Twitter User Gender Classification using SparkML

Project Report

By Venkatesh Jagannathan

Data Processing Steps

List of Data Issues found in raw data

1. New lines, Non Ascii characters in text data
2. Few malformed CSV lines
3. “Unknown” gender value
4. Gender judgement by reviewers with lower confidence
5. Text with multiple words, common stop words

How did you tackle the issues mentioned above?

1. Used `.option("mode", "DROPMALFORMED")` while loading csv data to clear malformed and problematic records
2. Filtered records with predicted gender value & good confidence for better accuracy (~61% records net obtained)
3. **RegexTokenizer** to tokenize the words, **explode** inbuilt udf to split words into multiple rows and **StopWordsRemover** to remove common words

```
Dataset<Row> reduced = userKnowDf.select(
        col("_golden"),
        col("gender"),
        col("gender:confidence"),
        col("gender_gold"),
        col("description"),
        col("sidebar_color"),//.cast("long"),
        col("link_color"),//.cast("long"),
        col("text"),
        col("tweet_location"),
        col("user_timezone")
        )
        .filter("`gender:confidence` == 1") //filter low
        //gender confidence records for better model accuracy
        .filter("gender == 'male' or gender == 'female' or gender == 'brand'"); //eliminate
        //unreliable data

RegexTokenizer tokenizer = new RegexTokenizer()
  .setInputCol("text")
  .setOutputCol("words")
    .setPattern("\\W");

 //remove commonly used words
 StopWordsRemover remover = new StopWordsRemover()
          .setInputCol("words")
          .setOutputCol("filtered_words");
```

```
Dataset<Row> word_result = filtered.sqlContext()
        .sql("SELECT _golden, gender,`gender:confidence`,gender_gold"
                + ",sidebar_color,link_color,tweet_location,user_timezone, word "
                + "FROM parent LATERAL VIEW explode(filtered_words) childTable AS word");
```

Input Records

_unit_id	_golden	_unit_state	_trusted_jud	_last_judgm	gender	gender:confi	text
815719226	FALSE	finalized	3	10/26/15 23:2	male	1	Robbie E Responds To Critics After Win Against Eddie Edwards In The #WorldTitleSe
815719227	FALSE	finalized	3	10/26/15 23:3	male	1	���It felt like they were my friends and I was living the story with them� https:
815719228	FALSE	finalized	3	10/26/15 23:3	male	0.6625	i absolutely adore when louis starts the songs it hits me hard but it feels good
815719229	FALSE	finalized	3	10/26/15 23:1	male	1	Hi @JordanSpieth - Looking at the url - do you use @IFTTT?! Don't typically see an
815719245	FALSE	finalized	3	10/26/15 22:1	unknown	0.3527	

Output

```
+-------+------+-----------------+-----------+-------------+----------+---------------+--------------------+----------------+
|_golden|gender|gender:confidence|gender_gold|sidebar_color|link_color| tweet_location|       user_timezone|            word|
+-------+------+-----------------+-----------+-------------+----------+---------------+--------------------+----------------+
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|          robbie|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|               e|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|        responds|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|         critics|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|             win|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|           eddie|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|         edwards|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|worldtitleseries|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|           https|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|              co|
|  FALSE|  male|                1|       null|       FFFFFF|    08C2C2|main; @Kan1shk3|             Chennai|      nsybbmvjkz|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|            felt|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|            like|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|         friends|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|          living|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|           story|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|           https|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|              co|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|      arnge0yhno|
|  FALSE|  male|                1|       null|       C0DEED|    0084B4|           null|Eastern Time (US ...|         retired|
+-------+------+-----------------+-----------+-------------+----------+---------------+--------------------+----------------+
```

Model Building

List of Models used

1. Decision Tree Algorithm
   I. Columns
      a) Label
         i. gender
      b) Features
         i. text
         ii. sidebar_color (4-5% improved accuracy)
         iii. link_color (4-5% improved accuracy)
   II. HyperParameters
      a) maxDepth – 20 (Test Data Accuracy = 52%, Training Data Accuracy = 52%)
         i. Increasing further to 25 caused model to *overfit* (test=53%,train=54%)
         ii. Lowering to 15 caused to *underfit* (test=50%,train=50%)

```
DecisionTreeClassifier dt = new DecisionTreeClassifier().setLabelCol("label").setFeaturesCol("features").setSeed(0);
dt.setMaxDepth(20);
dt.setMinInfoGain(0.0);
dt.setMinInstancesPerNode(1);
dt.setCacheNodeIds(false);
dt.setMaxBins(3000);
DecisionTreeClassificationModel Model = dt.fit(trainingData);
System.out.println("Learned Decision tree" + Model.toDebugString());

Dataset<Row>  rawPredictions = Model.transform(testData);
Dataset<Row> predictions = predConverter.transform(labelConverter.transform(rawPredictions));
predictions.select("predictionStr", "labelStr", "features").show(5);
//***Test data predictions***
Dataset<Row> trainRawPredictions = Model.transform(trainingData);
Dataset<Row> trainPredictions = predConverter.transform(labelConverter.transform(trainRawPredictions));
trainPredictions.select("predictionStr", "labelStr", "features").show(5);
```

