IS 688-WEB MINING PROJECT

Analysis on Twitter Data

Priyanka Bongale
Sachin Mathew Jose
Chaitanya Shah
Pushkar Gadgil

Agenda

- What is Sentiment Analysis
- Data Cleaning Process
- Word Clouds : Positive and Negative Sentiments
- Naive Bayes Model: Multinomial and Bernoulli
- Logistic Regression Model
- SGDClassifier Model
- LinearSVC Model
- Random Forest Classifier Model
- MLPClassifier Model
- Most Efficient Model
- Challenges

Objective

- The objective of the project is to analyze the twitter data to predict the positive and negative sentiments in tweets

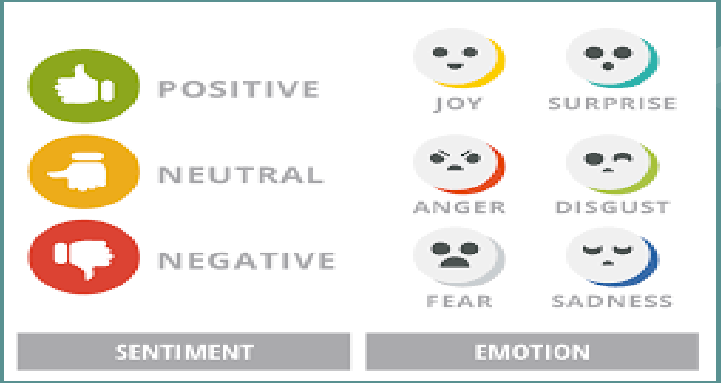
- To prepare and train a model based on Logistic Regression, Naive Bayes classifier, SVM, Neural Network, Random Forest.

- To compare these models to determine which model has the best accuracy results

What is Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping to understand the social sentiment of a brand, product or service while monitoring online conversations

Sentiment is often framed as a binary distinction (positive vs. negative), but it can also be a more fine-grained, like identifying the specific emotion an author is expressing (like fear, joy or anger).



Our approach to sentiment analysis (Bag of Words)

- Clean the data
- Remove stop words.
- Create a list of words and their frequencies.
- Create bigrams and their frequencies.
- Select the top 5k features from the the above two lists.
- Vectorize the sentence using these features.
- Randomly select train and test data (80:20).
- Train and test 7 model using these features.
- Repeat 10 times using another randomly selected data
- AVerage the results obtained in all the 10 iterations

