upon arrival. My second kit arrived and I am just now finding a second lense is broken, unfortunately while I am on vacation. I put my camera around me neck. I looked down to grab it a shoot a photo to discover this lense. It was NOT dropped. I literally just took it out of the case. I am utterly disappointed with the quality of materials and while this is outside of amazon's return policy (I've had it for 2 mo), I will be contacting the seller for a full refund.	1

DATA

- Product reviews data from Amazon.com
- Six Categories of Products
- Four sizes of datasets
- 6*4= 24 datasets
- Each dataset has equal number of positive and negative examples



Split Data randomly and evenly by +ve and -ve Ratings for all Product Categories for different data sizes

MOBILE

CAMERA

LAPTOP

TABLET

Video

TV

size

2000 Instances

4000 Instances

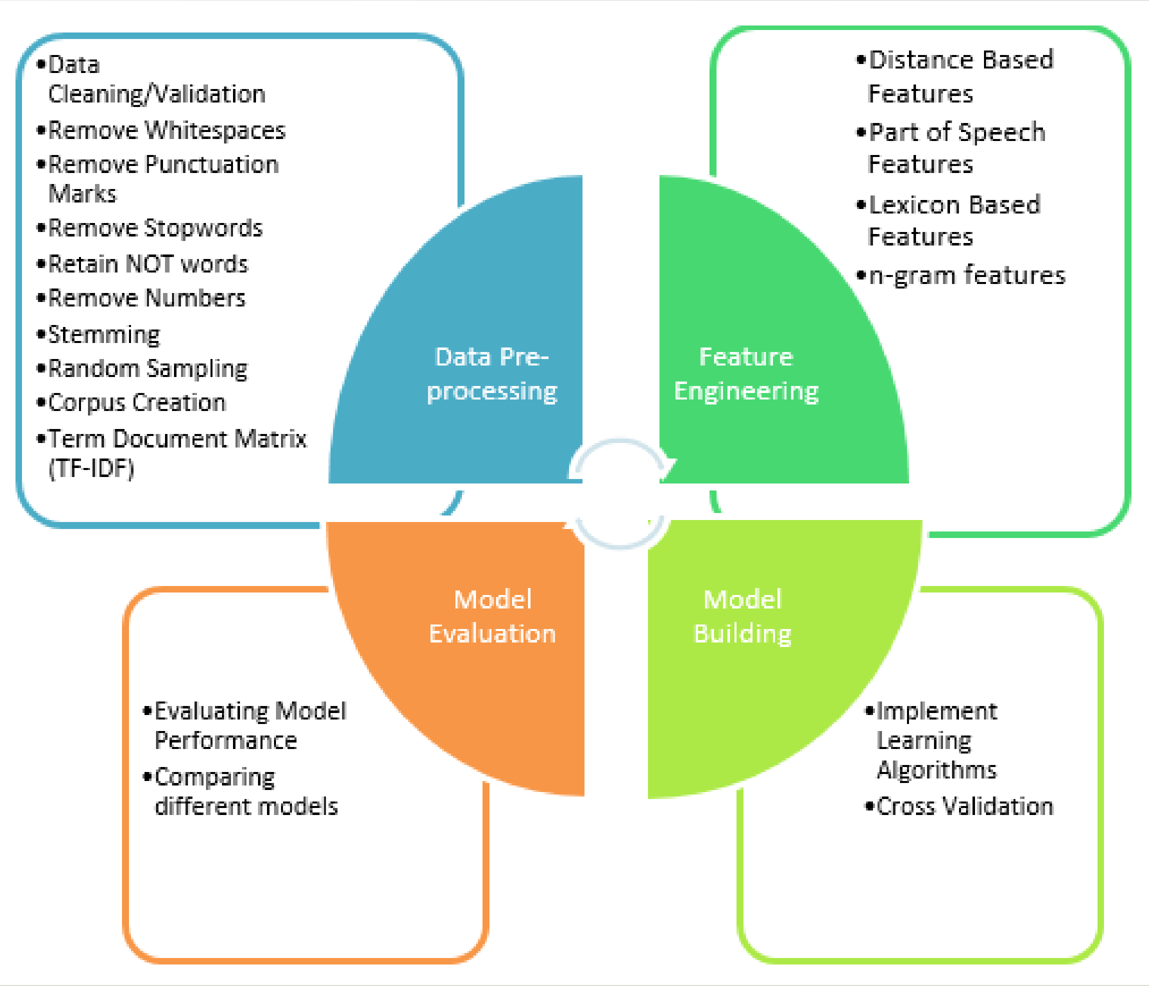
6000 Instances

8000 Instances

For each category

RESEARCH FRAMEWORK

- Data Import
- Data Pre-Processing
- Feature Engineering
- Predictive Model Building
- Model Evaluation and Comparison



Data Pre-processing

- Data Cleaning/Validation
- Remove Whitespaces
- Remove Punctuation Marks
- Remove Stopwords
- Retain NOT words
- Remove Numbers
- Stemming
- Random Sampling
- Corpus Creation
- Term Document Matrix (TF-IDF)

Feature Engineering

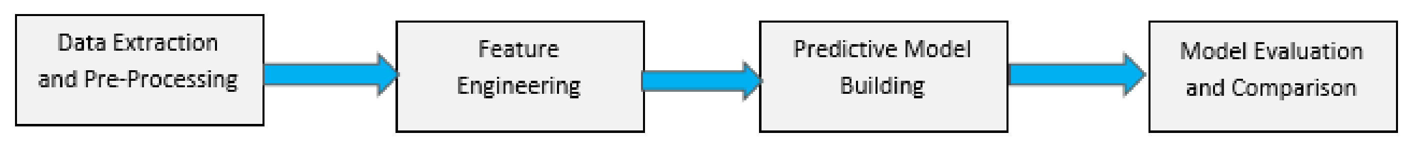
- Distance Based Features
- Part of Speech Features
- Lexicon Based Features
- n-gram features

