SOCCER COMMENTARY MINING

Szymon Maksymiuk

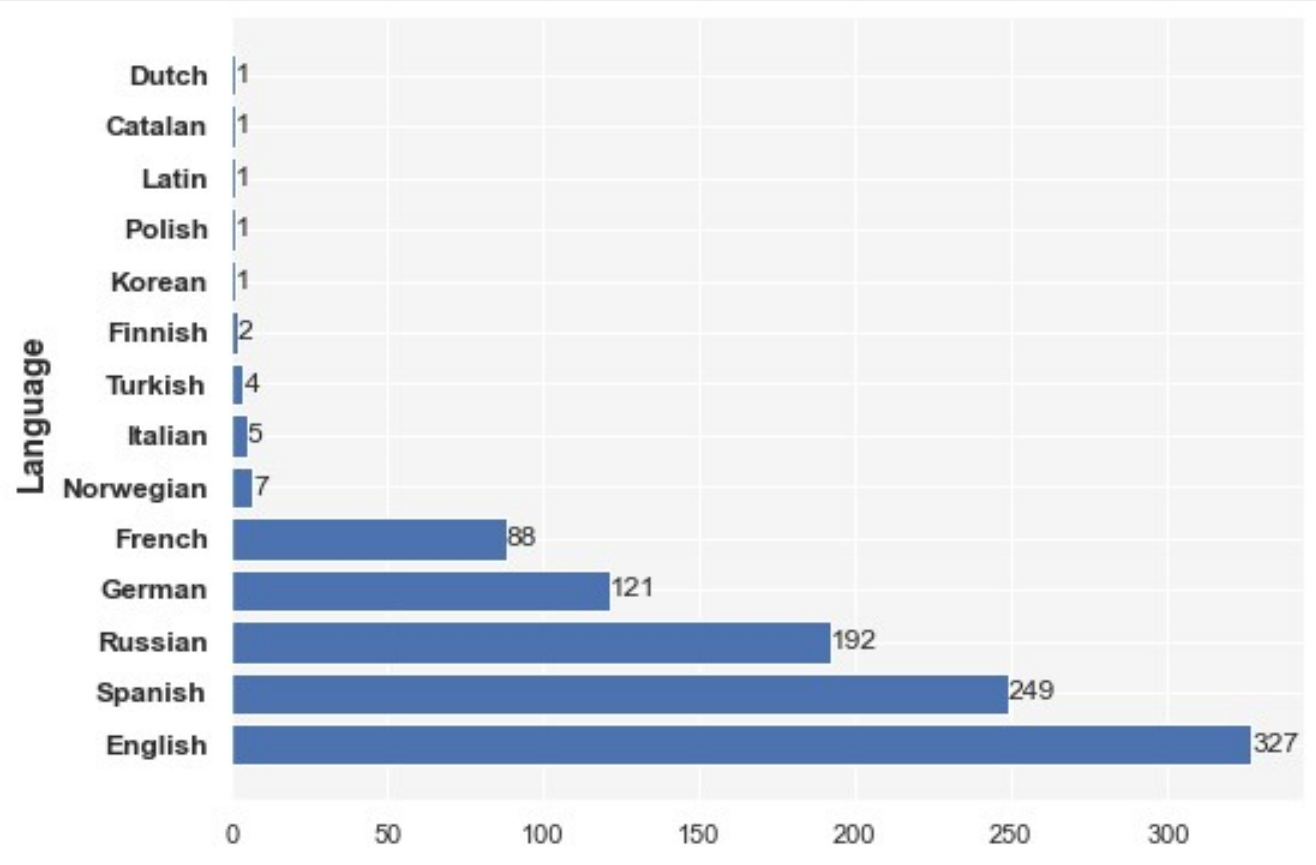
Adam Narożniak

Władysław Olejnik

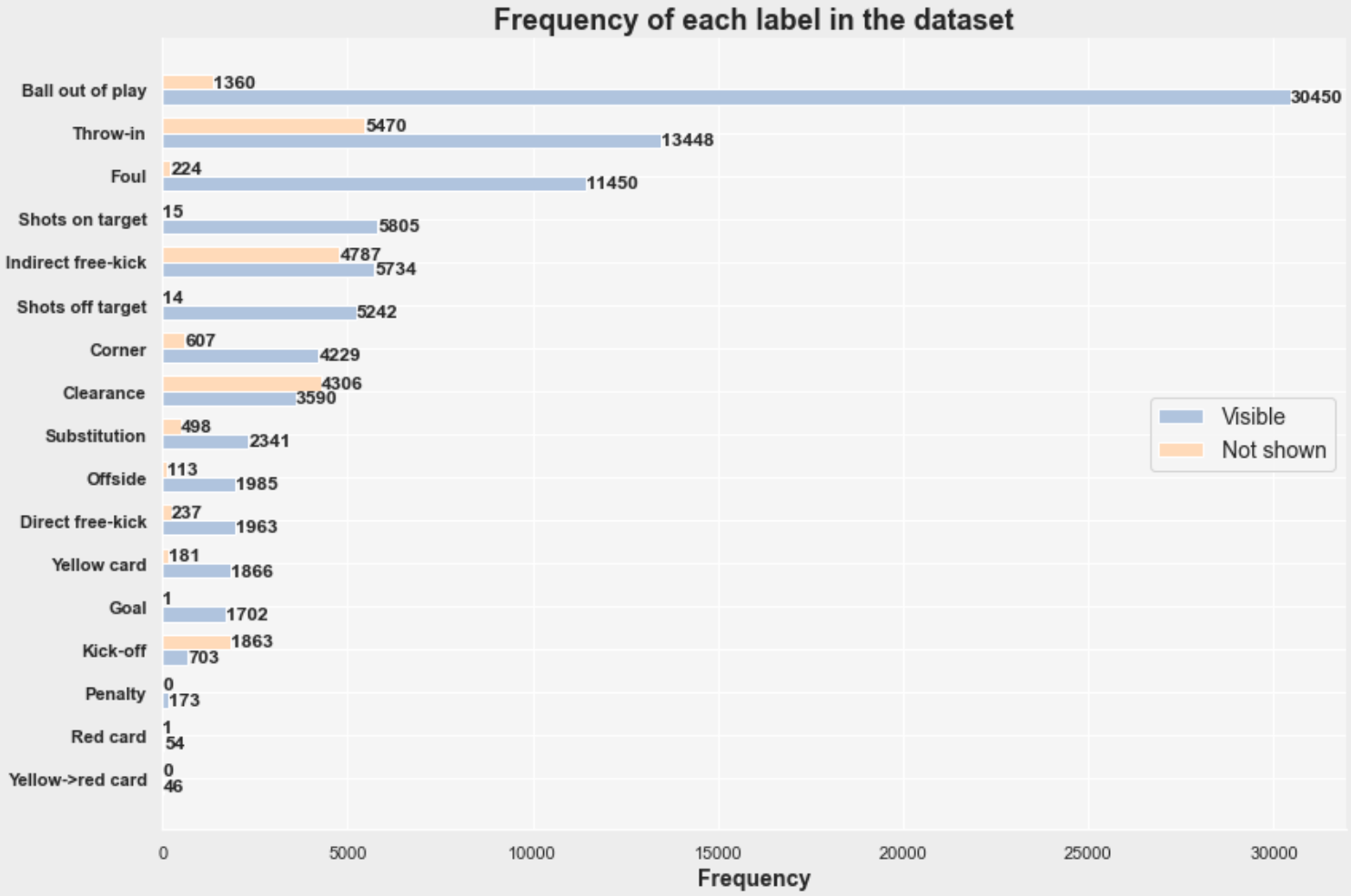
Patryk Świątek

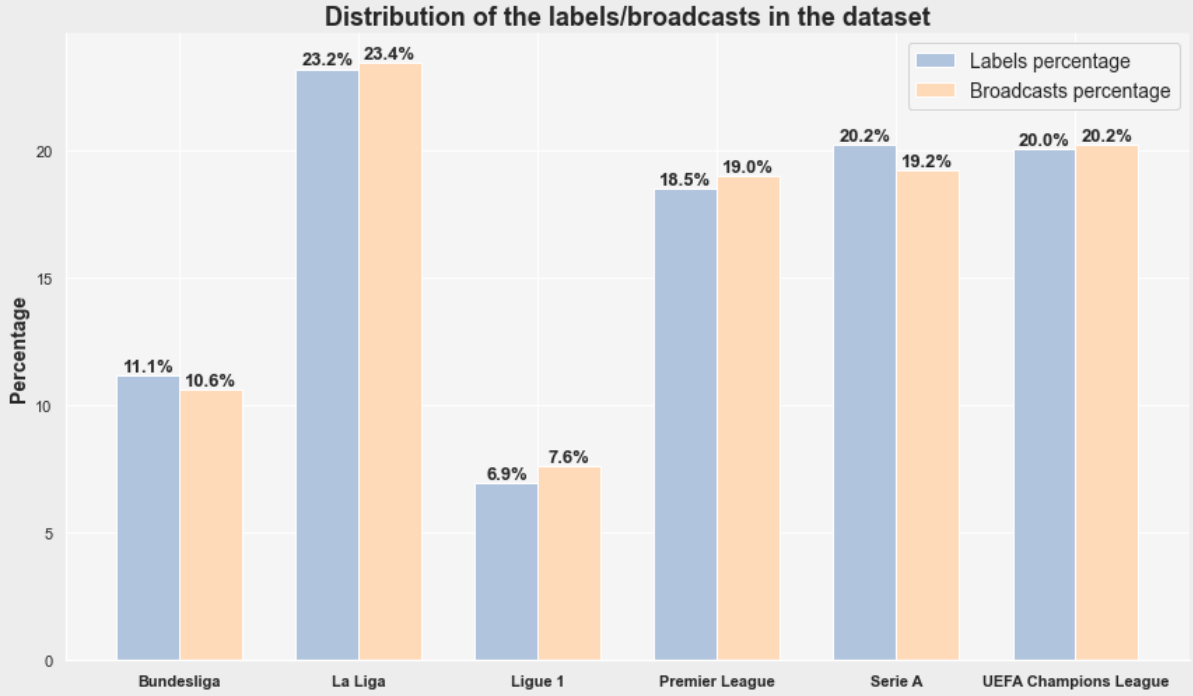
Introduction: basic information about datasets