Data Set

1. Sentiments
2. ID
3. Date
4. Username
5. Tweets

	A	B	C	D	E	F
1	0	1467810369	Mon Apr 06 22:19:45 PDT	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpi
2	0	1467810672	Mon Apr 06 22:19:49 PDT	NO_QUERY	scotthamilton	is upset that he can't upda
3	0	1467810917	Mon Apr 06 22:19:53 PDT	NO_QUERY	mattycus	@Kenichan I dived many t
4	0	1467811184	Mon Apr 06 22:19:57 PDT	NO_QUERY	ElleCTF	my whole body feels itchy
5	0	1467811193	Mon Apr 06 22:19:57 PDT	NO_QUERY	Karoli	@nationwideclass no, it's
6	0	1467811372	Mon Apr 06 22:20:00 PDT	NO_QUERY	joy_wolf	@Kwesidei not the whole
7	0	1467811592	Mon Apr 06 22:20:03 PDT	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03 PDT	NO_QUERY	coZZ	@LOLTrish hey long time
9	0	1467811795	Mon Apr 06 22:20:05 PDT	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they dic
10	0	1467812025	Mon Apr 06 22:20:09 PDT	NO_QUERY	mimismo	@twittera que me muera
11	0	1467812416	Mon Apr 06 22:20:16 PDT	NO_QUERY	erinx3leannexo	spring break in plain city...
12	0	1467812579	Mon Apr 06 22:20:17 PDT	NO_QUERY	pardonlauren	I just re-pierced my ears
13	0	1467812723	Mon Apr 06 22:20:19 PDT	NO_QUERY	TLeC	@caregiving I couldn't bea
14	0	1467812771	Mon Apr 06 22:20:19 PDT	NO_QUERY	robrobbierobert	@octolinz16 It it counts, i
15	0	1467812784	Mon Apr 06 22:20:20 PDT	NO_QUERY	bayofwolves	@smarrison i would've be
16	0	1467812799	Mon Apr 06 22:20:20 PDT	NO_QUERY	HairByJess	@iamjazzyfizzle I wish I g
17	0	1467812964	Mon Apr 06 22:20:22 PDT	NO_QUERY	lovesongwriter	Hollis' death scene will hu
18	0	1467813137	Mon Apr 06 22:20:25 PDT	NO_QUERY	armotley	about to file taxes
19	0	1467813579	Mon Apr 06 22:20:31 PDT	NO_QUERY	starkissed	@LettyA ahh ive always w
20	0	1467813782	Mon Apr 06 22:20:34 PDT	NO_QUERY	gi_gi_bee	@FakerPattyPattz Oh dea
21	0	1467813985	Mon Apr 06 22:20:37 PDT	NO_QUERY	quanvu	@alydesigns i was out mc
22	0	1467813992	Mon Apr 06 22:20:38 PDT	NO_QUERY	swinspeedx	one of my friend called m
23	0	1467814119	Mon Apr 06 22:20:40 PDT	NO_QUERY	cooliodoc	@angry_barista I baked yc
24	0	1467814180	Mon Apr 06 22:20:40 PDT	NO_QUERY	viJILLante	this week is not going as i
25	0	1467814192	Mon Apr 06 22:20:41 PDT	NO_QUERY	Ljelli3166	blagh class at 8 tomorrow
26	0	1467814438	Mon Apr 06 22:20:44 PDT	NO_QUERY	ChicagoCubbie	I hate when I have to call
27	0	1467814783	Mon Apr 06 22:20:50 PDT	NO_QUERY	KatieAngell	Just going to cry myself to
28	0	1467814883	Mon Apr 06 22:20:52 PDT	NO_QUERY	gagoo	im sad now Miss.Lilly
29	0	1467815199	Mon Apr 06 22:20:56 PDT	NO_QUERY	abel209	ooooh.... LOL that leslie...
30	0	1467815753	Mon Apr 06 22:21:04 PDT	NO_QUERY	BaptisteTheFool	Meh... Almost Lover is the
31	0	1467815923	Mon Apr 06 22:21:07 PDT	NO_QUERY	fatkat309	some1 hacked my account
32	0	1467815924	Mon Apr 06 22:21:07 PDT	NO_QUERY	EmCDL	@alielayus I want to go to
33	0	1467815988	Mon Apr 06 22:21:09 PDT	NO_QUERY	merisssa	thought sleeping in was a
34	0	1467816149	Mon Apr 06 22:21:11 PDT	NO_QUERY	Pbearfox	@julieebaby awe i love yo
35	0	1467816665	Mon Apr 06 22:21:21 PDT	NO_QUERY	jsoo	@HumpNinja I cry my asia
36	0	1467816749	Mon Apr 06 22:21:20 PDT	NO_QUERY	scarletletterm	ok I'm sick and spent an h
37	0	1467817225	Mon Apr 06 22:21:27 PDT	NO_QUERY	crosland_12	@cocomix04 ill tell ya the
38	0	1467817374	Mon Apr 06 22:21:30 PDT	NO_QUERY	ajaxpro	@MissXu sorry! bed time
39	0	1467817502	Mon Apr 06 22:21:32 PDT	NO_QUERY	Tmttq86	@fleurylis I don't either. I
40	0	1467818007	Mon Apr 06 22:21:39 PDT	NO_QUERY	Anthony_Nguyen	Bed. Class 8-12. Work 12-

- Remove xml encoding
- Remove links with ‘http://’ and ‘www.’
- Converted words into lower case
- Change words like ‘isn’t’ to ‘is not’
- Remove utf-8 encoded signs
- Removed Special characters
- Eliminated Digits
- Eliminate unnecessary spaces
- Scrapped Stopwords