Model Building

- Implement Learning Algorithms
- Cross Validation

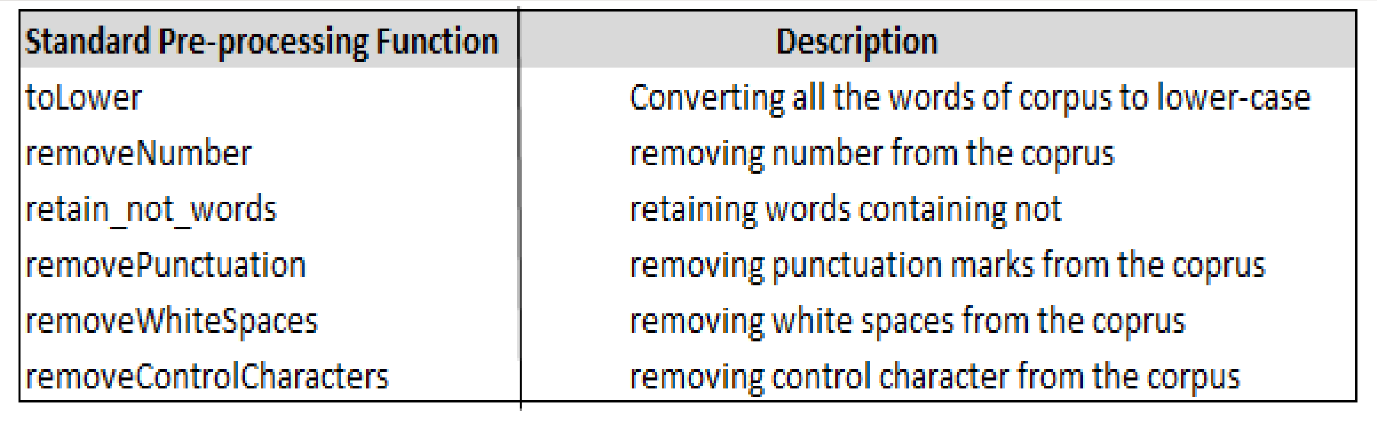
Model Evaluation

- Evaluating Model Performance
- Comparing different models

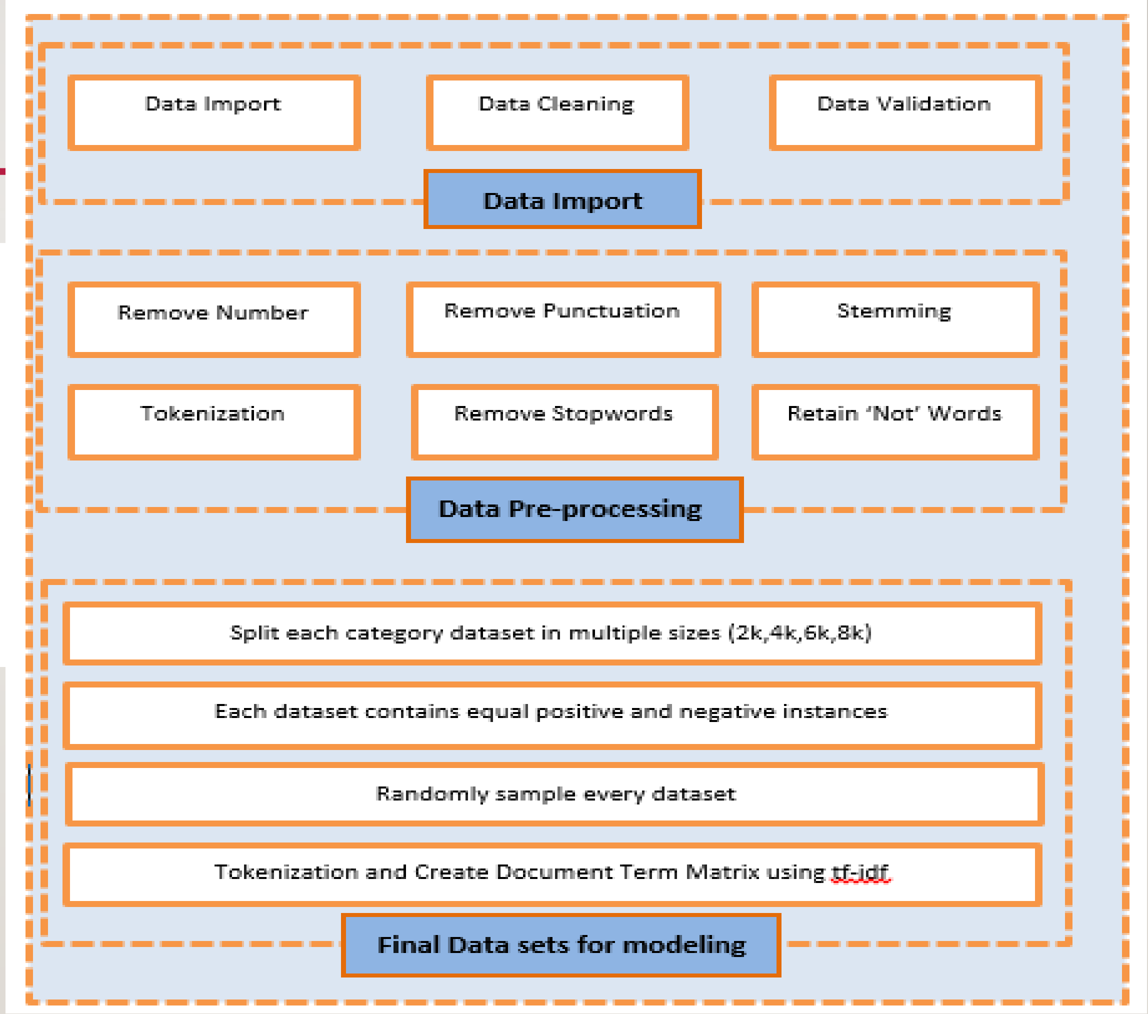


Data Extraction and Pre-Processing → Feature Engineering → Predictive Model Building → Model Evaluation and Comparison

DATA PRE-PROCESSING



Standard Pre-processing Function	Description
toLower	Converting all the words of corpus to lower-case
removeNumber	removing number from the coprus
retain_not_words	retaining words containing not
removePunctuation	removing punctuation marks from the coprus
removeWhiteSpaces	removing white spaces from the coprus
removeControlCharacters	removing control character from the corpus



Data Import

Data Cleaning

Data Validation

Data Import

Remove Number

Remove Punctuation

Stemming

Tokenization

Remove Stopwords

Retain 'Not' Words

Data Pre-processing

Split each category dataset in multiple sizes (2k,4k,6k,8k)

