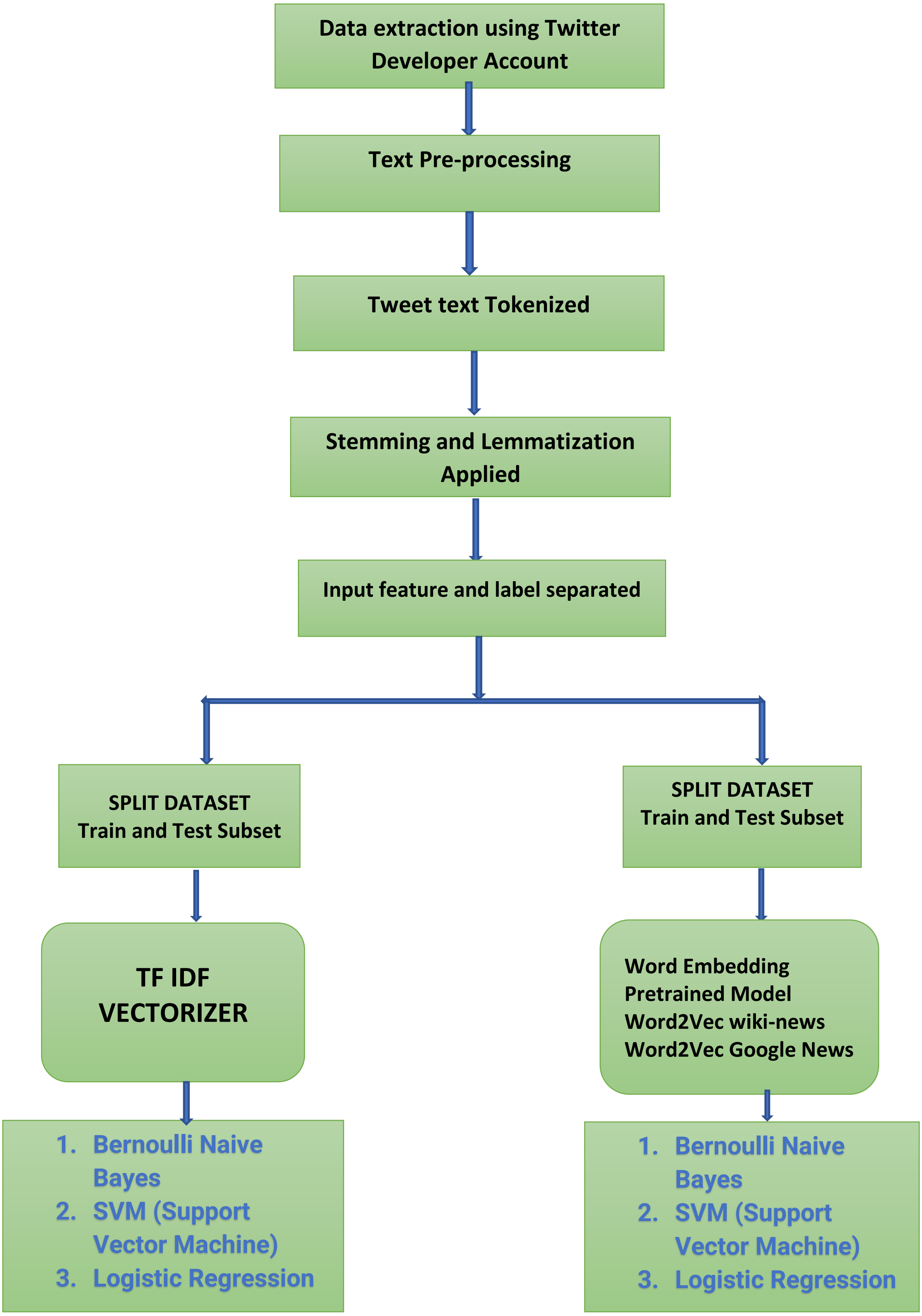
Natural Language Processing Project: Write Up.

Machine Learning on Text Data

Name: Ramya Mercy Rajan
Roll number: AA.SC.P2MCA2107434

Natural Language Processing Project: Machine Learning on Text Data.