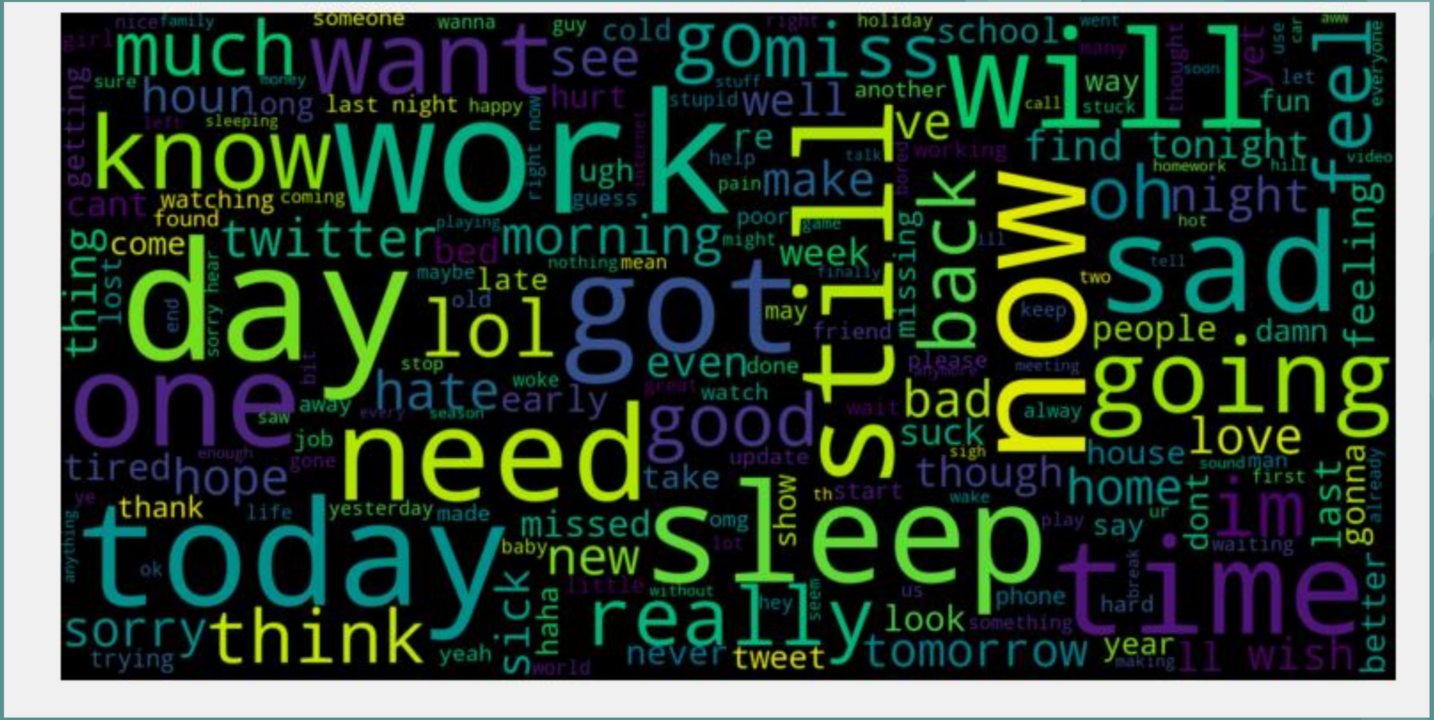
Data after Cleaning

tweet	target
awww that s a bummer you shoulda got david carr of third day to do it d	0
is upset that he can not update his facebook by texting it and might cry as a result school today also blah	0
i dived many times for the ball managed to save the rest go out of bounds	0
my whole body feels itchy and like its on fire	0
no it s not behaving at all i m mad why am i here because i can not see you all over there	0
not the whole crew	0
need a hug	0
hey long time no see yes rains a bit only a bit lol i m fine thanks how s you	0
k nope they did not have it	0
que me muera	0
spring break in plain city it s snowing	0
i just re pierced my ears	0
i could not bear to watch it and i thought the ua loss was embarrassing	0
it it counts idk why i did either you never talk to me anymore	0
i would ve been the first but i did not have a gun not really though zac snyder s just a doucheclown	0
i wish i got to watch it with you i miss you and how was the premiere	0
hollis death scene will hurt me severely to watch on film wry is directors cut not out now	0
about to file taxes	0
ahh ive always wanted to see rent love the soundtrack	0
oh dear were you drinking out of the forgotten table drinks	0
i was out most of the day so did not get much done	0
one of my friend called me and asked to meet with her at mid valley today but i ve no time sigh	0
barista i baked you a cake but i ated it	0
this week is not going as i had hoped	0
blagh class at tomorrow	0
i hate when i have to call and wake people up	0
just going to cry myself to sleep after watching marley and me	0
im sad now miss lilly	0
ooooh lol that leslie and ok i will not do it again so leslie will not get mad again	0
meh almost lover is the exception this track gets me depressed every time	0

6000 Tweets

Word Cloud: Positive Sentiments

Word Cloud: Negative Sentiments



Multinomial Naiive Bayes

- Naive Bayes classifier for multinomial models
- This is suitable for classification with discrete features (e.g., word counts for text classification)

MultinomialNB (*alpha=1.0, fit_prior=True, class_prior=None*)

```
MultinomialNB
--------------------------------
        Avg. Accuracy: 73.08%
        Avg. F1 Score: 72.93
        Avg. precision Score: 73.11
        Avg. recall Score: 73.08
        Avg. Confusion Matrix:

[[441.6 157.2]
 [165.8 435.4]]
```

Bernoulli Naiive Bayes