Each dataset contains equal positive and negative instances

Randomly sample every dataset

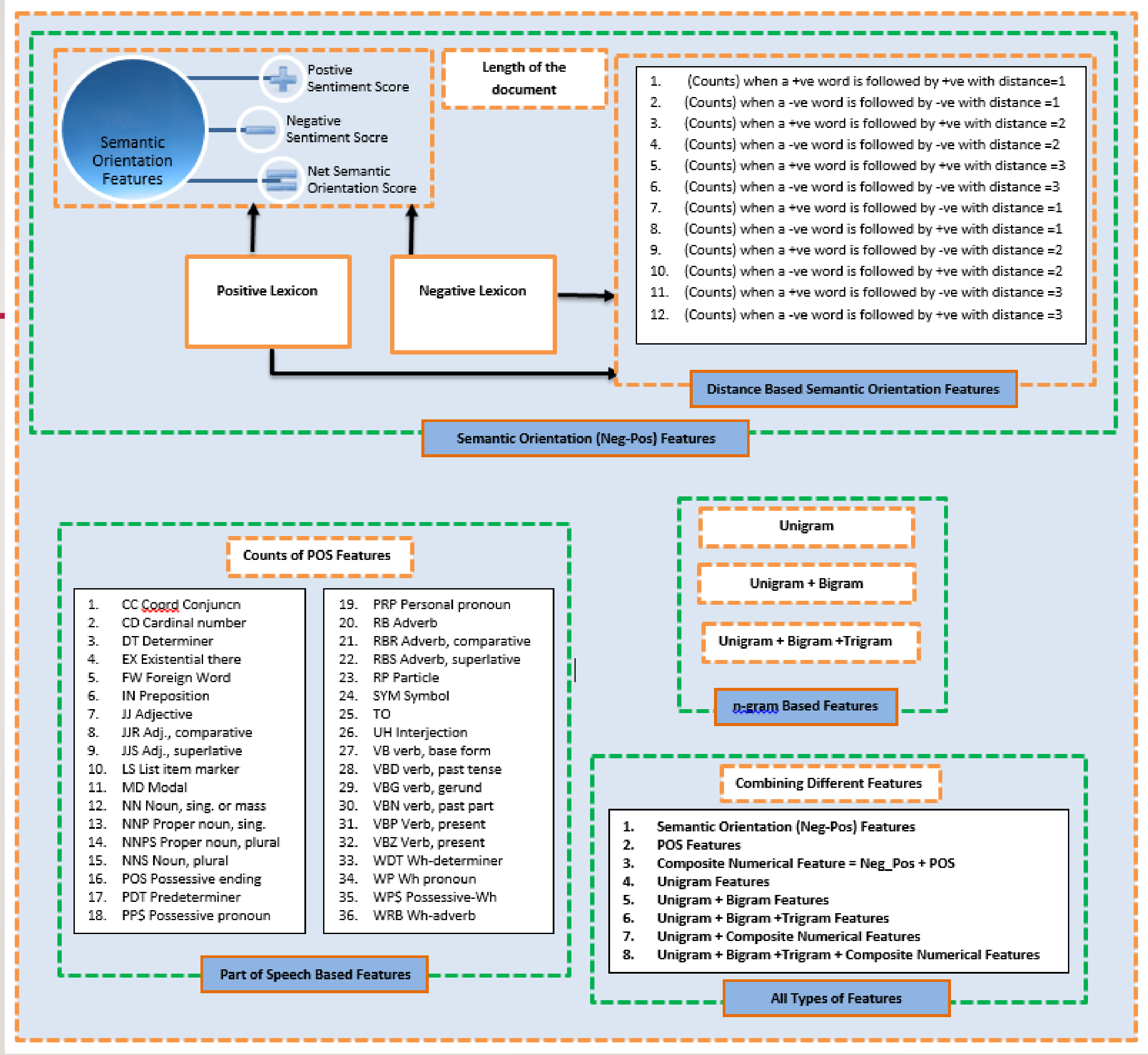
Tokenization and Create Document Term Matrix using tf-idf

Final Data sets for modeling

FEATURE ENGINEERING

Two types of features have been created:

- Semantic Features – Words polarity, Semantic Orientation
- Syntactic Features- Structure and form of the sentence



Semantic Orientation Features

Postive Sentiment Score

Negative Sentiment Socre

Net Semantic Orientation Score

Length of the document

Positive Lexicon

Negative Lexicon

1. (Counts) when a +ve word is followed by +ve with distance=1
2. (Counts) when a -ve word is followed by -ve with distance =1
3. (Counts) when a +ve word is followed by +ve with distance =2
4. (Counts) when a -ve word is followed by -ve with distance =2
5. (Counts) when a +ve word is followed by +ve with distance =3
6. (Counts) when a -ve word is followed by -ve with distance =3
7. (Counts) when a +ve word is followed by -ve with distance =1
8. (Counts) when a -ve word is followed by +ve with distance =1
9. (Counts) when a +ve word is followed by -ve with distance =2
10. (Counts) when a -ve word is followed by +ve with distance =2
11. (Counts) when a +ve word is followed by -ve with distance =3
12. (Counts) when a -ve word is followed by +ve with distance =3

Distance Based Semantic Orientation Features

Semantic Orientation (Neg-Pos) Features

Counts of POS Features

1. CC Coord Conjuncn
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign Word
6. IN Preposition
7. JJ Adjective
8. JJR Adj., comparative
9. JJS Adj., superlative
10. LS List item marker
11. MD Modal
12. NN Noun, sing. or mass
13. NNP Proper noun, sing.
14. NNPS Proper noun, plural
15. NNS Noun, plural
16. POS Possessive ending
17. PDT Predeterminer
18. PP$ Possessive pronoun
19. PRP Personal pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO
26. UH Interjection
27. VB verb, base form
28. VBD verb, past tense
29. VBG verb, gerund
30. VBN verb, past part
31. VBP Verb, present
32. VBZ Verb, present
33. WDT Wh-determiner
34. WP Wh pronoun
35. WP$ Possessive-Wh
36. WRB Wh-adverb