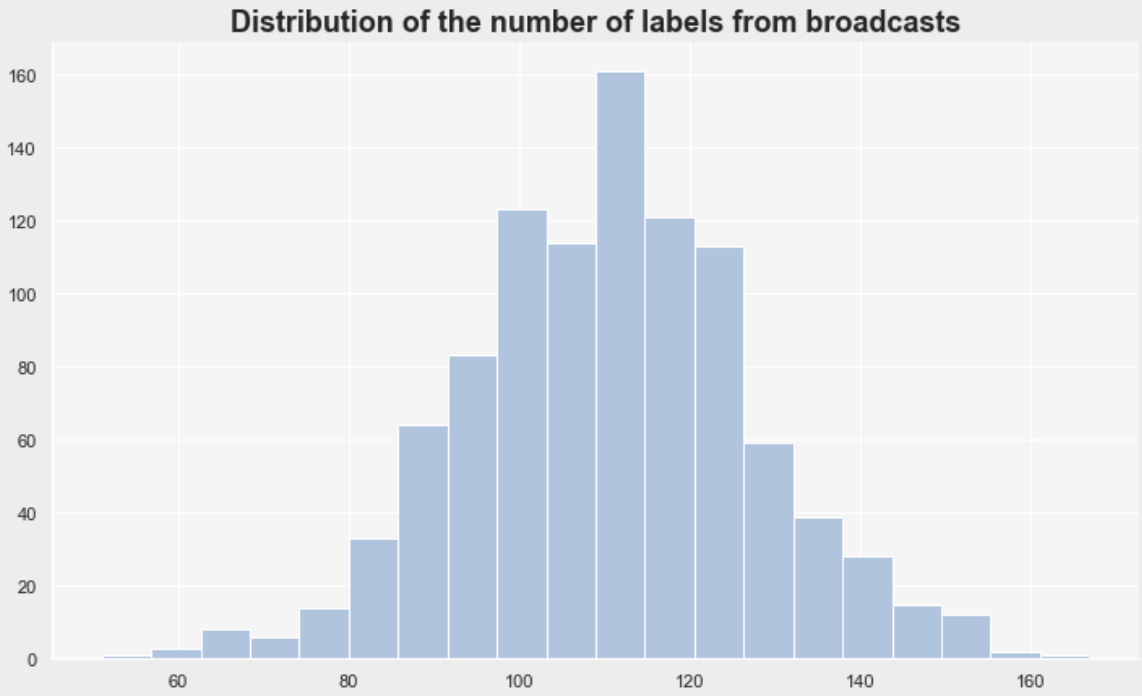
Transcriptions



Labels of match events

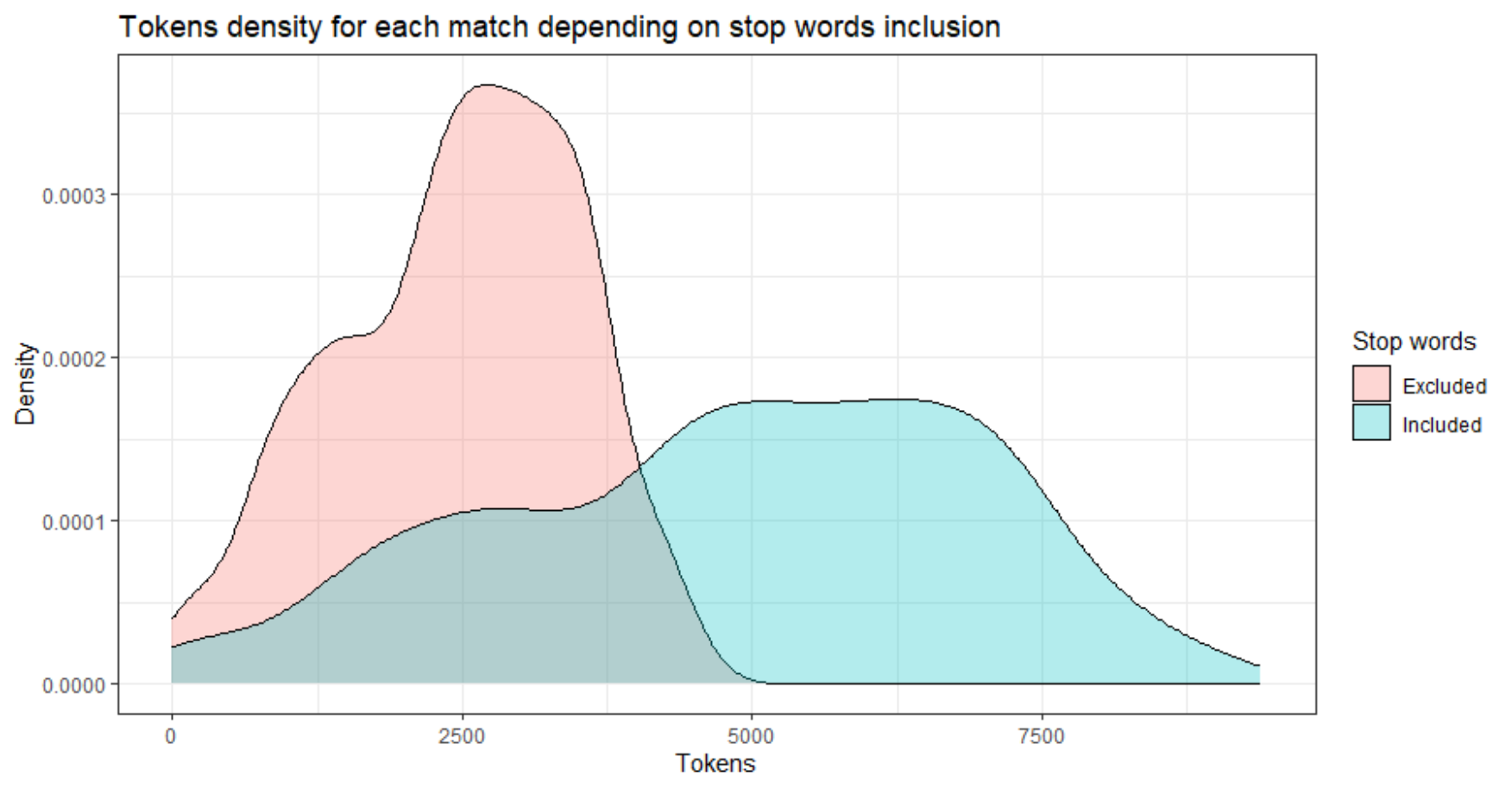



Distribution of the labels/broadcasts in the dataset
Labels percentage
Broadcasts percentage
Percentage
0
5
10
15
20
11.1%
10.6%
23.2%
23.4%
6.9%
7.6%
18.5%
19.0%
20.2%
19.2%
20.0%
20.2%
Bundesliga
La Liga
Ligue 1
Premier League
Serie A
UEFA Champions League

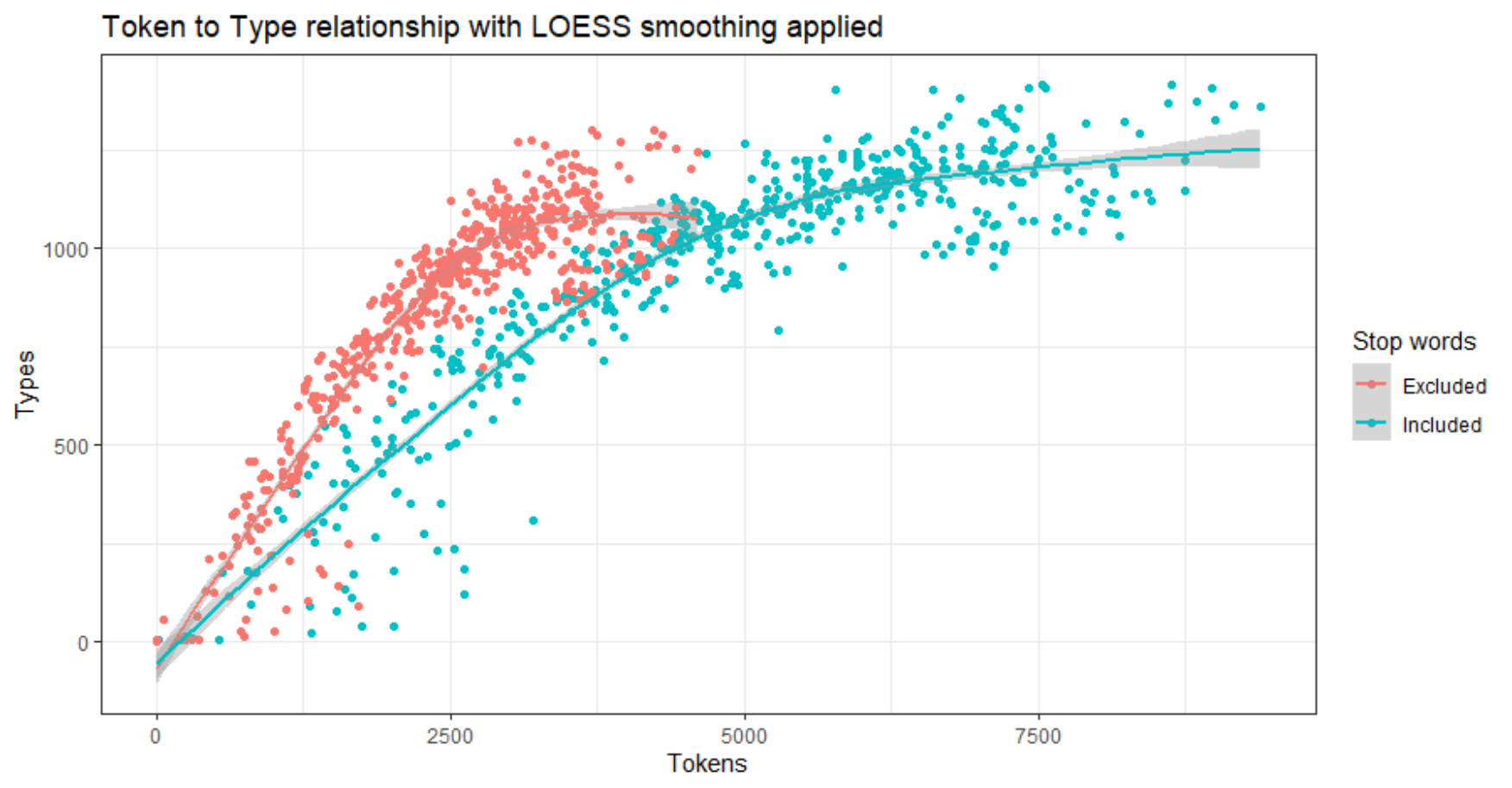

Distribution of the number of labels from broadcasts
0
20
40
60
80
100
120
140
160
60
80
100
120
140
160

EDA: text analysis

Tokens density analysis



Token to type relation



Text complexity - Flesch reading ease score

$$Y = 206.835 - 1.015(X_1) - 84.6(X_2)$$

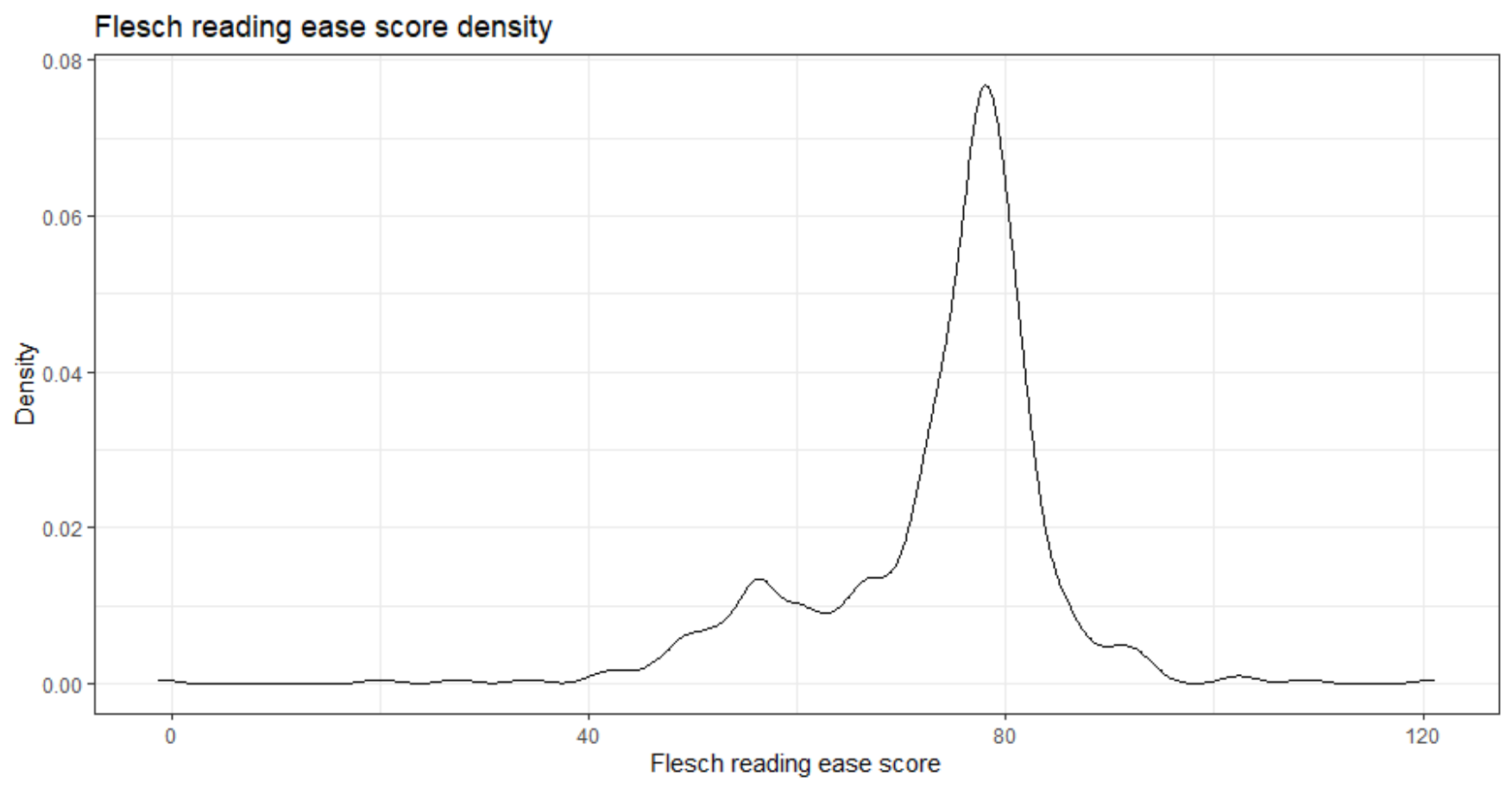
Y= Reading Ease Score

X_1 = Average Sentence Length

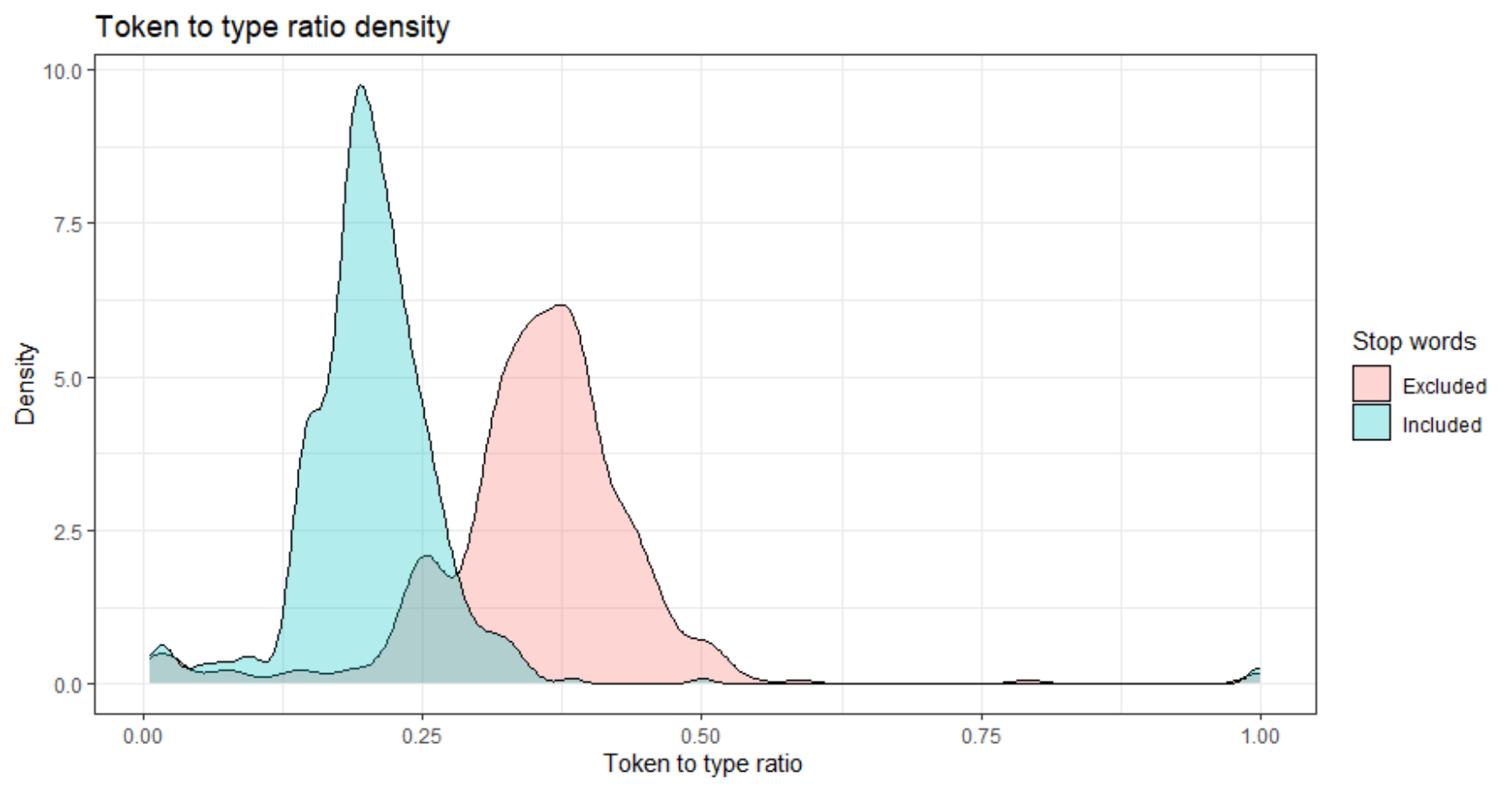
X_2 = Average number of syllables per word