- Performs better on datasets, especially those with shorter documents.
- This implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions.
- This requires samples to be represented as binary-valued feature vectors.

BernoulliNB (*alpha=1.0, binarize=0.0, fit_prior=True, class_prior=None*)

```
BernoulliNB
------------------------------
        Avg. Accuracy: 72.97%
        Avg. F1 Score: 74.32
        Avg. precision Score: 73.23
        Avg. recall Score: 72.97
        Avg. Confusion Matrix:

[[406.4 192.4]
 [131.9 469.3]]
```

LogisticRegression Model

- Linear model for classification rather than regression.
- Probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

```
LogisticRegression(penalty='l2', dual=False, tol=0.0001,
                   C=1.0, fit_intercept=True, inter-
                   cept_scaling=1, class_weight=None,
                   random_state=None, solver='warn',
                   max_iter=100, multi_class='warn', ver-
                   bose=0, warm_start=False, n_jobs=None,
                   l1_ratio=None)
```

```
LogisticRegression
-----------------------------------
        Avg. Accuracy: 73.13%
        Avg. F1 Score: 73.66
        Avg. precision Score: 73.18
        Avg. recall Score: 73.13
        Avg. Confusion Matrix:

[[426.5 172.3]
 [150.1 451.1]]
```

SGDClassifier Model

- This implements a regularised linear models with Stochastic Gradient Descent learning routine which supports different loss functions and penalties for classification.

```
SGDClassifier (loss='hinge', penalty='l2', alpha=0.0001,
               l1_ratio=0.15, fit_intercept=True, max_iter=None,
               tol=None, shuffle=True, verbose=0, epsilon=0.1,
               n_jobs=None, random_state=None, learn-
               ing_rate='optimal', eta0=0.0, power_t=0.5,
               early_stopping=False, validation_fraction=0.1,
               n_iter_no_change=5, class_weight=None,
               warm_start=False, average=False, n_iter=None)
```

```
SGDClassifier
-----------------------------------
        Avg. Accuracy: 69.31%
        Avg. F1 Score: 69.32
        Avg. precision Score: 69.46
        Avg. recall Score: 69.31
        Avg. Confusion Matrix:

[[413.6 185.2]
 [183.1 418.1]]
```

LinearSVC Model

- The linear-SVC uses a linear kernel for the basis function
- This class supports both dense and sparse input and the multiclass support is handled according to a one vs the rest scheme.