Data extraction using Twitter Developer Account
Text Pre-processing
Tweet text Tokenized
Stemming and Lemmatization Applied
Input feature and label separated
SPLIT DATASET Train and Test Subset
SPLIT DATASET Train and Test Subset
TF IDF VECTORIZER
Word Embedding Pretrained Model Word2Vec wiki-news Word2Vec Google News
1. Bernoulli Naive Bayes 2. SVM (Support Vector Machine) 3. Logistic Regression
1. Bernoulli Naive Bayes 2. SVM (Support Vector Machine) 3. Logistic Regression

Name: Ramya Mercy Rajan
Roll number: AA.SC.P2MCA2107434

Gathering Data

The raw Twitter tweets were obtained with the help of creating a twitter developer account. For most text mining or classification projects, extracting tweets is one of the most important and initial steps. The well-known method is extracting tweets with tweepy and creating a developer account in twitter.

This project includes extraction of tweets from twitter and have performed pre-processing and text representation. Machine Learning Algorithms have been applied on it.
Length of dataset is 13192
Tweets datasets consist of 13192 rows × 3 columns
Columns consist of id, text and label.
Labels are those 5 classes (labels) selected and extracted tweets from twitter.

The 5 classes include

1. Digital electronics
2. Robotics
3. Artificial Intelligence
4. Computer Vision
5. Bioinformatics

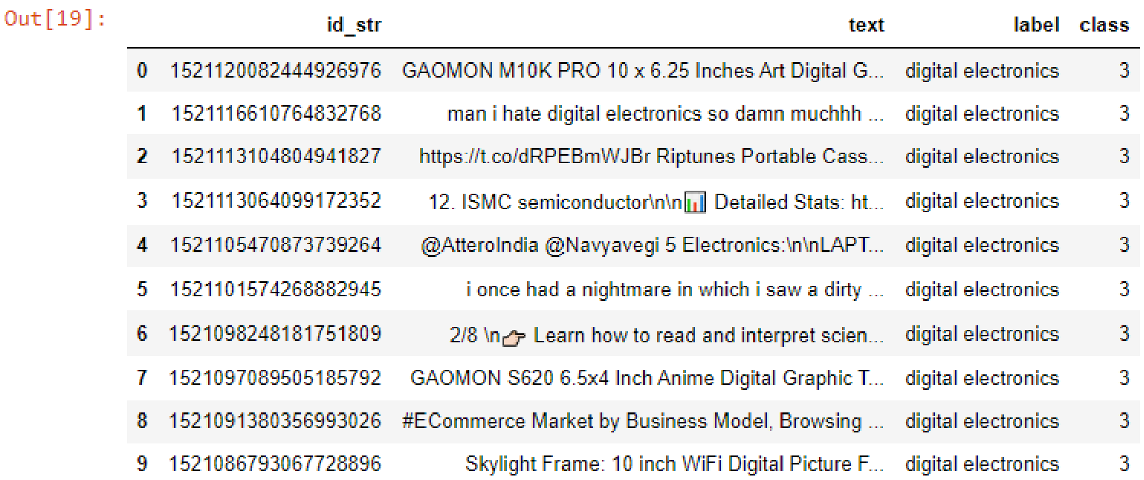
Raw tweets contain some special characters and symbols that the model will not be able to process, hence the data needs to be cleaned before it can be passed to our model. This task and much of the subsequent tasks were undertaken with significant guidance from the demo and live recorded lectures of my respected Lecturer Ms. Jayashree.

Pre-processing Tweets

1. The basic rules we used to pre-process the tweets were:
2. The stopwords was cleaned and removed from the tweets text.
3. URL's, punctuations, numeric numbers were removed as a part of Preprocessing stage.
4. Cleaning and removing punctuations
5. Cleaning and removing Numeric numbers

Most of the pre-processing steps were done with the help of a python library called Regular Expression operator.

Label Encoder was used to convert my target class to numbers.



Out[19]:

	id_str	text	label	class
0	1521120082444926976	GAOMON M10K PRO 10 x 6.25 Inches Art Digital G...	digital electronics	3
1	1521116610764832768	man i hate digital electronics so damn muchhh ...	digital electronics	3
2	1521113104804941827	https://t.co/dRPEBmWJBr Riptunes Portable Cass...	digital electronics	3
3	1521113064099172352	12. ISMC semiconductor\n\n📊 Detailed Stats: ht...	digital electronics	3
4	1521105470873739264	@AtteroIndia @Navyavegi 5 Electronics:\n\nLAPT...	digital electronics	3
5	1521101574268882945	i once had a nightmare in which i saw a dirty ...	digital electronics	3
6	1521098248181751809	2/8 \n👉 Learn how to read and interpret scien...	digital electronics	3
7	1521097089505185792	GAOMON S620 6.5x4 Inch Anime Digital Graphic T...	digital electronics	3
8	1521091380356993026	#ECommerce Market by Business Model, Browsing ...	digital electronics	3
9	1521086793067728896	Skylight Frame: 10 inch WiFi Digital Picture F...	digital electronics	3