Reading Ease Score	Descriptive Categories	Estimated Reading Grade
90 – 100	Very Easy	5th Grade
80 – 90	Easy	6th Grade
70 – 80	Fairly Easy	7th Grade
60 – 70	Standard / Plain English	8th and 9th Grade
50 – 60	Fairly Difficult	10th to 12th Grade (High School Sophomore to Senior)
30 – 50	Difficult	In College
0 - 30	Very Difficult	College Graduate

Text complexity - Flesch reading ease score

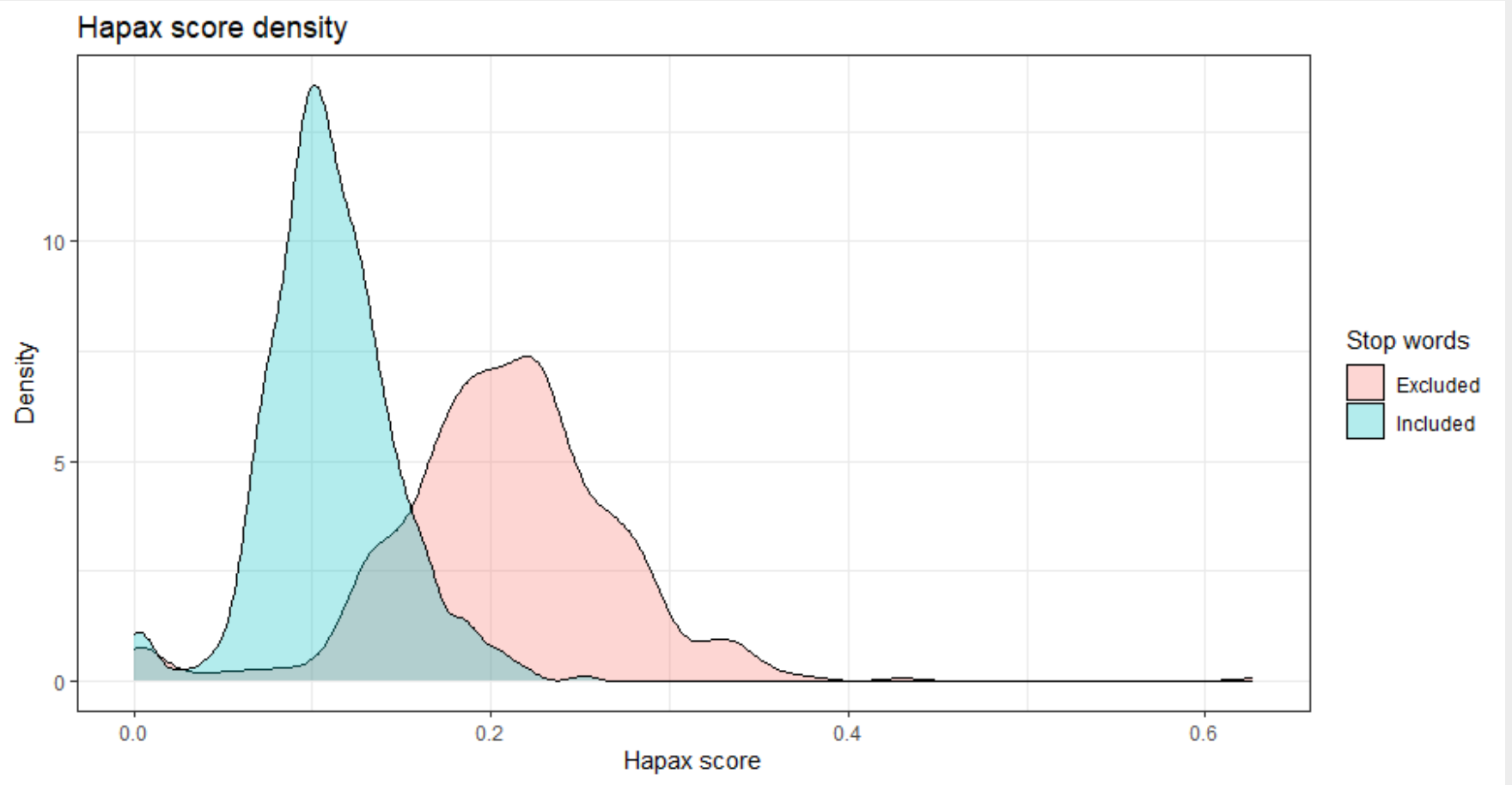


Text richness – token to type ratio



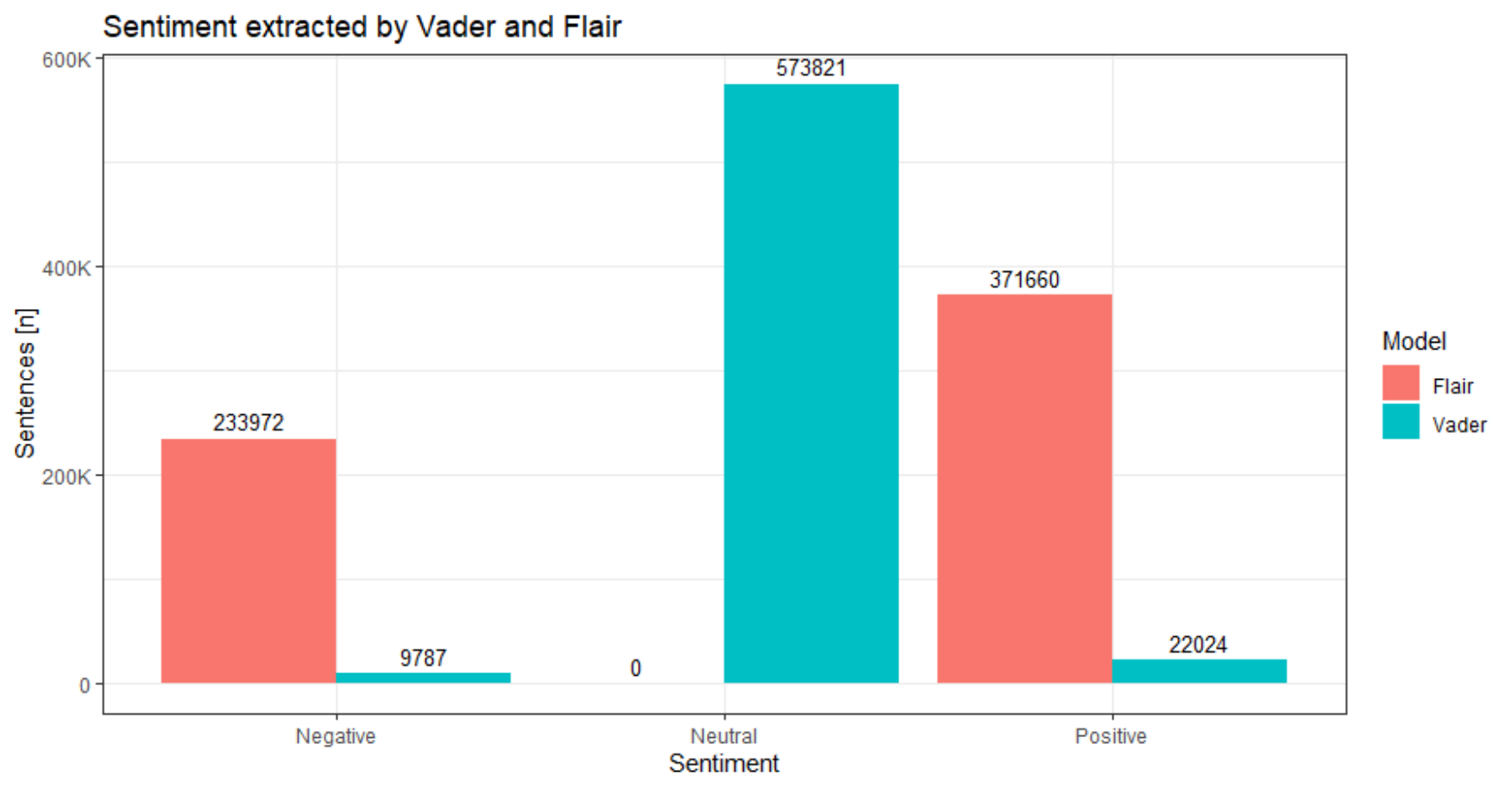
Text richness – Hapax score

Richenss measure defined as the number of words that occured only once by the number of all the words in the text