Part of Speech Based Features

Unigram

Unigram + Bigram

Unigram + Bigram +Trigram

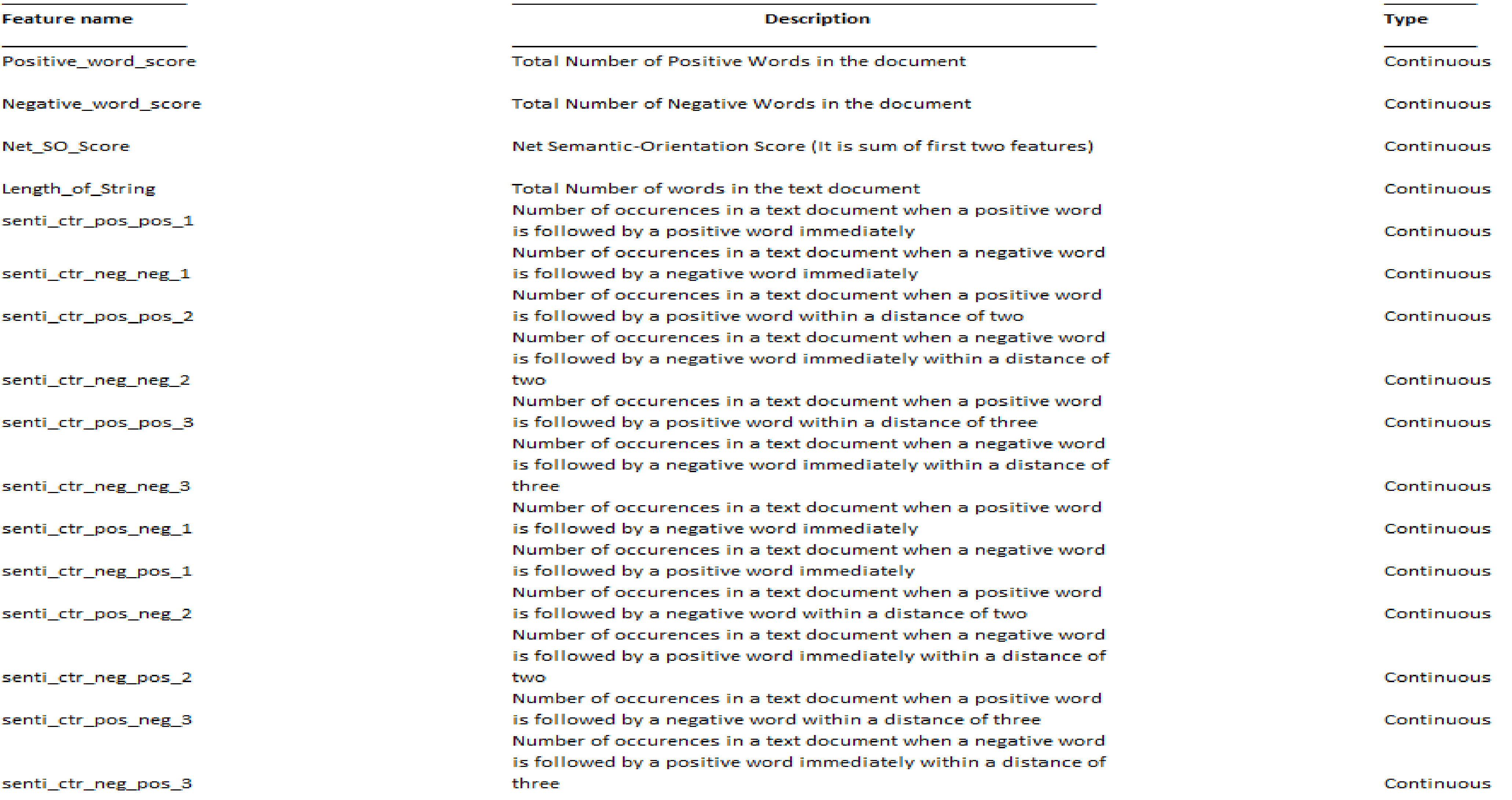
n-gram Based Features

Combining Different Features

1. Semantic Orientation (Neg-Pos) Features
2. POS Features
3. Composite Numerical Feature = Neg_Pos + POS
4. Unigram Features
5. Unigram + Bigram Features
6. Unigram + Bigram +Trigram Features
7. Unigram + Composite Numerical Features
8. Unigram + Bigram +Trigram + Composite Numerical Features

