Twitter Sentiment Analysis

Data Preparation

Data is downloaded from
https://www.kaggle.com/kazanova/sentiment140

It contains 1,600,000 tweets.

It contains the following 6 fields:

1. target: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. ids: The id of the tweet (*2087*)
3. date: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
4. flag: The query (*lyx*). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (*robotickilldozr*)
6. text: the text of the tweet (*Lyx is cool*)

0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine thanks , how's you ?
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing . . . . .
0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore
0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really though, zac snyder's just a doucheclown.
0	1467812799	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	HairByJess	@iamjazzyfizzle I wish I got to watch it with you!! I miss you and @iamlilnicki how was the premiere?!
0	1467812964	Mon Apr 06 22:20:22 PDT 2009	NO_QUERY	lovesongwriter	Hollis' death scene will hurt me severely to watch on film wry is directors cut not out now?
0	1467813137	Mon Apr 06 22:20:25 PDT 2009	NO_QUERY	armotley	about to file taxes
0	1467813579	Mon Apr 06 22:20:31 PDT 2009	NO_QUERY	starkissed	@LettyA ahh ive always wanted to see rent love the soundtrack!!
0	1467813782	Mon Apr 06 22:20:34 PDT 2009	NO_QUERY	gi_gi_bee	@FakerPattyPattz Oh dear. Were you drinking out of the forgotten table drinks?
0	1467813985	Mon Apr 06 22:20:37 PDT 2009	NO_QUERY	quanvu	@alydesigns i was out most of the day so didn't get much done
0	1467813992	Mon Apr 06 22:20:38 PDT 2009	NO_QUERY	swinspeedx	one of my friend called me, and asked to meet with her at Mid Valley today...but i've no time *sigh*
0	1467814119	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	cooliodoc	@angry_barista I baked you a cake but I ated it
0	1467814180	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	viJILLante	this week is not going as i had hoped

So firstly we have create schema using case class.

```
case class TweetLabel(label : Int, tweet : String)
```

Then we load the data along with cleaning it. We load only those columns which are needed i.e. label and tweet. We removed special characters, website links, unnecessary spaces at starting and end.

```
val data = sc.textFile("/home/harsh/Desktop/twitter sentiment/2477_4140_bundle_archive/training.1600000.processed.noemoticon.csv").map(_.split(",")).map(attributes => TweetLabel(attributes(0).replace("\"","").toInt, attributes(5).replace("\"","").toLowerCase()
      .replaceAll("\n", "")
      .replaceAll("rt\\s+", "")
      .replaceAll("\\s+@\\w+", "")
      .replaceAll("@\\w+", "")
      .replaceAll("\\s+#\\w+", "")
      .replaceAll("#\\w+", "")
      .replaceAll("(?:https?|http?)://[\\w/%.-]+", "")
      .replaceAll("(?:https?|http?)://[\\w/%.-]+\\s+", "")
      .replaceAll("(?:https?|http?)//[\\w/%.-]+\\s+", "")
      .replaceAll("(?:https?|http?)//[\\w/%.-]+", "")
      .trim()
  )).toDF()
```

After loading, we need to convert tweets into feature vectors.

```
val tokenizer = new Tokenizer().setInputCol("tweet").setOutputCol("words")
val wordsData = tokenizer.transform(data)
```

```
val hashingTF = new HashingTF()

.setInputCol("words").setOutputCol("rawFeatures").setNumFeatures(1000)

val featurizedData = hashingTF.transform(wordsData)

val idf = new IDF().setInputCol("rawFeatures").setOutputCol("features")

val idfModel = idf.fit(featurizedData)

val rescaledData = idfModel.transform(featurizedData)
```

Then we split the transformed data into two subsets i.e. training and test(ratio 0.7:0.3)

```
val Array(training, test) = rescaledData.randomSplit(Array(0.7,0.3), seed=1234L)
```

Model Selection and Model Tuning

We tried Naïve Bays and Gradient Boosted Trees for classification.

```
val nb = new NaiveBayes()

val paramGrid = new ParamGridBuilder()

                    .addGrid(nb.modelType, Array("multinomial","complement","gaussian"))

                    .build()
```

```
val cv = new CrossValidator()
            .setEstimator(nb)
            .setEvaluator(new BinaryClassificationEvaluator())
            .setEstimatorParamMaps(paramGrid)
            .setNumFolds(3)
            .setParallelism(2)

val cvModel = cv.fit(training)
val nb_predictions = cvModel.transform(test)

val gbt = new GBTClassifier()
                .setLabelCol("label")
                .setFeaturesCol("features")
               .setMaxIter(15)
               .setFeatureSubsetStrategy("all")
               .setMaxDepth(10)
val paramGrid2 = new ParamGridBuilder()
          .addGrid(gbt.featureSubsetStrategy, Array("auto","all"))
                 .addGrid(gbt.maxDepth, Array(5,10))
                 .addGrid(gbt.maxIter, Array(10,15))
                 .build()
```

```
val cv2 = new CrossValidator()
        .setEstimator(gbt)
        .setEvaluator(new BinaryClassificationEvaluator())
        .setEstimatorParamMaps(paramGrid2)
        .setNumFolds(3)
        .setParallelism(2)

val model2 = cv2.fit(training)
val gbt_predictions = model.transform(test)
```

Conclusion

We evaluated accuracy for both Naïve Bays and Gradient boosted trees using MultiClassClassification Evaluator and got 68 % accuracy for Naïve Bays and 71 %

```
val evaluator = new MulticlassClassificationEvaluator()
                .setLabelCol("label")
                .setPredictionCol("prediction")
                .setMetricName("accuracy")
val nb_accuracy = evaluator.evaluate(nb_predictions)
val gbt_accuracy = evaluator.evaluate(gbt_predictions)
```