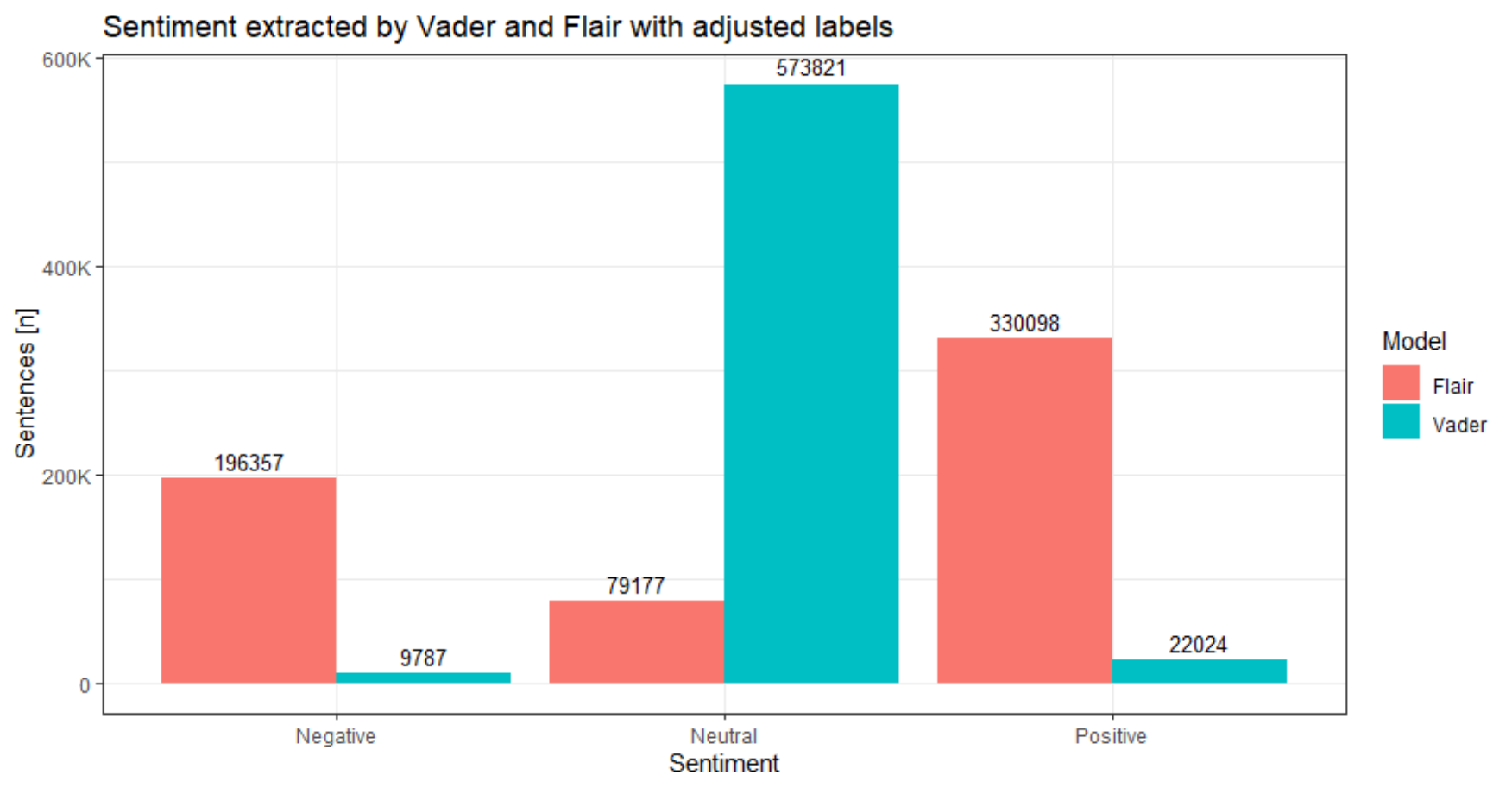


Results and statistical analysis

Sentiment histogram

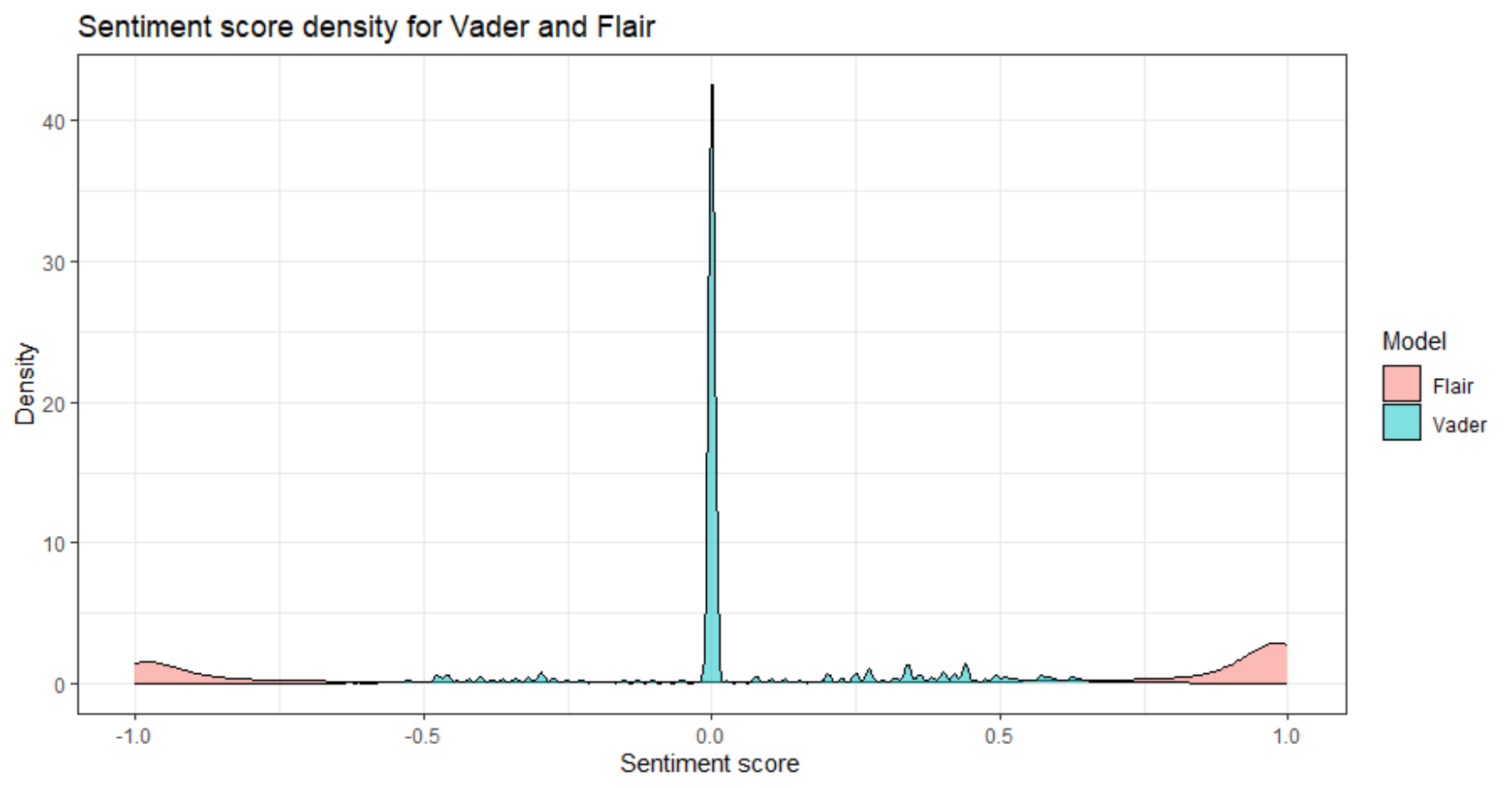


Sentiment histogram – modified flair

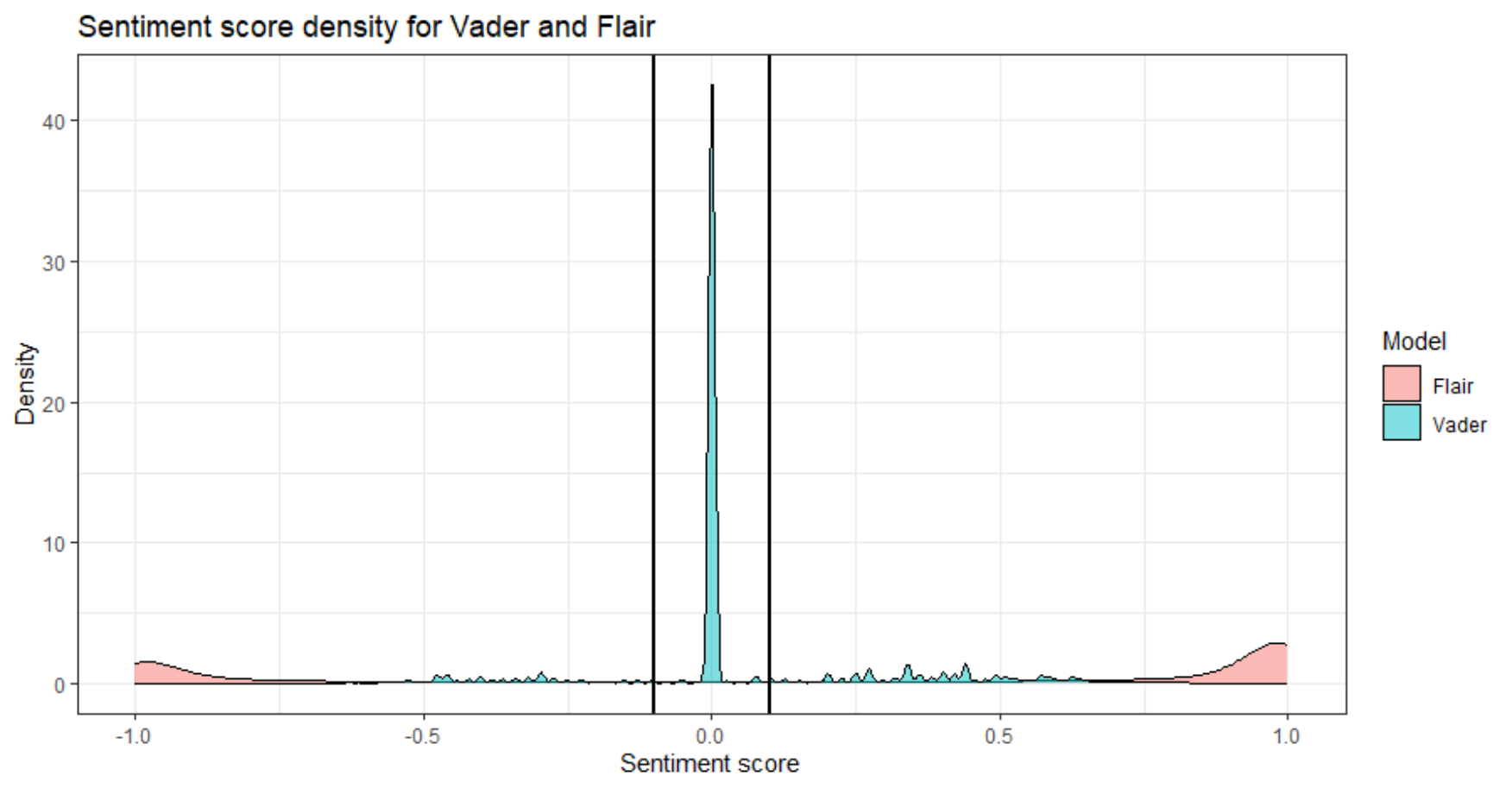


Wordclouds for sentiments extracted by Vader