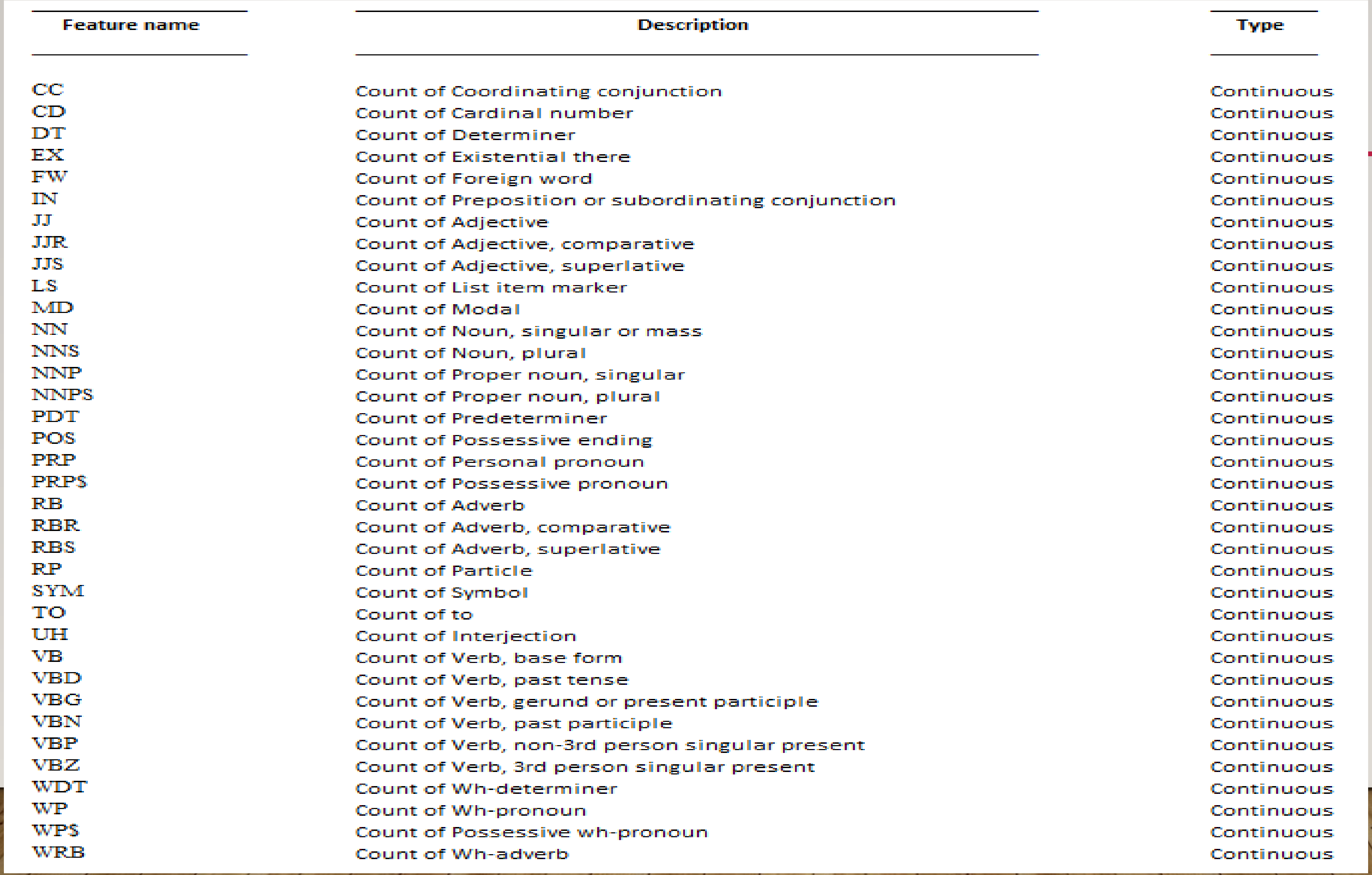
All Types of Features

SEMANTIC AND SYNTACTIC FEATURES



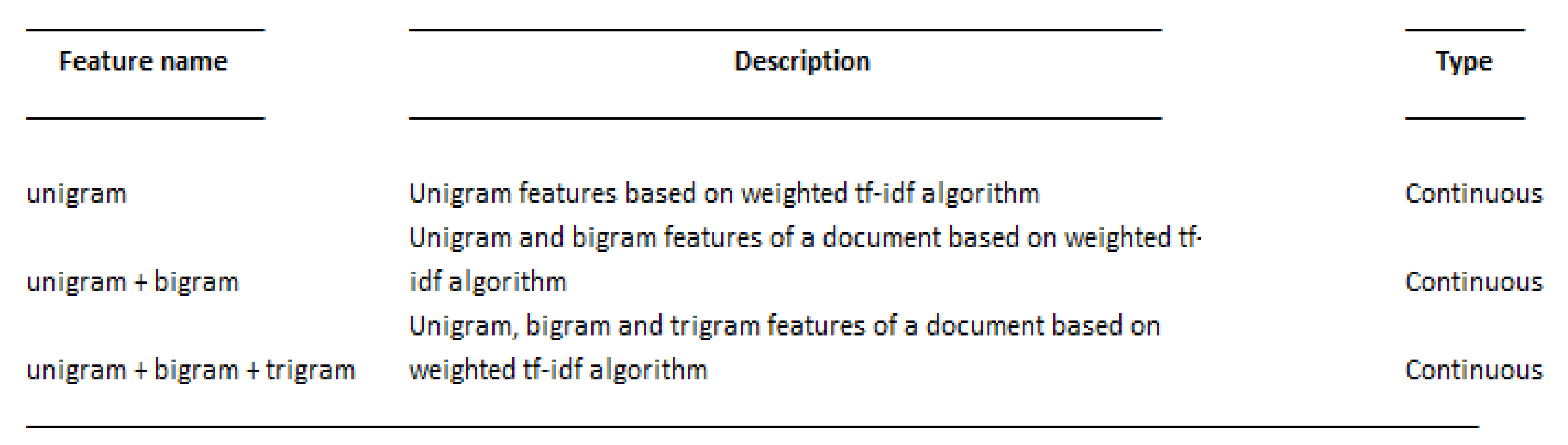
Feature name	Description	Type
Positive_word_score	Total Number of Positive Words in the document	Continuous
Negative_word_score	Total Number of Negative Words in the document	Continuous
Net_SO_Score	Net Semantic-Orientation Score (It is sum of first two features)	Continuous
Length_of_String	Total Number of words in the text document	Continuous
senti_ctr_pos_pos_1	Number of occurences in a text document when a positive word is followed by a positive word immediately	Continuous
senti_ctr_neg_neg_1	Number of occurences in a text document when a negative word is followed by a negative word immediately	Continuous
senti_ctr_pos_pos_2	Number of occurences in a text document when a positive word is followed by a positive word within a distance of two	Continuous
senti_ctr_neg_neg_2	Number of occurences in a text document when a negative word is followed by a negative word immediately within a distance of two	Continuous
senti_ctr_pos_pos_3	Number of occurences in a text document when a positive word is followed by a positive word within a distance of three	Continuous
senti_ctr_neg_neg_3	Number of occurences in a text document when a negative word is followed by a negative word immediately within a distance of three	Continuous
senti_ctr_pos_neg_1	Number of occurences in a text document when a positive word is followed by a negative word immediately	Continuous
senti_ctr_neg_pos_1	Number of occurences in a text document when a negative word is followed by a positive word immediately	Continuous
senti_ctr_pos_neg_2	Number of occurences in a text document when a positive word is followed by a negative word within a distance of two	Continuous
senti_ctr_neg_pos_2	Number of occurences in a text document when a negative word is followed by a positive word immediately within a distance of two	Continuous
senti_ctr_pos_neg_3	Number of occurences in a text document when a positive word is followed by a negative word within a distance of three	Continuous
senti_ctr_neg_pos_3	Number of occurences in a text document when a negative word is followed by a positive word immediately within a distance of three	Continuous