2. Random Forest Algorithm
   I. Columns
      a) Label
         i. gender
      b) Features
         i. text
         ii. sidebar_color (4-5% improved accuracy)
         iii. link_color (4-5% improved accuracy)
   II. HyperParameters
      a) maxDepth – 25 (Accuracy = 55%) (Test Data Accuracy = 52%, Training Data Accuracy = 52%)
         i. Increasing further to 30 caused no improvement (test=55%,train=55%)
         ii. No overfitting observed in this model
         iii. Lowering to 20 caused to *underfit* (test=54%,train=54%)
      b) Increasing minInfoGain to 1.0, minInstancePerNode=5 and setCacheNodeIds=false reduced accuracy to 37%

```
//************Random Forest****************//
RandomForestClassifier rf = new RandomForestClassifier();
rf.setLabelCol("label").setFeaturesCol("features").setSeed(0);
rf.setMaxDepth(25);
rf.setMinInfoGain(0.0);
rf.setMinInstancesPerNode(1);
rf.setCacheNodeIds(false);
rf.setMaxBins(3000);

RandomForestClassificationModel rfModel = rf.fit(trainingData);
System.out.println("Learned Random Forest Decision tree" + rfModel.toDebugString());
Dataset<Row> rfRawPredictions = rfModel.transform(testData);
Dataset<Row> rfPredictions = predConverter.transform(labelConverter.transform(rfRawPredictions));
rfPredictions.select("predictionStr", "labelStr", "features").show(5);

//***Test data predictions***//
Dataset<Row> rfTrainRawPredictions = rfModel.transform(trainingData);
Dataset<Row> rfTrainPredictions = predConverter.transform(labelConverter.transform(rfTrainRawPredictions));
rfTrainPredictions.select("predictionStr", "labelStr", "features").show(5);
```

Evaluation Metrics

<table>
<tr><th rowspan="2">Model</th><th colspan="4">Evaluation Scores</th><th rowspan="2">Confusion Matrix</th></tr>
<tr><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1</th></tr>
<tr><td>Decision Tree</td><td>52%</td><td>Female:0.69
Male:0.41
Brand:0.63

Weighted:0.57</td><td>Female:0.48
Male:0.73
Brand:0.33

Weighted:0.51</td><td>Female:0.57
Male:0.53
Brand:0.44

Weighted:0.51</td><td><pre>
+--------+------------+-----+
|labelStr|predictionStr|count|
+--------+------------+-----+
male	brand	1259
male	female	2070
female	female	6773
male	male	9050
brand	brand	4007
female	male	6216
brand	male	6522
female	brand	1038
brand	female	1428
+--------+------------+-----+
</pre></td></tr>
<tr><td>Random Forest</td><td>55%</td><td>Female:0.71
Male:0.43
Brand:0.69

Weighted:0.61</td><td>Female:0.49
Male:0.77
Brand:0.36

Weighted:0.55</td><td>Female:0.58
Male:0.55
Brand:0.47

Weighted:0.54</td><td><pre>
+--------+------------+-----+
|labelStr|predictionStr|count|
+--------+------------+-----+
male	brand	1086
male	female	1668
female	female	6996
male	male	9625
brand	brand	4316
female	male	6221
brand	male	6488
female	brand	810
brand	female	1153
+--------+------------+-----+
</pre></td></tr>
</table>

Model Fit

Model	Max Depth	Test Data Accuracy %	Training Data Accuracy %	Fit
Decision Tree	10	47	47	Under
Decision Tree	15	50	50	Under
Decision Tree	**20**	**52**	**52**	**Right**
Decision Tree	25	53	54	Over
Random Forest	10	49	49	Under
Random Forest	15	52	52	Under
Random Forest	20	54	54	Under
Random Forest	**25**	**55**	**55**	**Right**
Random Forest	30	55	55	No Change

Inferences and suggestions

1. **Results Comparision**

Model	Pros	Cons
Decision Tree	Fast Execution	Low Accuracy Prone to overfitment
Random Tree	Better Accuracy Less prove to overfitment	Slower compared to Decision Tree

Random Forest and clearly better algorithm than decision tree in terms of its prediction accuracy & overfitting issue. MaxDepth – Increasing max depth improved accuracy until a point after which it lead to overfitting. Random forest was less affected by overfitment than decision tree. In given data, there was no further improvement due to other hyperparameters: -

HyperParameter	Value	Algorithm	New Accuracy
minInfoGain	0 to 1	Decision Tree	37% (Reduced)
minInstancesPerNode	0 to 3	Decision Tree	52% (No change)
minInfoGain	0 to 1	Random Forest	55% (No change)
minInstancesPerNode	0 to 3	Random Forest	55% (No change)

2. **Improvisation techniques**
    1. Tuning more hyperparameters: -
        1. **MinInfoGain** – Increasing this parameter can improve stopping criteria there by increasing speed of anlaysis
        2. **MinInstancesPerNode** – By increasing minimum number of required split candidates in order to consider it for training, it can improve accuracy in case of deeper training algorithm like random forest
    2. Trying other type of classification algorithms like Naive Bayes, Boosted Trees, Neural Networks etc.
3. Choice of preferred Algorithm

Random Forest is preffered choice of algorithm among the two due to better accuracy & better fitment with same training data.