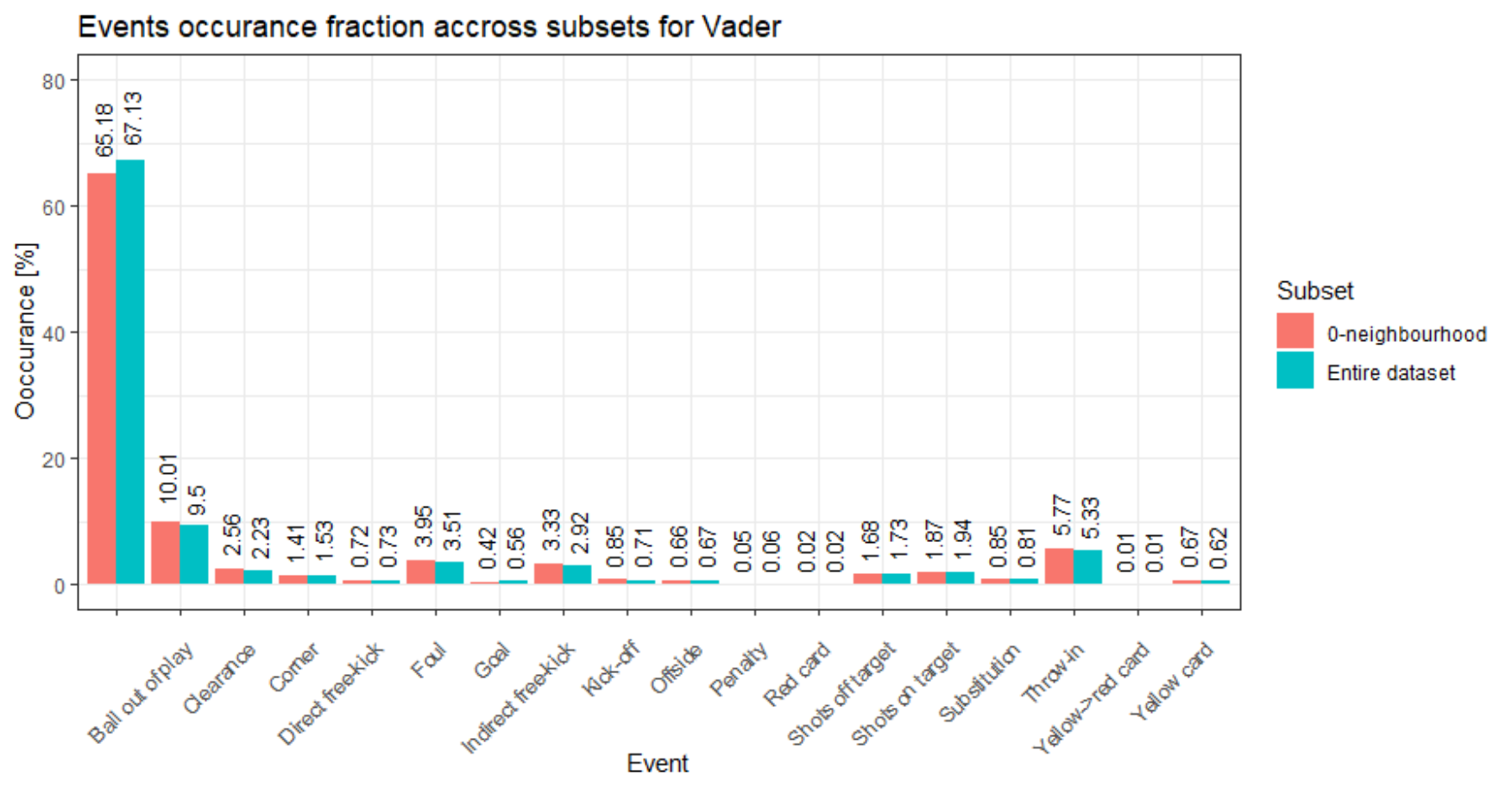
Sentiment density



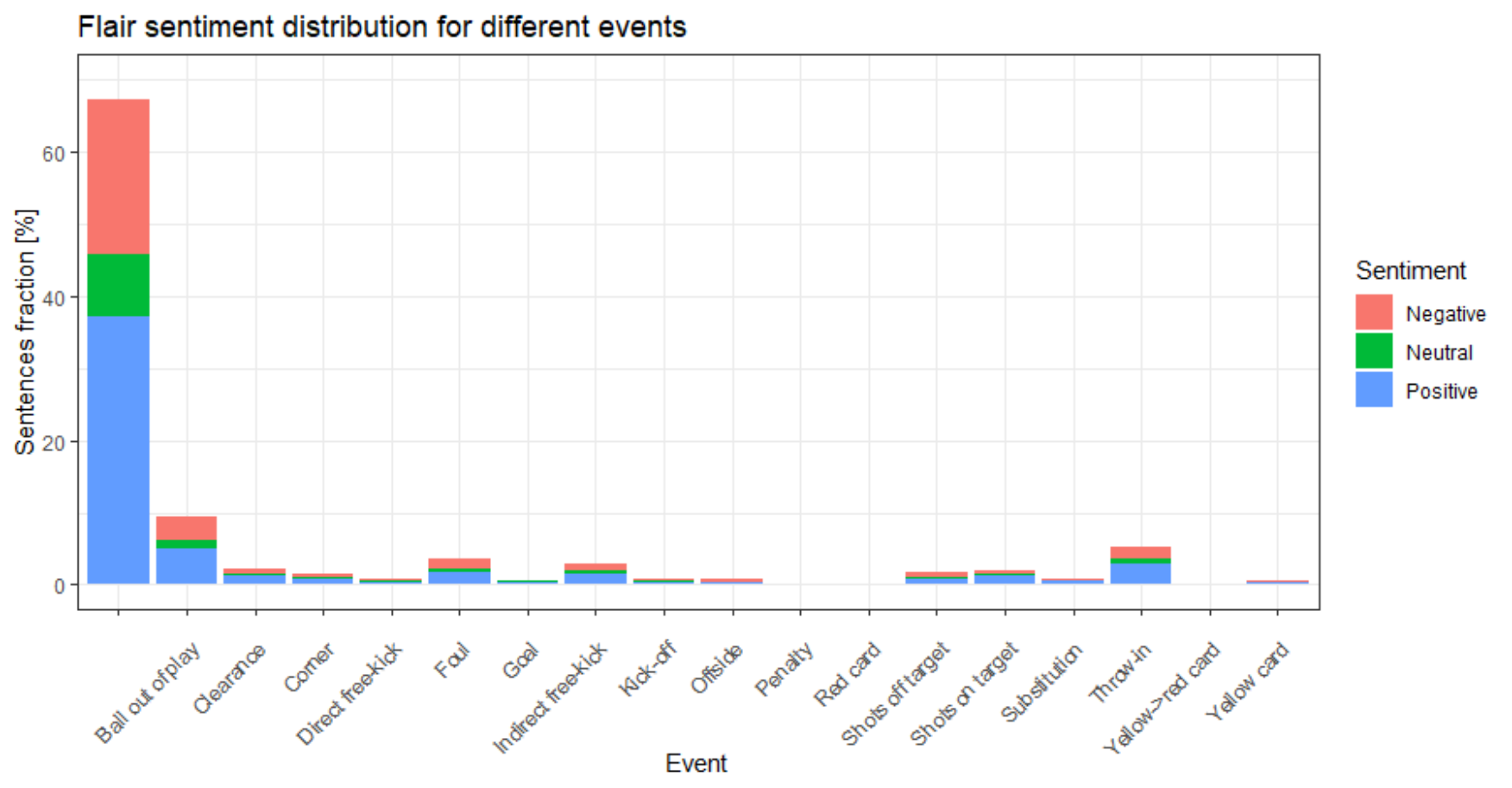
Sentiment density cutoff



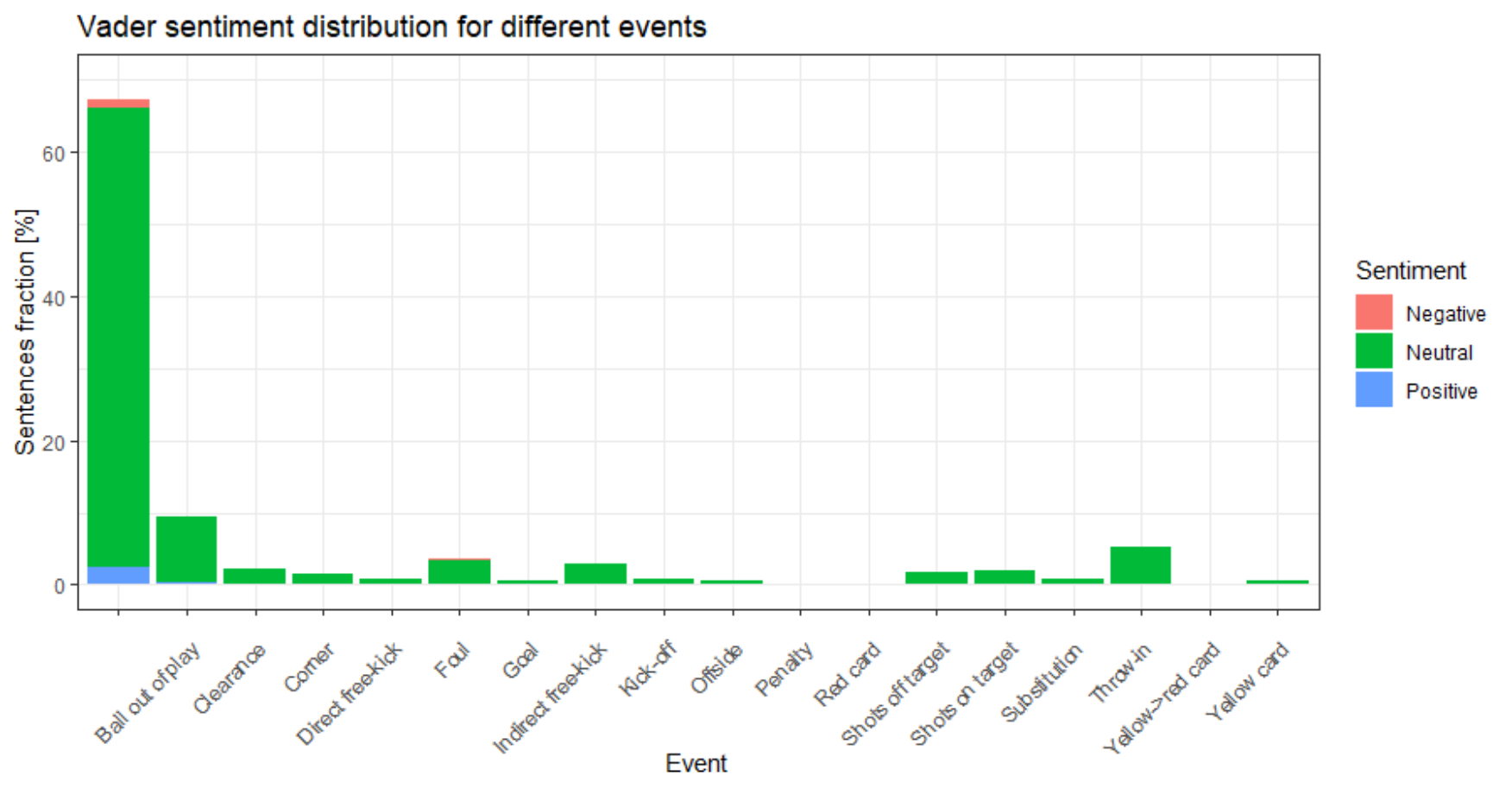
Vader – zero neighbourhood and whole dataset comparison



