Sports Articles Objectivity

Kovalev Evgeny, Chesakov Daniil, Gafurov Azamat

Skolkovo Institute of Science and Technology, "Introduction to Data Science"

Moscow, 2018

Task

- Binary classification of sports articles: objective vs subjective
- Betting is biased towards irrelevant information
- However, it is hard to identify (probably harder then semantic amalysis)
- Here is where ML comes to the rescue!

Examples

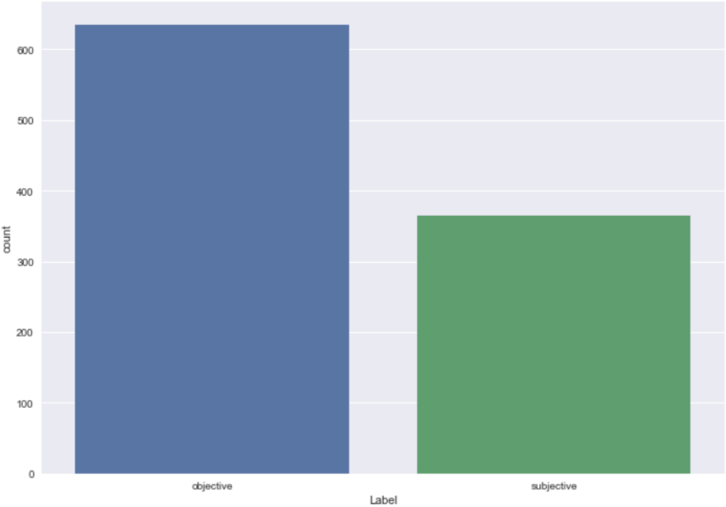
- Objective:

```
'Finalists in the Apertura play-offs, Toluca had drawn their first two Clausura games but got off to a good start when Edgar Be
nitez put them ahead in the 16th minute.\nMatias Britos levelled 20 minutes later but Lucas Silva netted 14 minutes from the en
d to ensure the visitors took all three points.\n  \tFranco Arizala scored 13 minutes from time to ensure Jaguares claimed thei
r first point with a 1-1 draw against Monterrey, who had opened the scoring through Aldo De Nigris (14).\n Hosts Jaguares also
had Jorge Rodriguez sent off in the closing moments.'
```

- Subjective:

```
'BARCELONA star Dani Alves has claimed Arsenal\'s Jack Wilshere can be as good as superstars Xavi and Andres Iniesta.\n\nThe Br
azilian recently played against Wilshere in a friendly clash with England.\n\nAnd Alves was so impressed with the midfielder\'s
performance, he has urged Barca chiefs to bid for the 21-year-old this summer.\n\n"He is a great player who we have met playing
against Arsenal and without doubt he can reach the height of the players we have here at Barcelona like Xavi and Iniesta," he s
aid.\n\nArsenal have lost stars Cesc Fabregas and Alex Song to the Spanish league leaders over the past two summers.\n\nAnd bos
s Arsene Wenger will be bracing himself for new attempts from Barca to nick his latest midfield linchpin.\n\n"[Wilshere] has a
lot of quality and a great personality. If I was given the chance to choose, he is a player that I would sign for Barcelona," a
dded Alves. '
```

Label distribution



Wordclouds

- Objective:

- Subjective:

Text preprocessing

- Before:

```
'Finalists in the Apertura play-offs, Toluca had drawn their first two Clausura games but got off to a good start when Edgar Be
nitez put them ahead in the 16th minute.\nMatias Britos levelled 20 minutes later but Lucas Silva netted 14 minutes from the en
d to ensure the visitors took all three points.\n  \tFranco Arizala scored 13 minutes from time to ensure Jaguares claimed thei
r first point with a 1-1 draw against Monterrey, who had opened the scoring through Aldo De Nigris (14).\n Hosts Jaguares also
had Jorge Rodriguez sent off in the closing moments.'
```

```
'BARCELONA star Dani Alves has claimed Arsenal\'s Jack Wilshere can be as good as superstars Xavi and Andres Iniesta.\n\nThe Br
azilian recently played against Wilshere in a friendly clash with England.\n\nAnd Alves was so impressed with the midfielder\'s
performance, he has urged Barca chiefs to bid for the 21-year-old this summer.\n\n"He is a great player who we have met playing
against Arsenal and without doubt he can reach the height of the players we have here at Barcelona like Xavi and Iniesta," he s
aid.\n\nArsenal have lost stars Cesc Fabregas and Alex Song to the Spanish league leaders over the past two summers.\n\nAnd bos
s Arsene Wenger will be bracing himself for new attempts from Barca to nick his latest midfield linchpin.\n\n"[Wilshere] has a
lot of quality and a great personality. If I was given the chance to choose, he is a player that I would sign for Barcelona," a
dded Alves. '
```

- After:

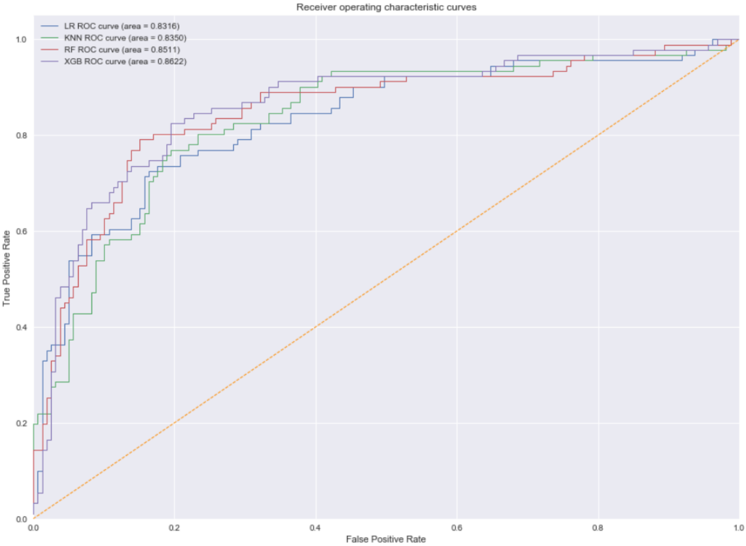
```
'finalist apertura toluca drawn first two clausura game got good start edgar benitez put ahead minut matia brito level minut la
ter luca silva net minut end ensur visitor took three point franco arizala score minut time ensur jaguar claim first point draw
monterrey open score aldo de nigri host jaguar also jorg rodriguez sent close moment'
```

```
'barcelona star dani alv claim arsenal jack wilsher good superstar xavi andr iniesta the brazilian recent play wilsher friend c
lash england and alv impress midfield perform urg barca chief bid summer he great player met play arsenal without doubt reach h
eight player barcelona like xavi iniesta said arsenal lost star cesc fabrega alex song spanish leagu leader past two summer and
boss arsene wenger brace new attempt barca nick latest midfield linchpin wilsher lot qualiti great person if i given chanc choo
s player i would sign barcelona ad alv'
```