Name: Ramya Mercy Rajan
Roll number: AA.SC.P2MCA2107434

1. Tweets text were tokenized, stemming and Lemmatization was applied.
2. Then the input feature and label were separated
3. Plotted a cloud of words for negative tweets.
4. Data was splitted into Train and Test subset
5. Data was transformed using TF-IDF Vectorizer.
6. Word2vec is another technique/model used to produce word embedding. Pretrained models were used for word embeddings

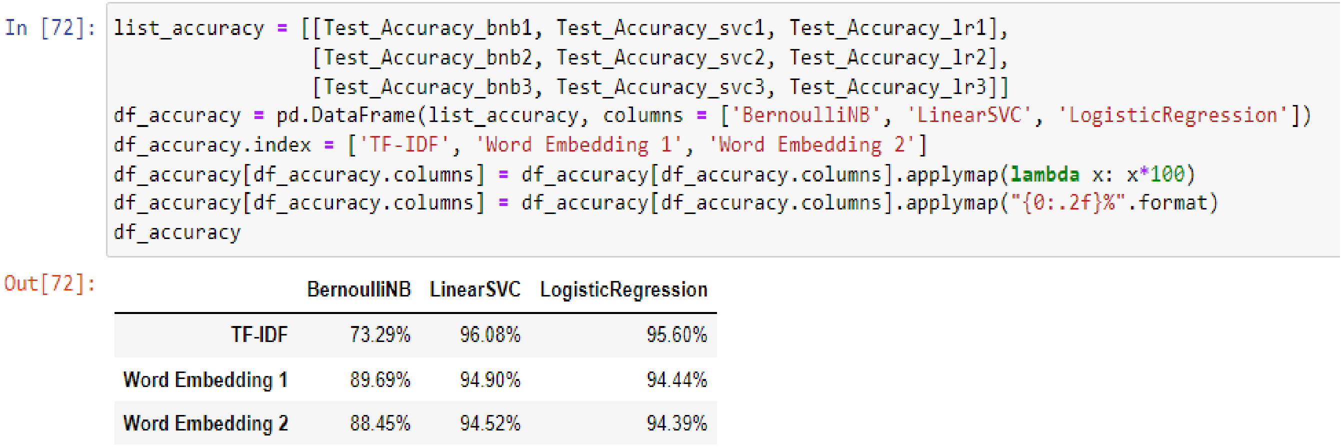
Classification Techniques

I have used 3 different models respectively:

1. Bernoulli Naive Bayes
2. SVM (Support Vector Machine)
3. Logistic Regression

The idea behind choosing these models is that I want to try all the classifiers on the dataset ranging from simple ones to complex models and then try to find out the one which gives the best performance among them.

A confusion matrix a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1score.



```
In [72]: list_accuracy = [[Test_Accuracy_bnb1, Test_Accuracy_svc1, Test_Accuracy_lr1],
                          [Test_Accuracy_bnb2, Test_Accuracy_svc2, Test_Accuracy_lr2],
                          [Test_Accuracy_bnb3, Test_Accuracy_svc3, Test_Accuracy_lr3]]
         df_accuracy = pd.DataFrame(list_accuracy, columns = ['BernoulliNB', 'LinearSVC', 'LogisticRegression'])
         df_accuracy.index = ['TF-IDF', 'Word Embedding 1', 'Word Embedding 2']
         df_accuracy[df_accuracy.columns] = df_accuracy[df_accuracy.columns].applymap(lambda x: x*100)
         df_accuracy[df_accuracy.columns] = df_accuracy[df_accuracy.columns].applymap("{0:.2f}%".format)
         df_accuracy
```

Out[72]:

	BernoulliNB	LinearSVC	LogisticRegression
TF-IDF	73.29%	96.08%	95.60%
Word Embedding 1	89.69%	94.90%	94.44%
Word Embedding 2	88.45%	94.52%	94.39%

Conclusion

The models Linear SVC and Logistic Regression have high and comparable accuracies with the training tweets that were used for testing. Considering the comparable accuracies, all these models are giving with the training tweets, these models would very likely give comparable accuracies with a much larger number of new tweets.

Name: Ramya Mercy Rajan
Roll number: AA.SC.P2MCA2107434