LinearSVC (*penalty='l2', loss='squared_hinge', dual=True, tol=0.0001, C=1.0, multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000*)

```
LinearSVC
--------------------------------
        Avg. Accuracy: 70.33%
        Avg. F1 Score: 70.98
        Avg. precision Score: 70.39
        Avg. recall Score: 70.33
        Avg. Confusion Matrix:

[[408.5 190.3]
 [165.7 435.5]]
```

RandomForestClassifier Model

- A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- When splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features.

RandomForestClassifier (*n_estimators='warn', criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None*)

```
RandomForestClassifier
--------------------------------
        Avg. Accuracy: 68.11%
        Avg. F1 Score: 67.12
        Avg. precision Score: 68.19
        Avg. recall Score: 68.11
        Avg. Confusion Matrix:

[[426.1 172.7]
 [210.  391.2]]
```

MLPClassifier Model

- It trains iteratively since at each time step the partial derivatives of the loss function with respect to the model parameters are computed to update the parameters.
- This implementation works with data represented as dense numpy arrays or sparse scipy arrays of floating point values.

MLPClassifier (*hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10*)

```
MLPClassifier
------------------------------
        Avg. Accuracy: 69.84%
        Avg. F1 Score: 70.32
        Avg. precision Score: 69.88
        Avg. recall Score: 69.84
        Avg. Confusion Matrix:

[[409.   189.8]
 [172.1 429.1]]
```

Most Efficient Models

```
LogisticRegression
------------------------------
        Avg. Accuracy: 73.13%
        Avg. F1 Score: 73.66
        Avg. precision Score: 73.18
        Avg. recall Score: 73.13
        Avg. Confusion Matrix:

[[426.5 172.3]
 [150.1 451.1]]
```