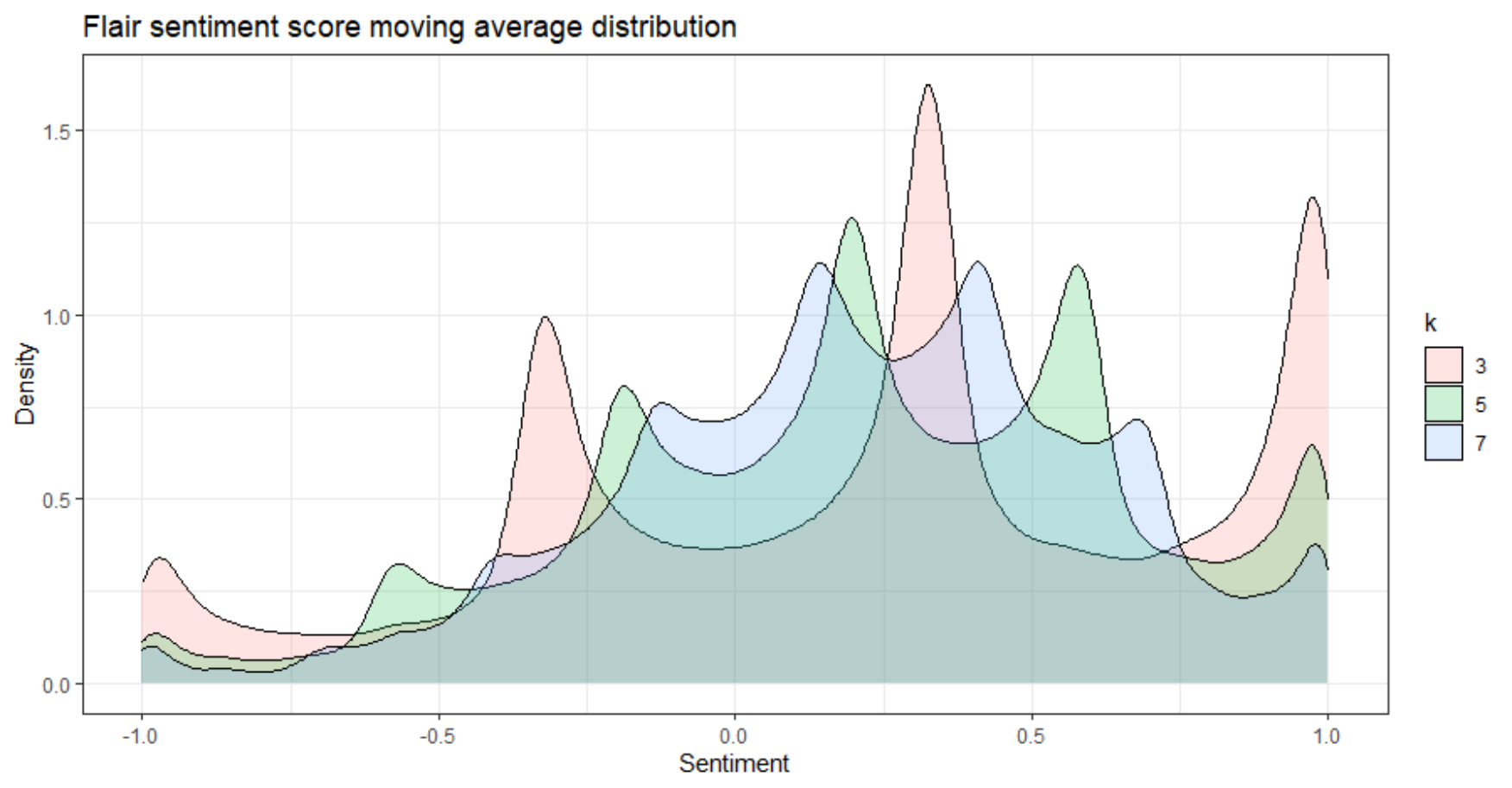
Flair – sentiment across events



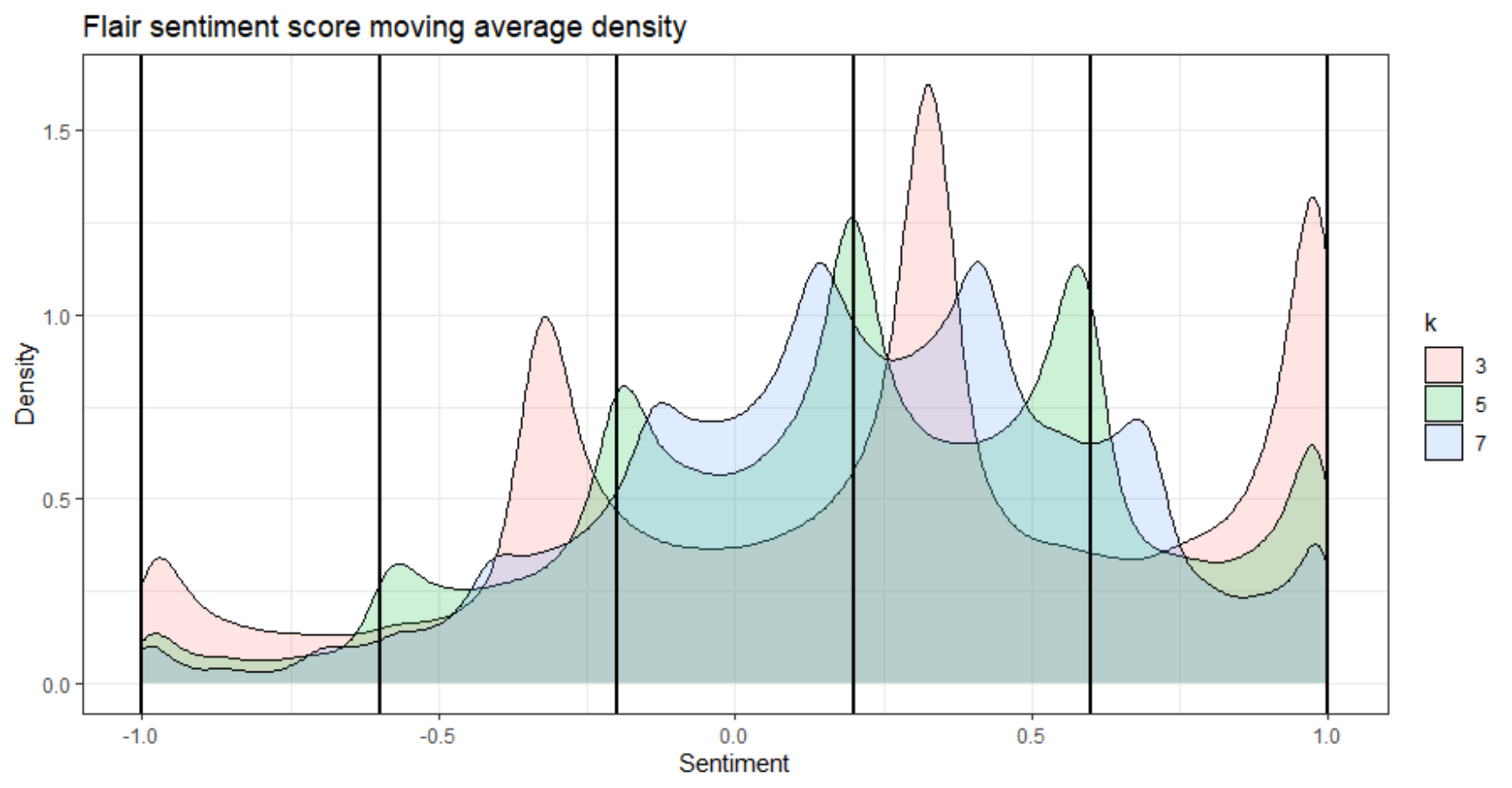
Vader – sentiment across events



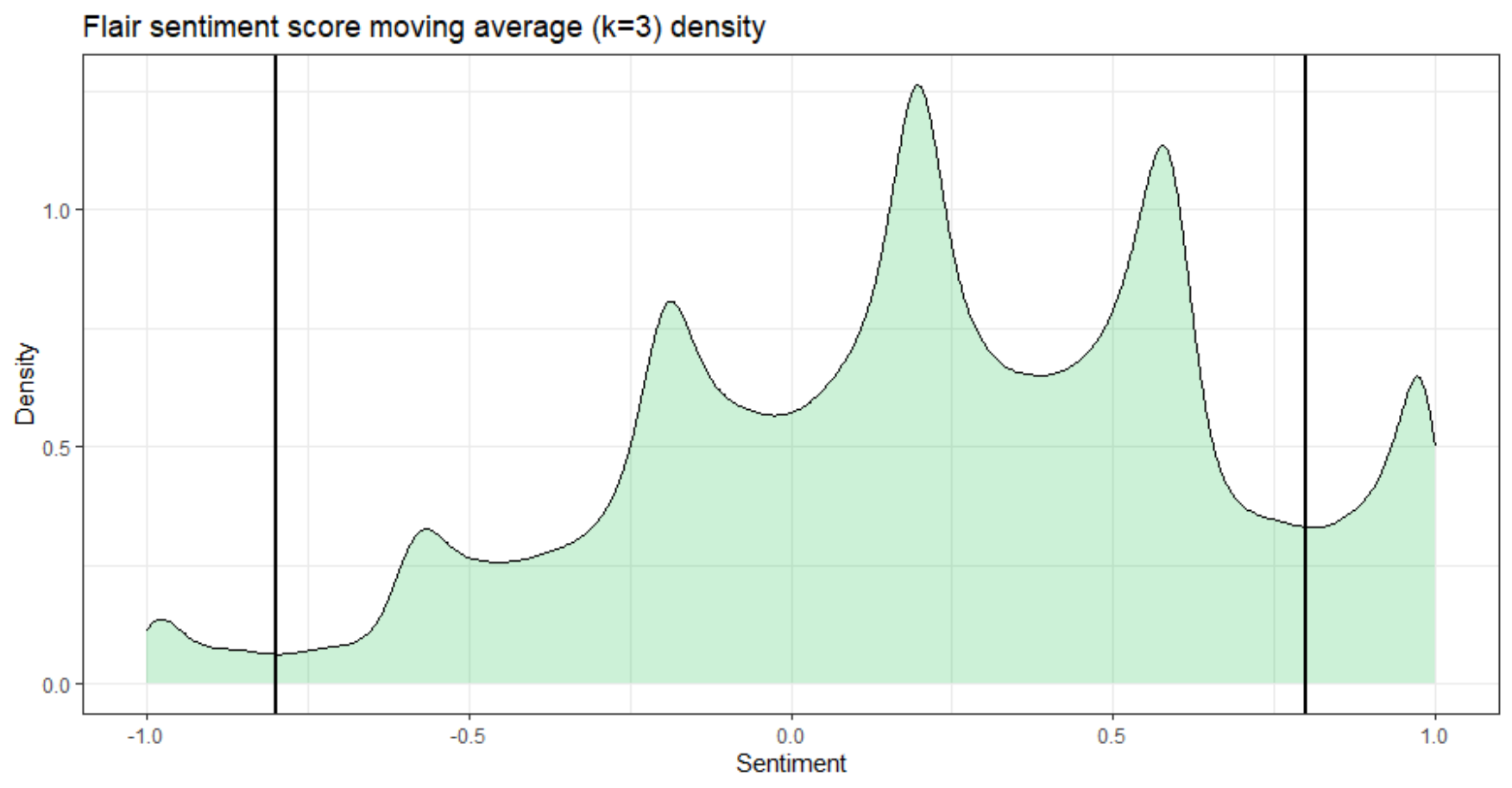
Flair – Moving Average Approach



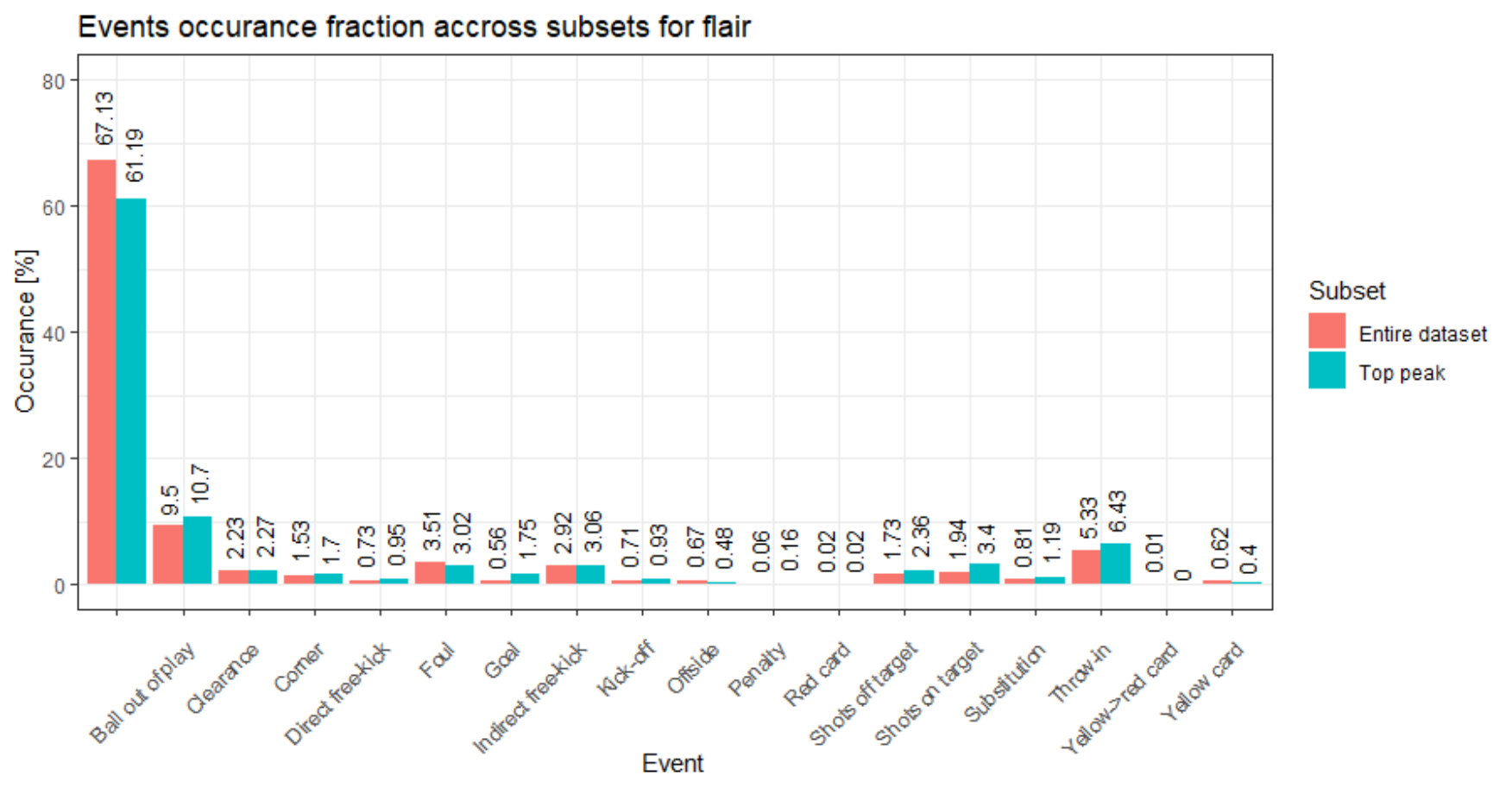
Flair – capturing the context



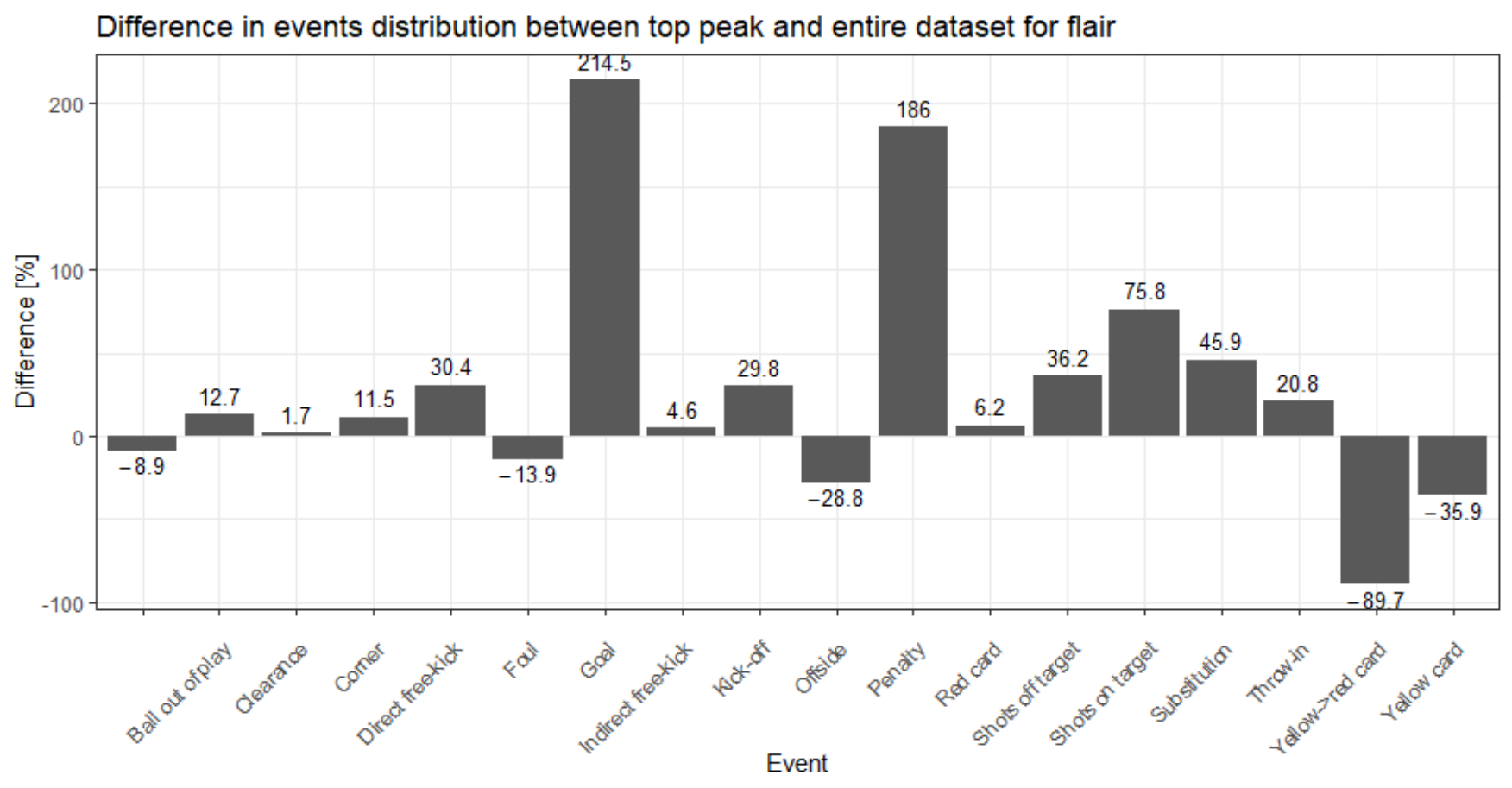
Flair – MA(5) cutoff



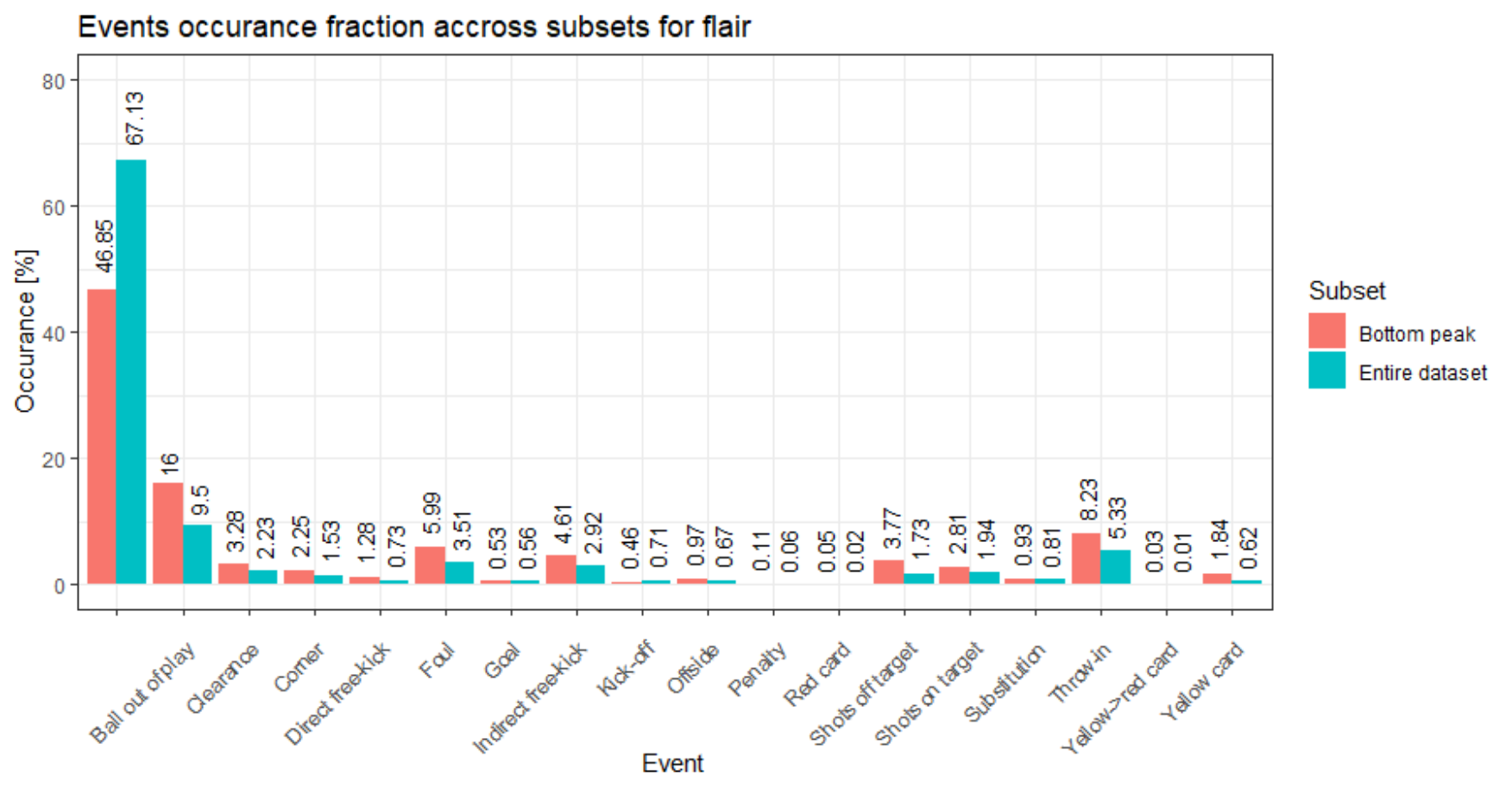
Events distribution in top peak



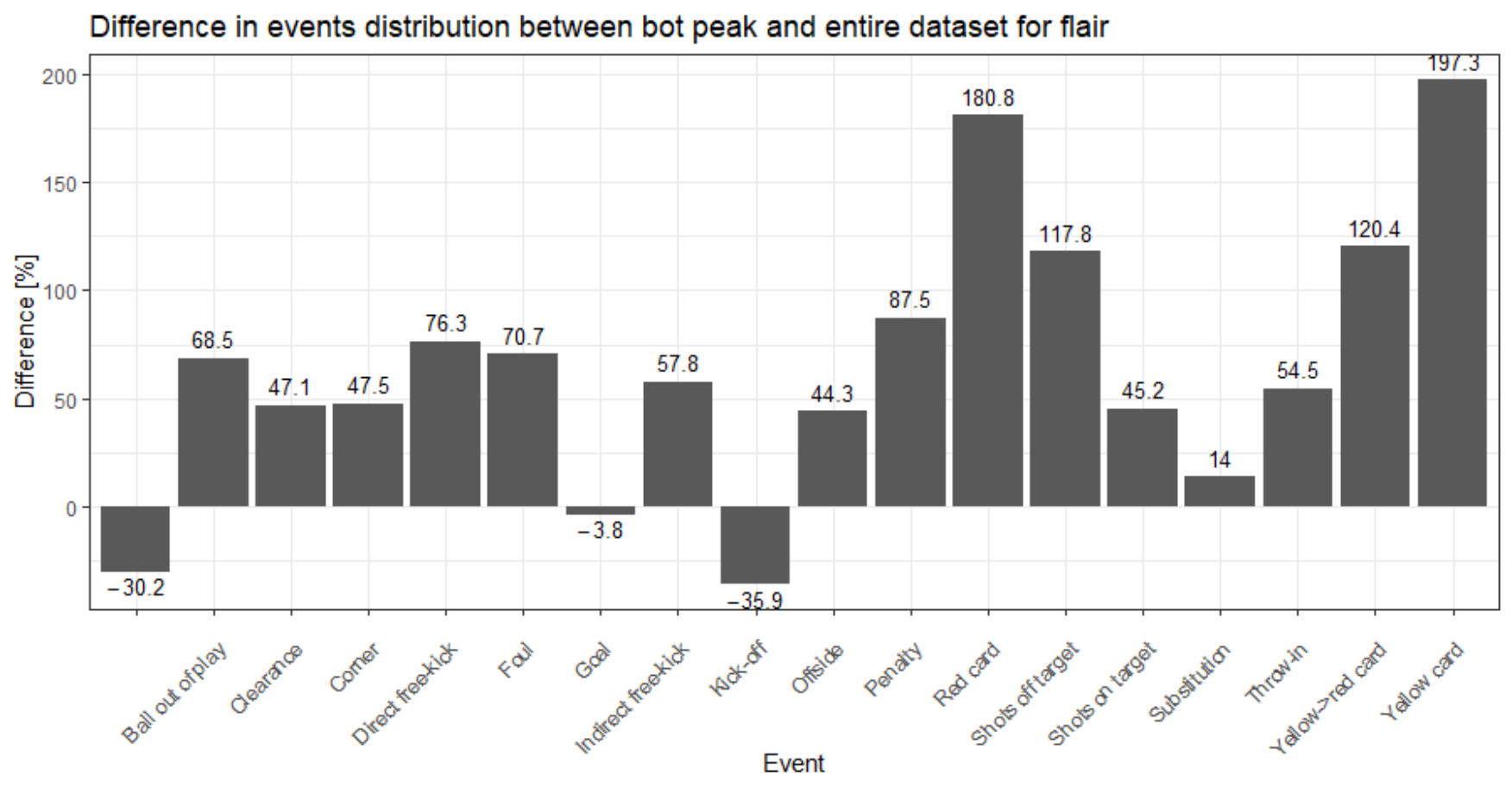
Difference in events distribution