Deep Learning

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_3 (InputLayer)         (None, 500)               0
_________________________________________________________________
embedding_3 (Embedding)      (None, 500, 50)           500000
_________________________________________________________________
bidirectional_3 (Bidirection (None, 500, 100)          40400
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 498, 50)           15050
_________________________________________________________________
global_max_pooling1d_3 (Glob (None, 50)                0
_________________________________________________________________
dense_5 (Dense)              (None, 50)                2550
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 51
=================================================================
Total params: 558,051
Trainable params: 558,051
Non-trainable params: 0
_________________________________________________________________
```

TF-IDF ROCs

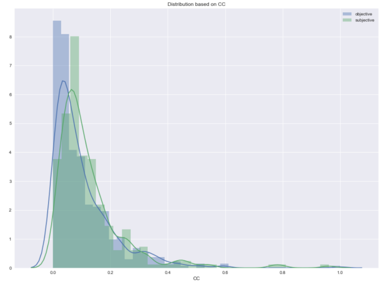

Receiver operating characteristic curves
LR ROC curve (area = 0.8316)
KNN ROC curve (area = 0.8350)
RF ROC curve (area = 0.8511)
XGB ROC curve (area = 0.8622)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

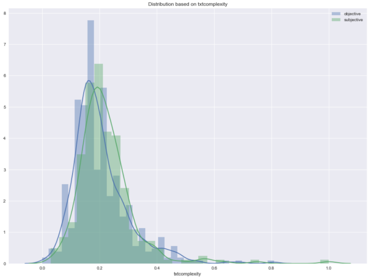
Feature extraction

- `symbols` - total number of symbols in raw text
- `sentences` - total number of sentences
- `unique_words_count` - number of unique words
- `unique_words_share` - ratio between number of unique words and number of total words
- `word_average_len` - average word length in text
- `stopwords_count` - total number of stopwords
- `stopwords_share` - ratio between number of stopwords and number of total words
- `polarity_raw`, `polarity_preprocessed` - polarity in raw and preprocessed text respectively using textblob
- `subjectivity_raw`, `subjectivity_preprocessed` - subjectivity in raw and preprocessed text respectively using textblob

Feature distributions

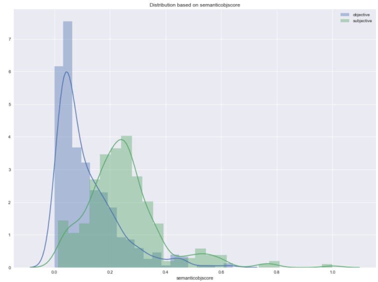
- Bad separation:

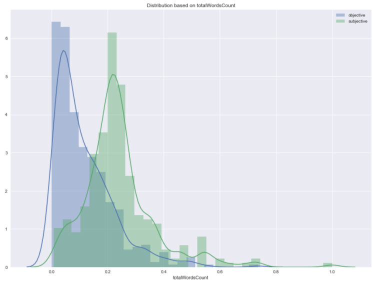




Feature distributions

- Good separation:





Low variance

```
WRB                  0.000000
NNP                  0.000000
ellipsis             0.000000
sentencelast         0.002663
JJS                  0.002814
colon                0.006744
semicolon            0.006916
pronouns1st          0.007173
TOs                  0.009206
exclamationmarks     0.010101
dtype: float64
```

Low variance

- Good to exclude:

```
X_train['JJS'].value_counts()
```

```
0.0    744
0.2      3
0.6      1
0.8      1
1.0      1
Name: JJS, dtype: int64
```

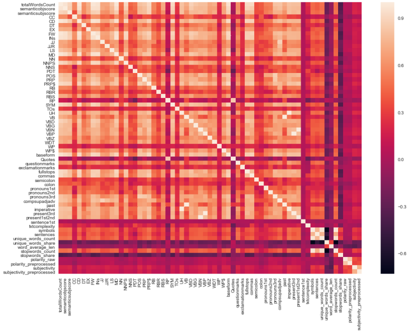
- Bad to exclude:

```
X_train['colon'].value_counts()
```

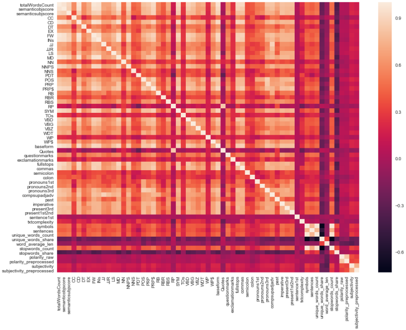
```
0.000000    349
0.026316    169
0.052632     80
0.078947     51
0.105263     32
0.131579     20
0.157895     13
0.184211      8
0.236842      7
0.210526      7
0.315789      3
0.289474      2
0.605263      1
0.421053      1
0.368421      1
0.710526      1
0.394737      1
0.263158      1
0.868421      1
0.342105      1
1.000000      1
Name: colon, dtype: int64
```

High correlation (1.0)

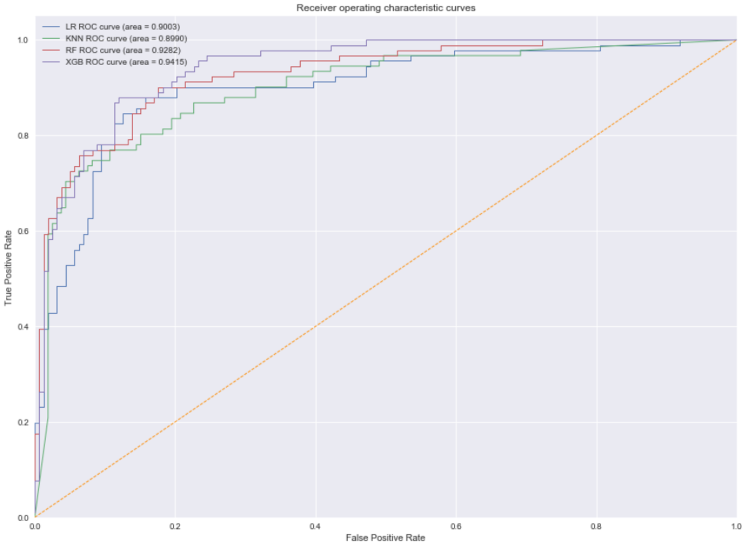
Before



After



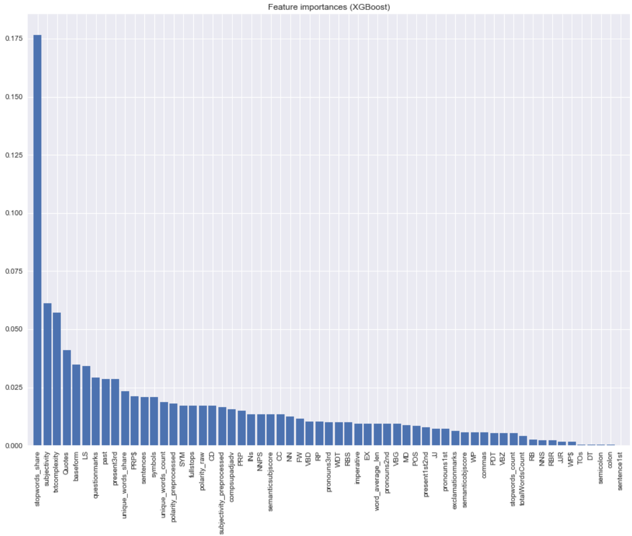
ROCs


Receiver operating characteristic curves
LR ROC curve (area = 0.9003)
KNN ROC curve (area = 0.8990)
RF ROC curve (area = 0.9282)
XGB ROC curve (area = 0.9415)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Results

Model	Data	AUC-ROC
XGBoost	**tabular**	**0.9415**
Random Forest	tabular	0.9282
Logistic Regression	tabular	0.9003
KNN	tabular	0.8990
XGBoost	**TF – IDF**	**0.8622**
Random Forest	TF-IDF	0.8511
LSTM + Conv	**texts**	**0.8460**
KNN	TF-IDF	0.8350
Logistic Regression	TF-IDF	0.8316
LSTM	texts	0.8269

Feature importances



Conclusion

- Different approaches were compared
  - DL on texts
    - LSTM+Conv was better than LSTM
    - The worst results though
    - Probably model architecture should be more complex
  - ML on TF-IDF matrix
    - Best: XGBoost
  - ML on tabular data
    - Best: XGBoost
    - The best approach
- EDA was performed
  - Low variance, high correlation features were excluded
- Feature extraction from texts
  - Golden feature: stopwords share
- Model is applicable to a real-life scenario
  - It is interpretable, the quality is good
  - But better to train it on larger dataset