1. LogisticRegression

2. Multinomial NB

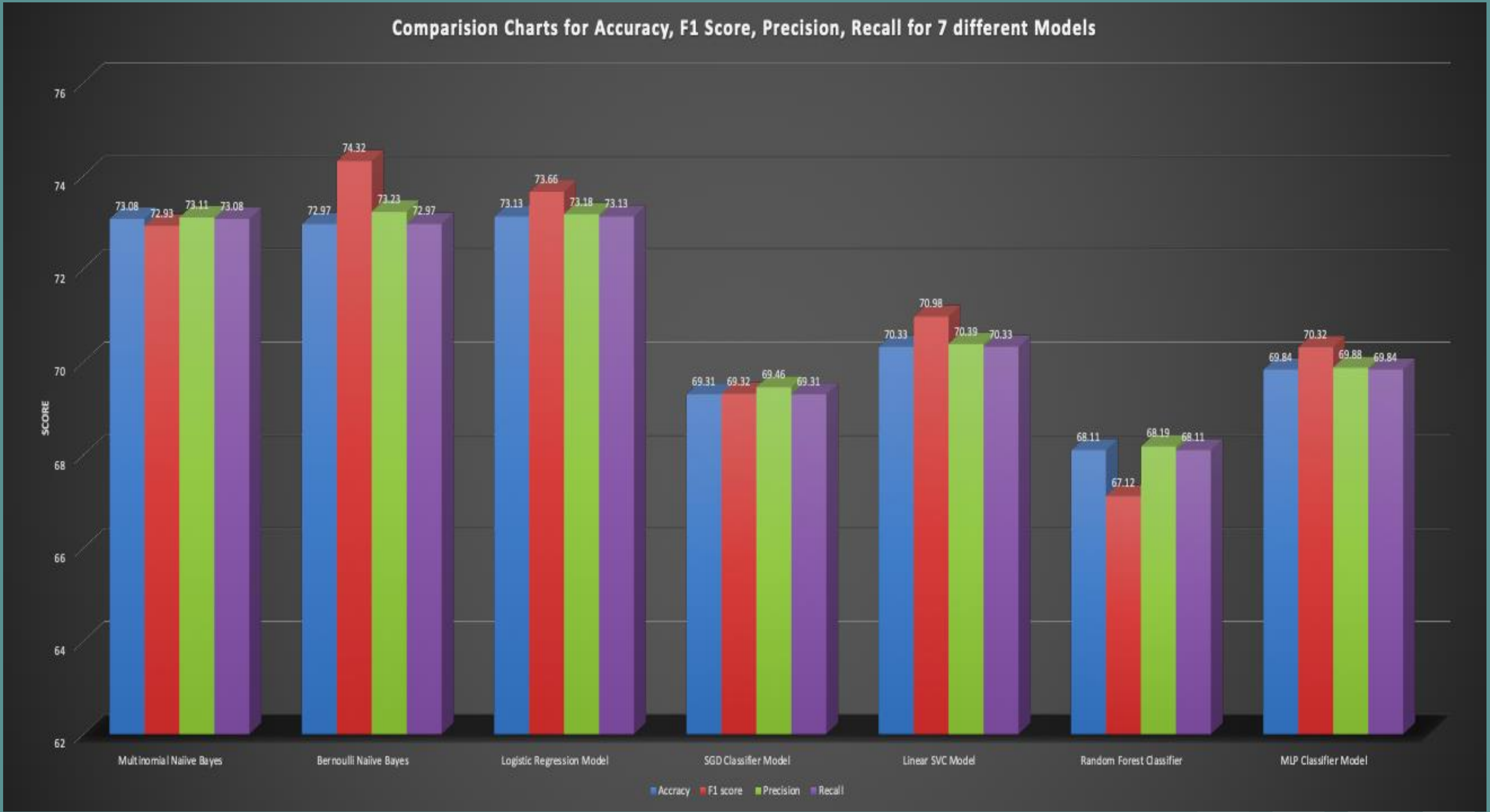
```
MultinomialNB
------------------------------
        Avg. Accuracy: 73.08%
        Avg. F1 Score: 72.93
        Avg. precision Score: 73.11
        Avg. recall Score: 73.08
        Avg. Confusion Matrix:

[[441.6 157.2]
 [165.8 435.4]]
```

3. Bernoulli NB

```
BernoulliNB
------------------------------
        Avg. Accuracy: 72.97%
        Avg. F1 Score: 74.32
        Avg. precision Score: 73.23
        Avg. recall Score: 72.97
        Avg. Confusion Matrix:

[[406.4 192.4]
 [131.9 469.3]]
```



Comparision Charts for Accuracy, F1 Score, Precision, Recall for 7 different Models
SCORE
76
74
72
70
68
66
64
62
73.08
72.93
73.11
73.08
72.97
74.32
73.23
72.97
73.13
73.66
73.18
73.13
69.31
69.32
69.46
69.31
70.33
70.98
70.39
70.33
68.11
67.12
68.19
68.11
69.84
70.32
69.88
69.84
Multinomial Naiive Bayes
Bernoulli Naiive Bayes
Logistic Regression Model
SGD Classifier Model
Linear SVC Model
Random Forest Classifier
MLP Classifier Model
Accracy
F1 score
Precision
Recall

Challenges

1. Data set is small so the accuracy will be low

1. Training is taking too much time for larger data

1. Using the right parameters for each models


рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
spas
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
thank you
gracias
хвала
asante
manana
kiitos
dankie
nanni
nandri
dhanyavad
maururu
bedankt
köszönöm
hvala
gràcie
bayarlalaa
akun dankon
aĉiū
obrigada
murakoze
tenki
enkosi
dziękuje
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
děkuji
sukriya
kop khun krap
taiku
obrigado
sagolun
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
kam sah hamnida
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
мерси
rahmat
xièxie
ευχαριστώ
merci
তোমাকে ধন্যবাদ
감사합니다