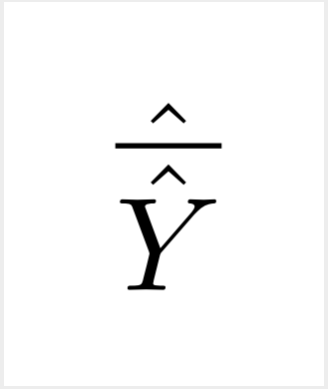
Events distribution in bottom peak



Difference in events distribution



Chasing the dreams

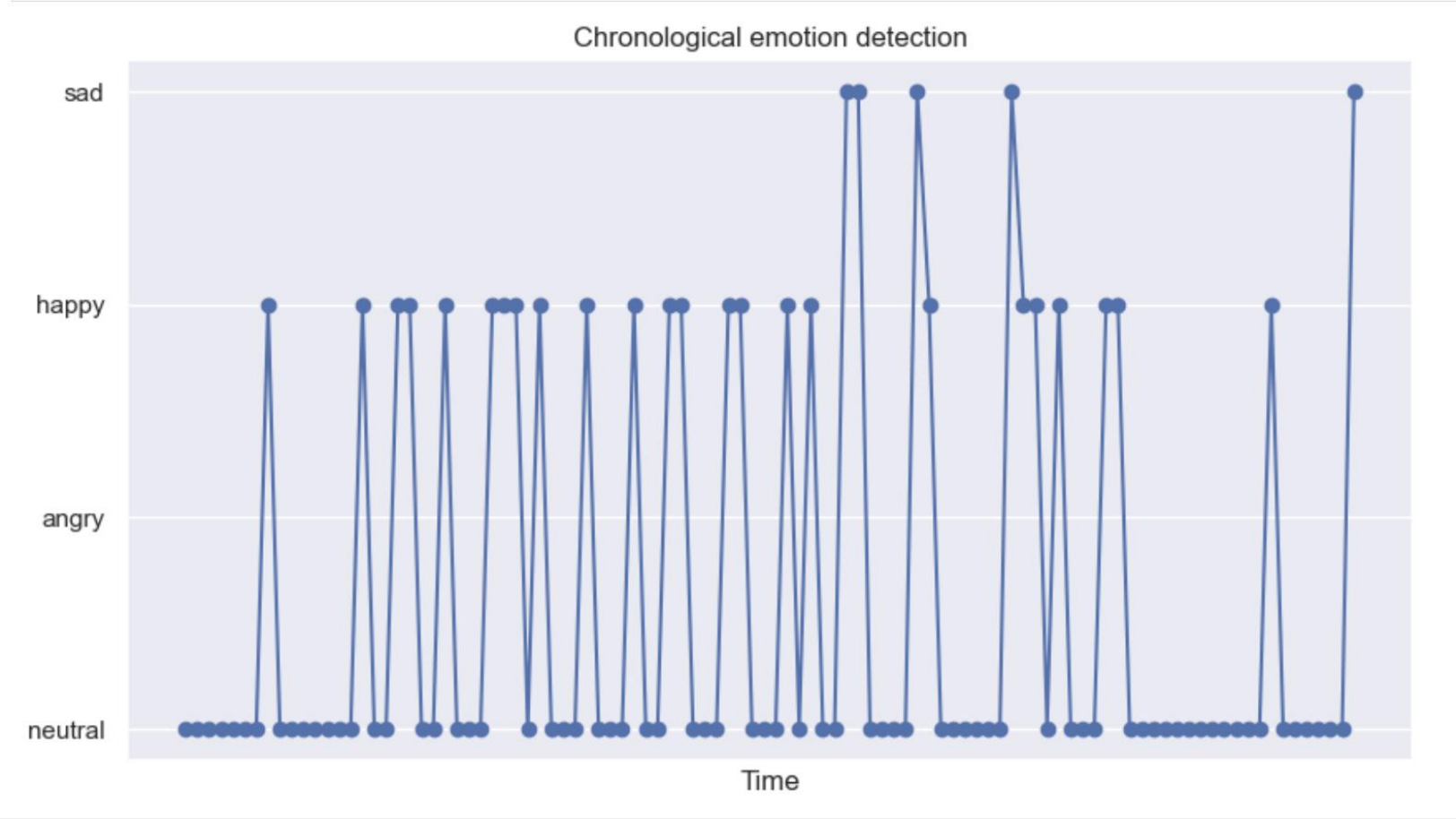


Text related research questions answers

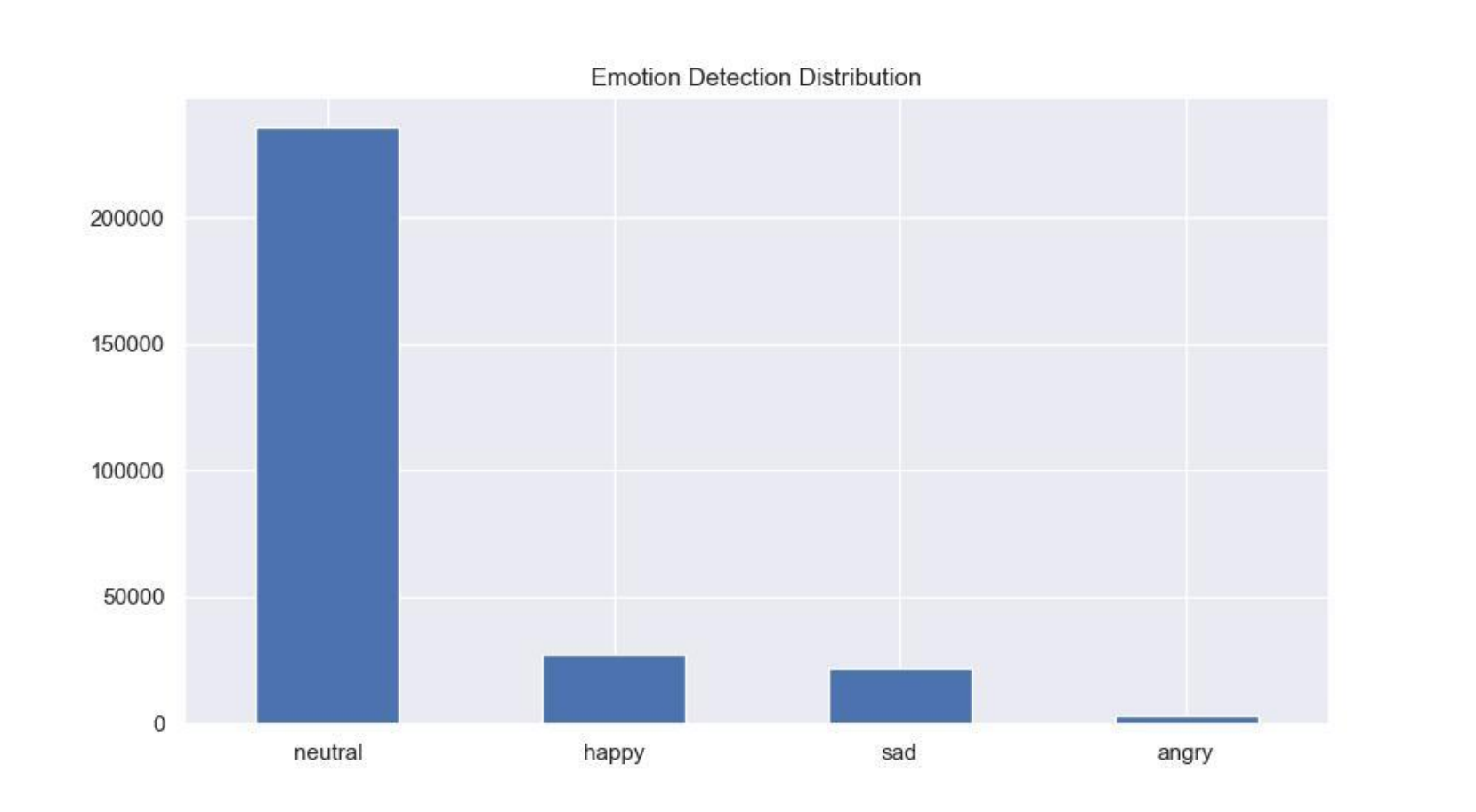
- How well does the tabulated data we have, explain the estimated sentiment?
  Answer: NO
- Statistical differences between lexical and ML approaches?
  Answer: NO
- Do we really need audio data to spot interesting relationships? Is text alone not enough?
  Answer: YES, WE NEED AUDIO!