PART-OF-SPEECH BASED FEATURES



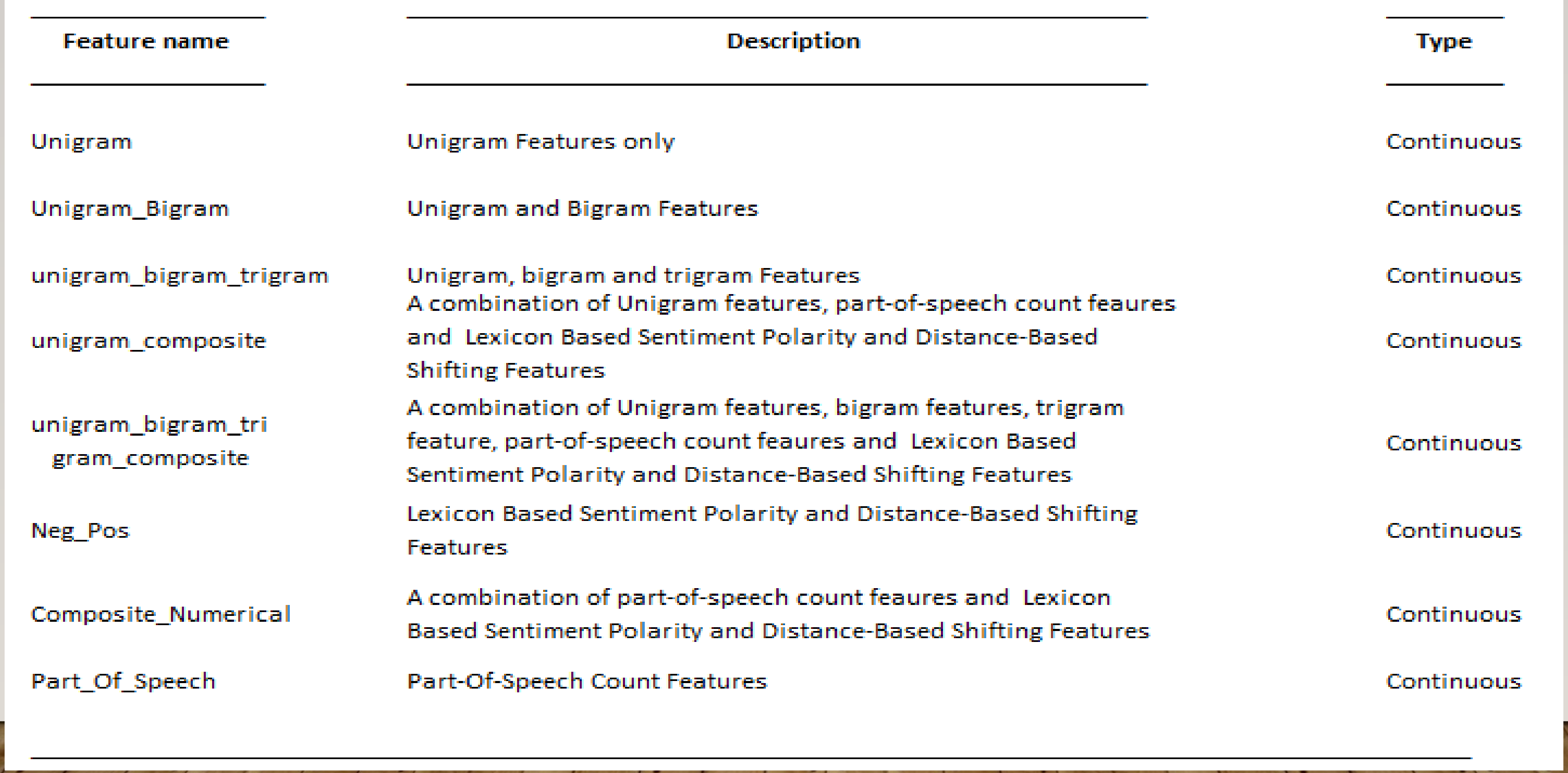
Feature name	Description	Type
CC	Count of Coordinating conjunction	Continuous
CD	Count of Cardinal number	Continuous
DT	Count of Determiner	Continuous
EX	Count of Existential there	Continuous
FW	Count of Foreign word	Continuous
IN	Count of Preposition or subordinating conjunction	Continuous
JJ	Count of Adjective	Continuous
JJR	Count of Adjective, comparative	Continuous
JJS	Count of Adjective, superlative	Continuous
LS	Count of List item marker	Continuous
MD	Count of Modal	Continuous
NN	Count of Noun, singular or mass	Continuous
NNS	Count of Noun, plural	Continuous
NNP	Count of Proper noun, singular	Continuous
NNPS	Count of Proper noun, plural	Continuous
PDT	Count of Predeterminer	Continuous
POS	Count of Possessive ending	Continuous
PRP	Count of Personal pronoun	Continuous
PRP$	Count of Possessive pronoun	Continuous
RB	Count of Adverb	Continuous
RBR	Count of Adverb, comparative	Continuous
RBS	Count of Adverb, superlative	Continuous
RP	Count of Particle	Continuous
SYM	Count of Symbol	Continuous
TO	Count of to	Continuous
UH	Count of Interjection	Continuous
VB	Count of Verb, base form	Continuous
VBD	Count of Verb, past tense	Continuous
VBG	Count of Verb, gerund or present participle	Continuous
VBN	Count of Verb, past participle	Continuous
VBP	Count of Verb, non-3rd person singular present	Continuous
VBZ	Count of Verb, 3rd person singular present	Continuous
WDT	Count of Wh-determiner	Continuous
WP	Count of Wh-pronoun	Continuous
WP$	Count of Possessive wh-pronoun	Continuous
WRB	Count of Wh-adverb	Continuous

N-GRAM FEAFTURES



Feature name	Description	Type
unigram	Unigram features based on weighted tf-idf algorithm	Continuous
unigram + bigram	Unigram and bigram features of a document based on weighted tf-idf algorithm	Continuous
unigram + bigram + trigram	Unigram, bigram and trigram features of a document based on weighted tf-idf algorithm	Continuous

FINAL FEATURES SET



Feature name	Description	Type
Unigram	Unigram Features only	Continuous
Unigram_Bigram	Unigram and Bigram Features	Continuous
unigram_bigram_trigram	Unigram, bigram and trigram Features	Continuous
unigram_composite	A combination of Unigram features, part-of-speech count feaures and Lexicon Based Sentiment Polarity and Distance-Based Shifting Features	Continuous
unigram_bigram_trigram_composite	A combination of Unigram features, bigram features, trigram feature, part-of-speech count feaures and Lexicon Based Sentiment Polarity and Distance-Based Shifting Features	Continuous
Neg_Pos	Lexicon Based Sentiment Polarity and Distance-Based Shifting Features	Continuous
Composite_Numerical	A combination of part-of-speech count feaures and Lexicon Based Sentiment Polarity and Distance-Based Shifting Features	Continuous
Part_Of_Speech	Part-Of-Speech Count Features	Continuous

MODEL BUILDING

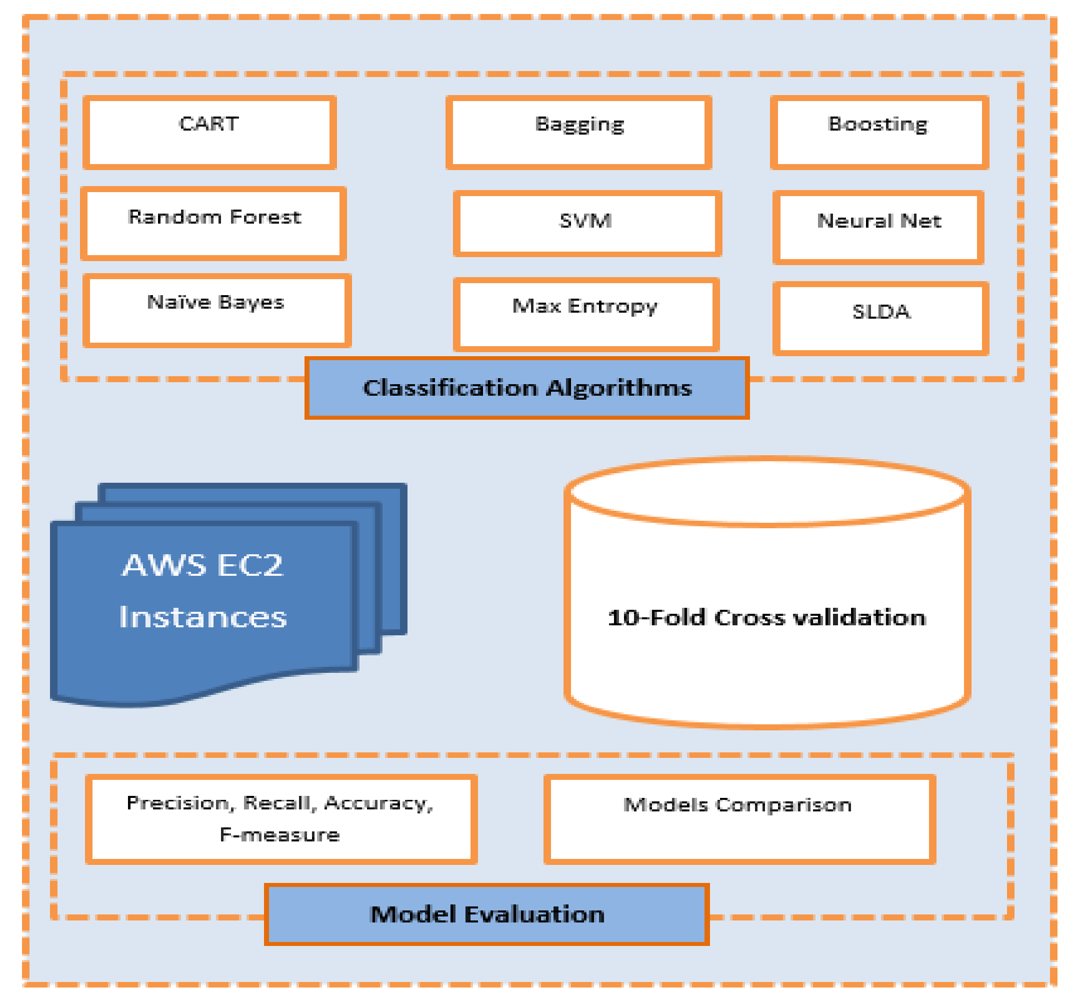
Multiple instances of EC2 were created to build multiple models simultaneously.

10-fold cross validation is performed to avoid overfitting of the model

All algorithms were run on each dataset.

Model output for each algorithm generates:

- Accuracy
- Precision
- Recall
- F-score



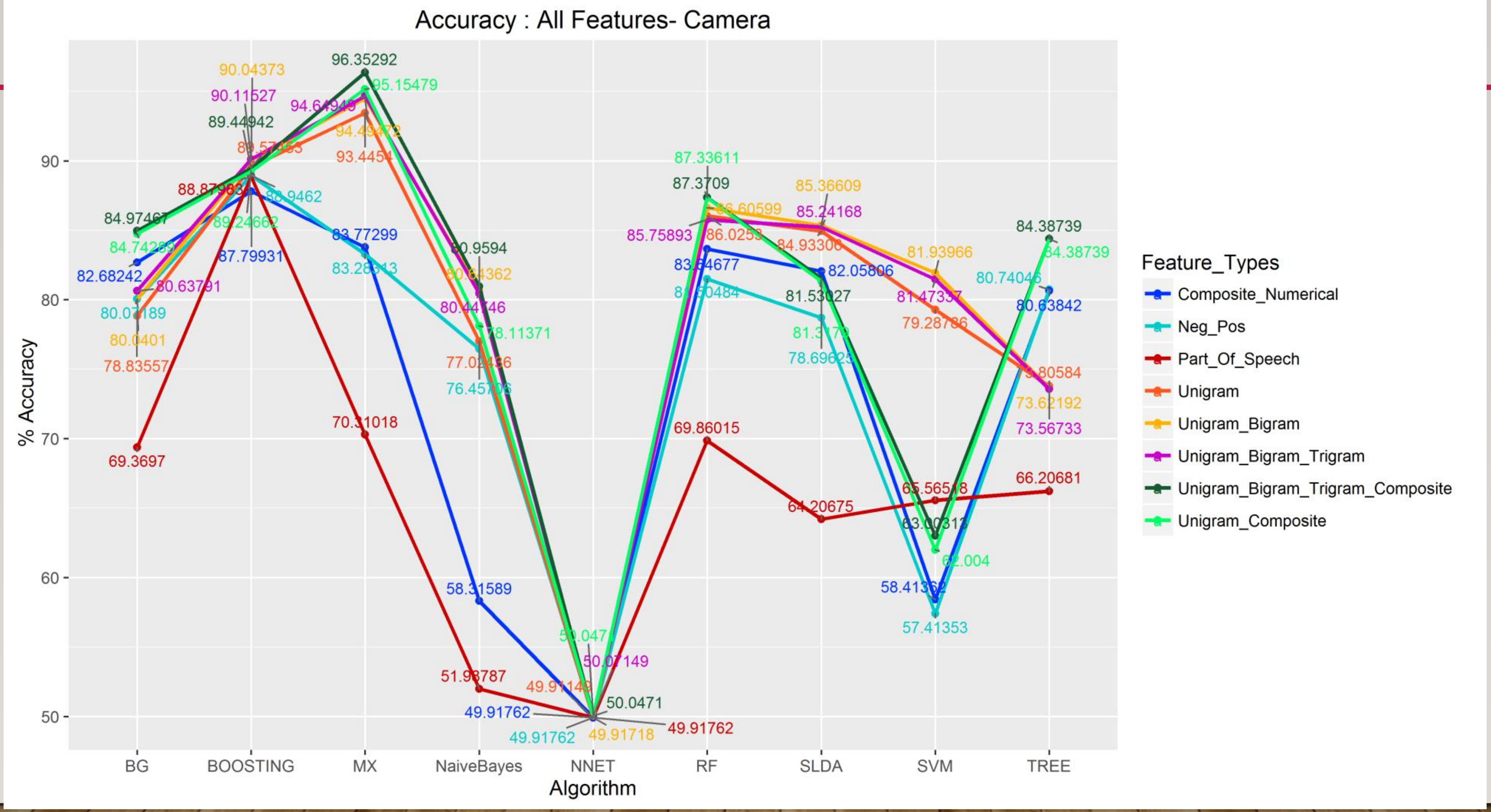
MODEL EVALUATION

- To evaluate the performance of different algorithms. Prediction of classification problems are primarily evaluated by four key measures as
  - **Precision**
  - **Recall**
  - **Accuracy**
  - **F-measure**

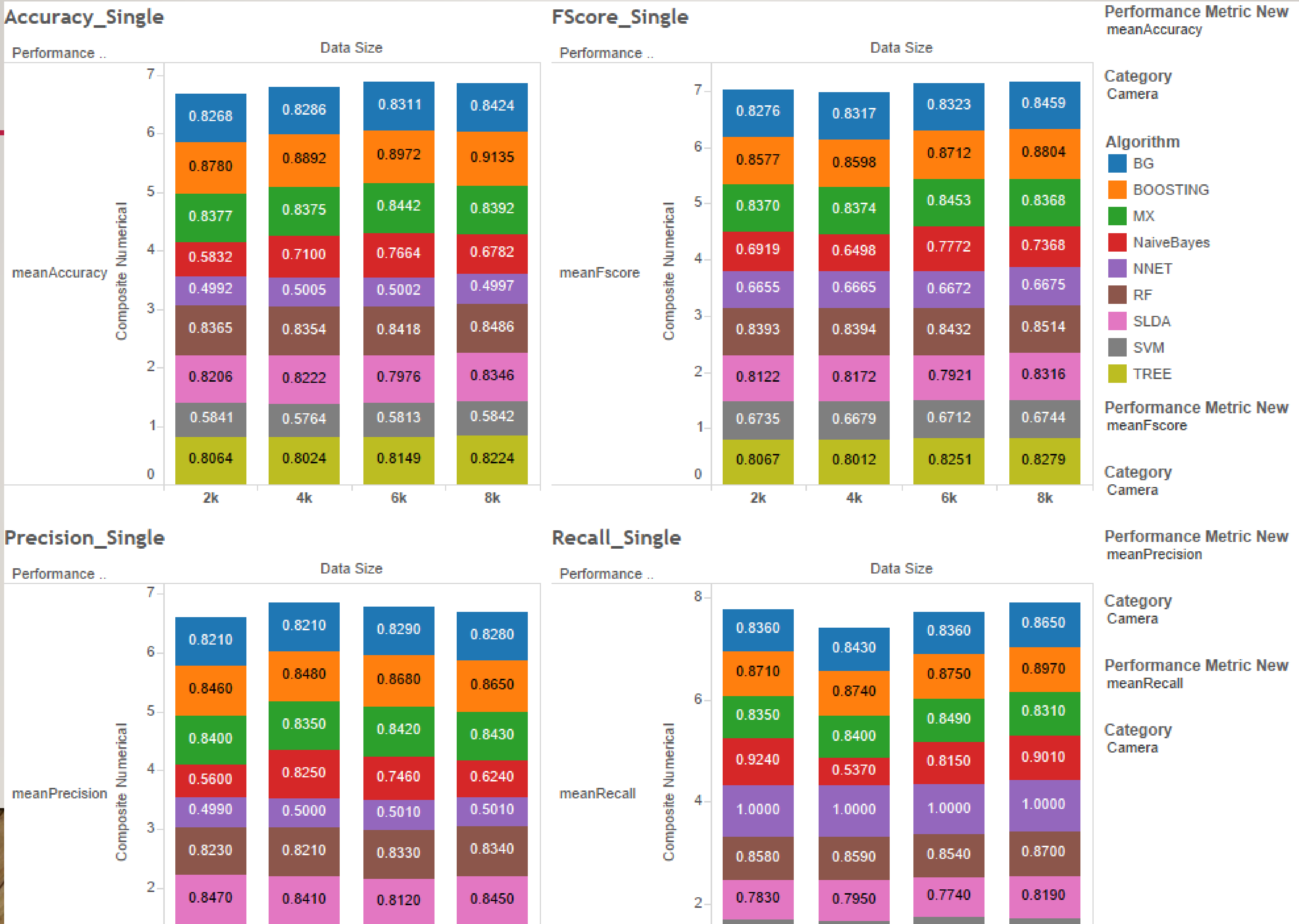
EXPERIMENT RESULTS

- Accuracy
  - **Unigram-Bigram- Trigram-Composite** feature with Maximum Entropy has achieved highest accuracy of 96.367% among all other features in all categories of dataset
  - Boosting algorithm has achieved **highest** accuracy and F-score with composite-numerical features.
  - Neural net has performed worst in all cases because of only 1 hidden layer.
  - Higher accuracy is achieved as the size of the dataset is increased in all the categories
  - All algorithms except Neural Net and Naïve Bayes have achieved accuracy of **more than 80%** across all categories and dataset sizes.
  - Recall and Precision has also improved as the size of dataset is increased in 4 out of 6 categories.

COMPARISON OF ALGORITHMS AND DATA SIZE



COMPARISON OF ALGORITHMS AND DATA SIZE



CONCLUSION AND FUTURE WORK

- High accuracy is achieved in the Sentiment Classification task on text data using
  - Efficient Feature engineering
  - Classic Supervised Machine Learning Algorithms
  - Large size datasets
- In future, more stress would be given on
  - the importance of each feature in the feature set
  - State-of the feature engineering frameworks and algorithms

LEARNING AND CHALLENGES

- Studied new and classic avenues of natural language processing
- Used latest visualization techniques to understand the findings
- Telecom-377 was a savior while running multiple instances on EC2
- Learning cloud technology like EC2 for large scale data processing was a scary task initially

THANK
YOU

Q
&
A