EDA: audio analysis

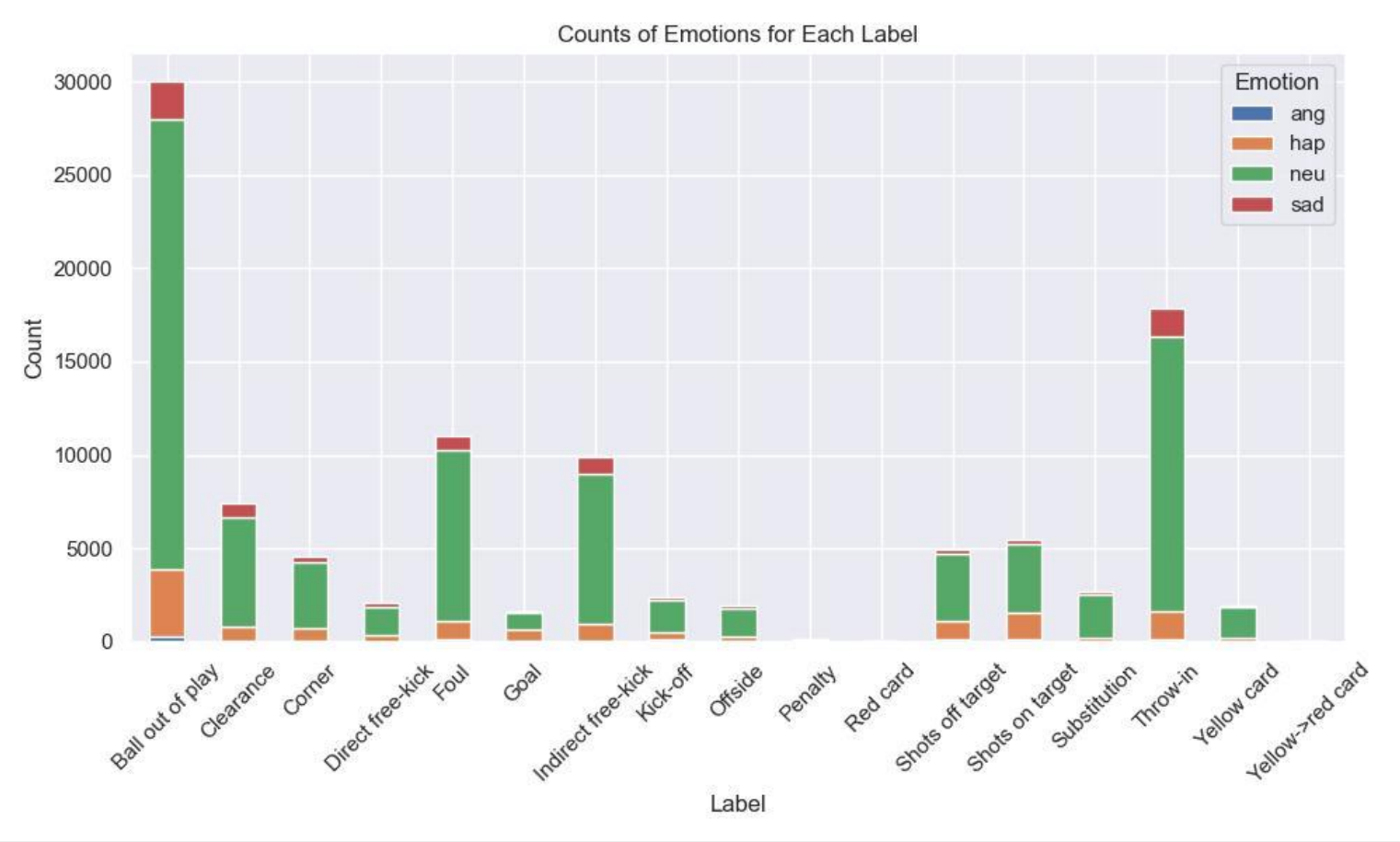
Single match example



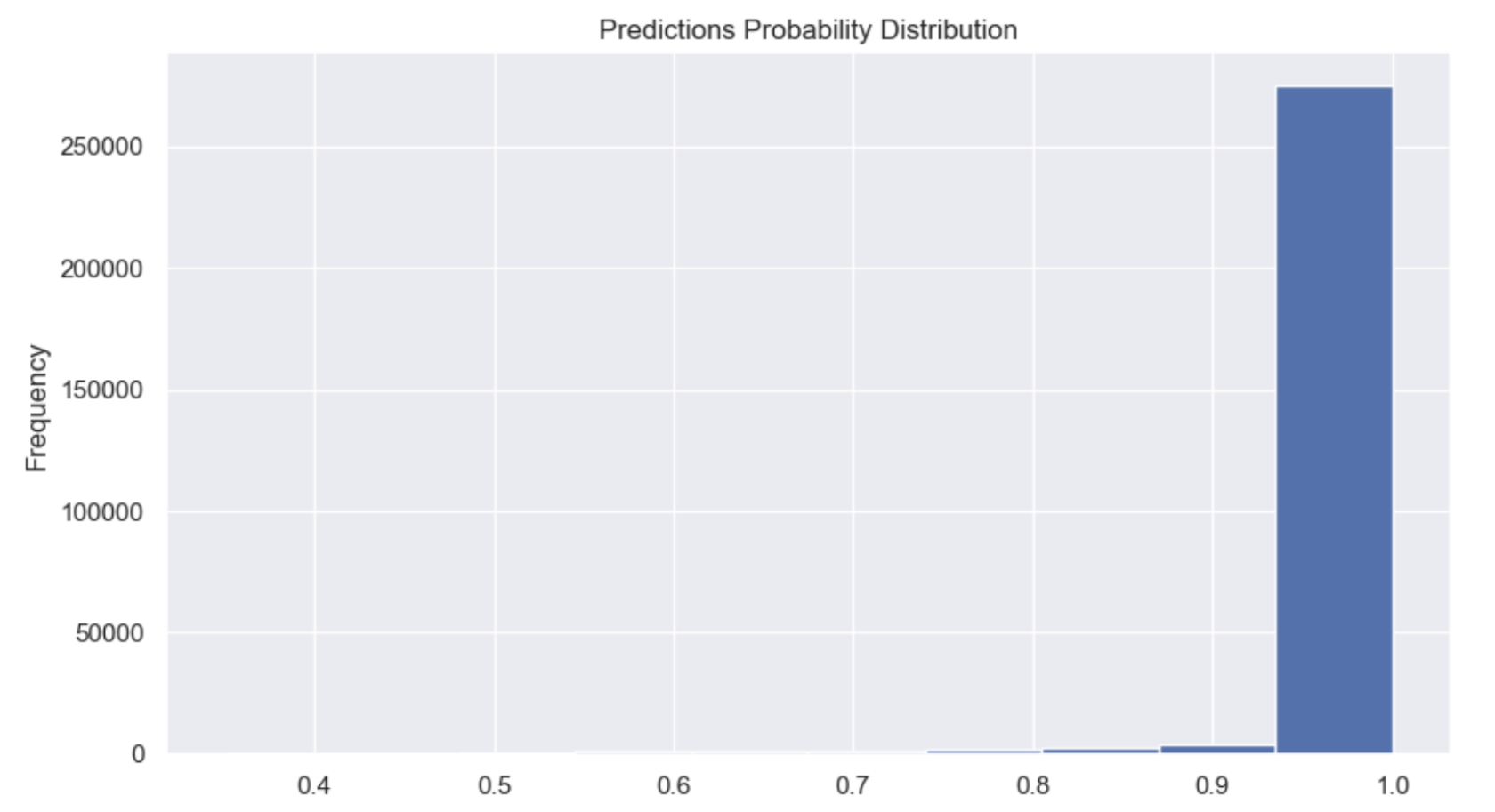
Emotion detection distribution



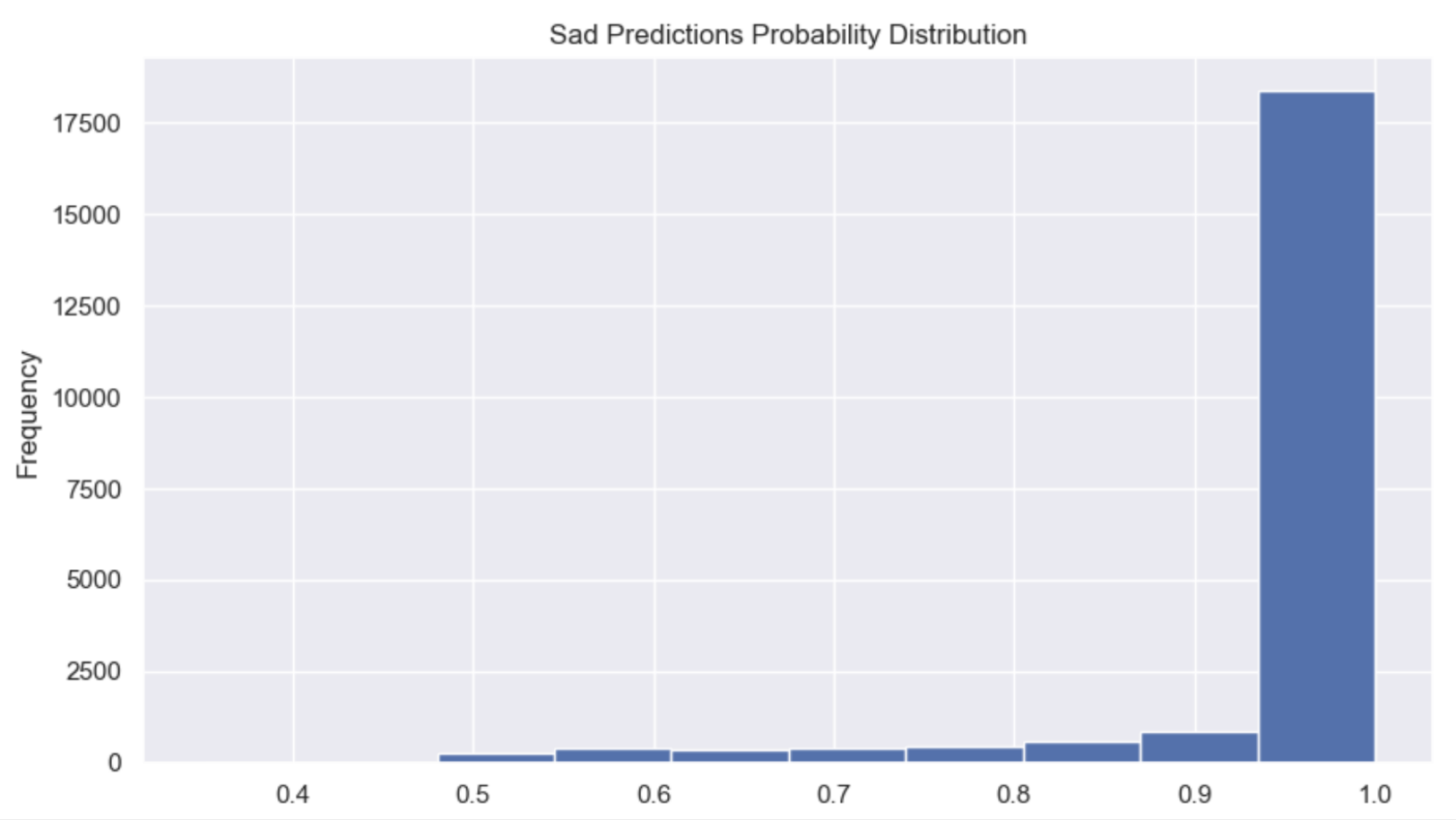
Emotion Detection across events



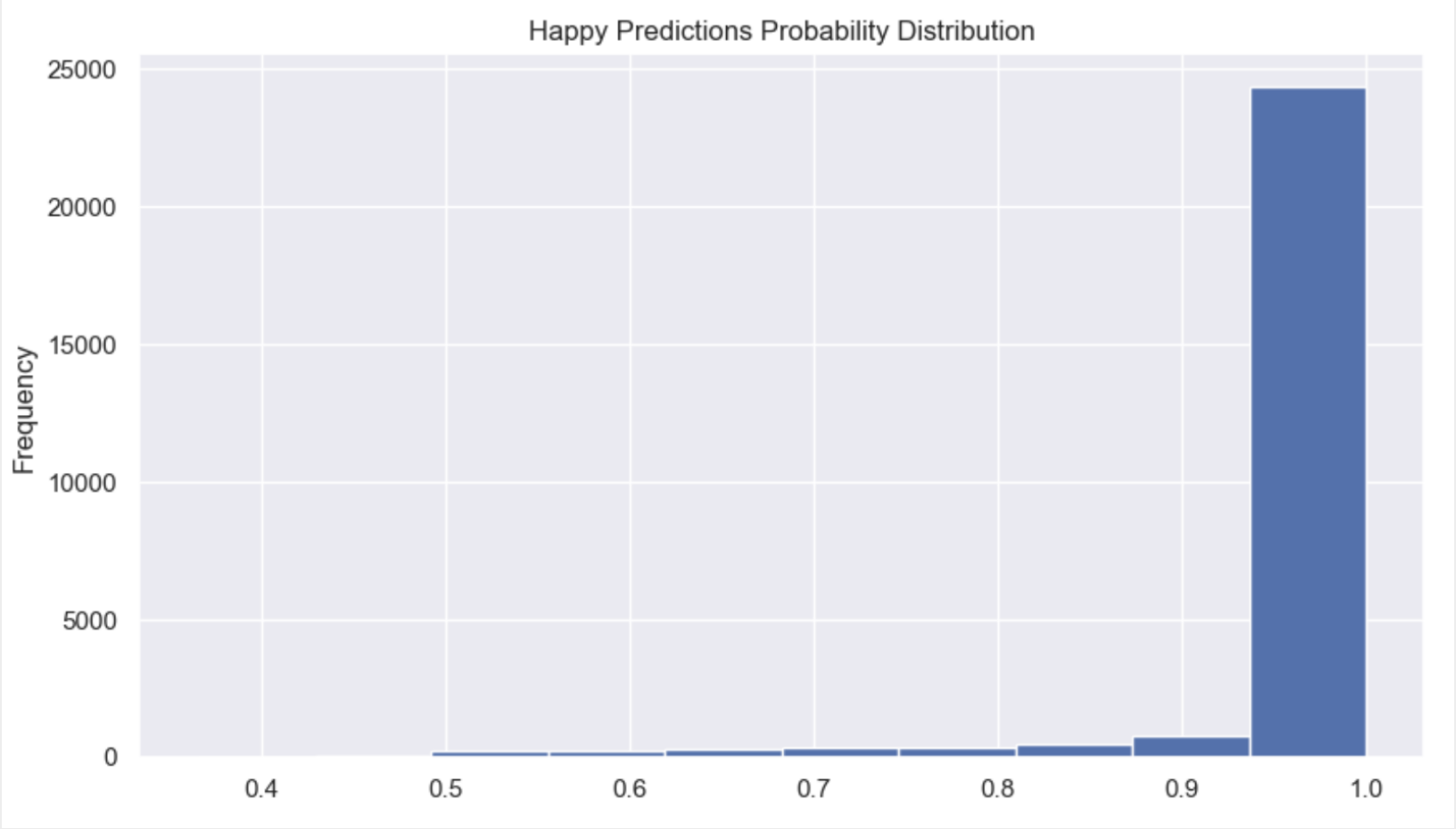
Predictions Probability Distribution



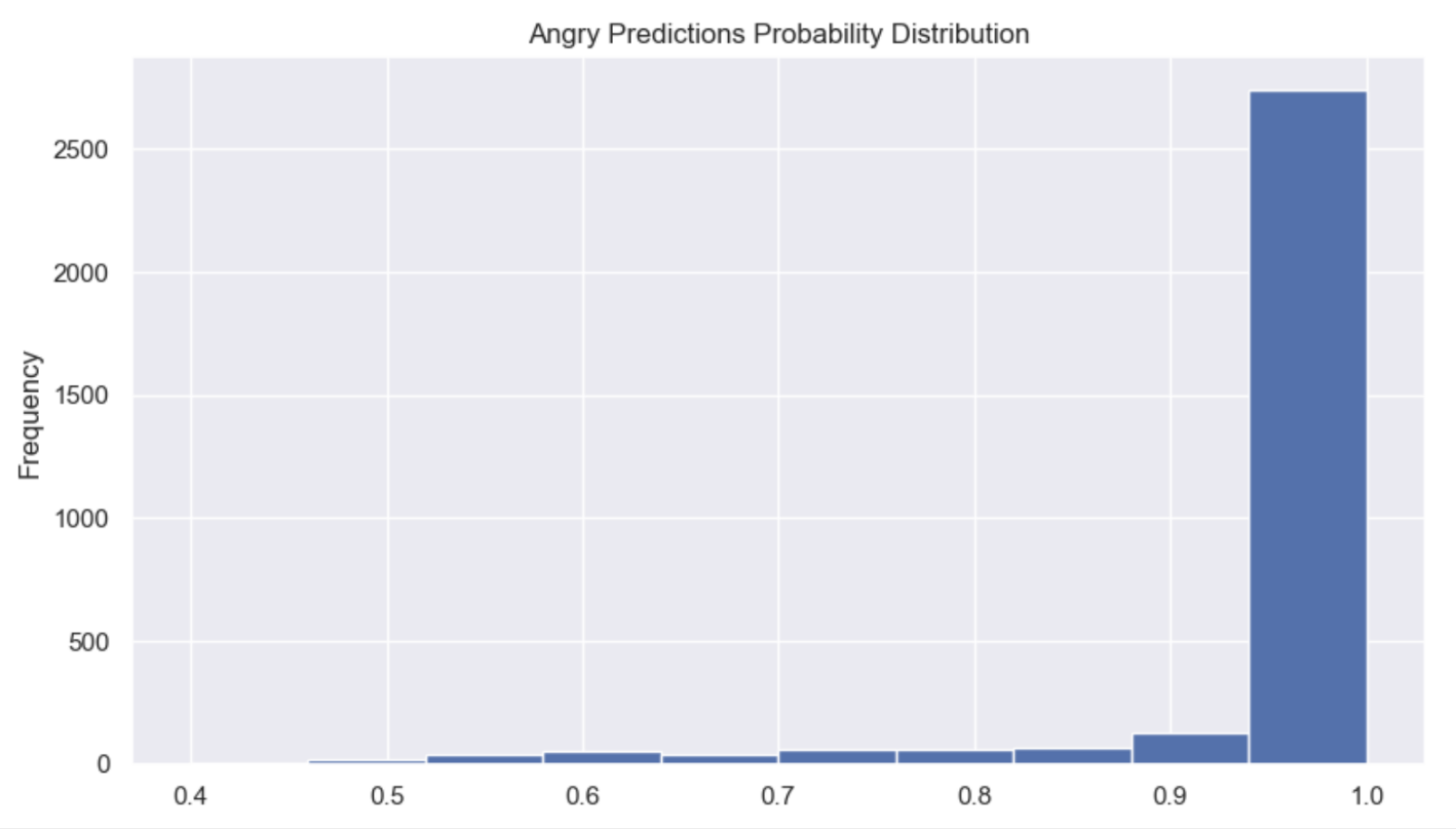
Sad Predictions Probability Distribution



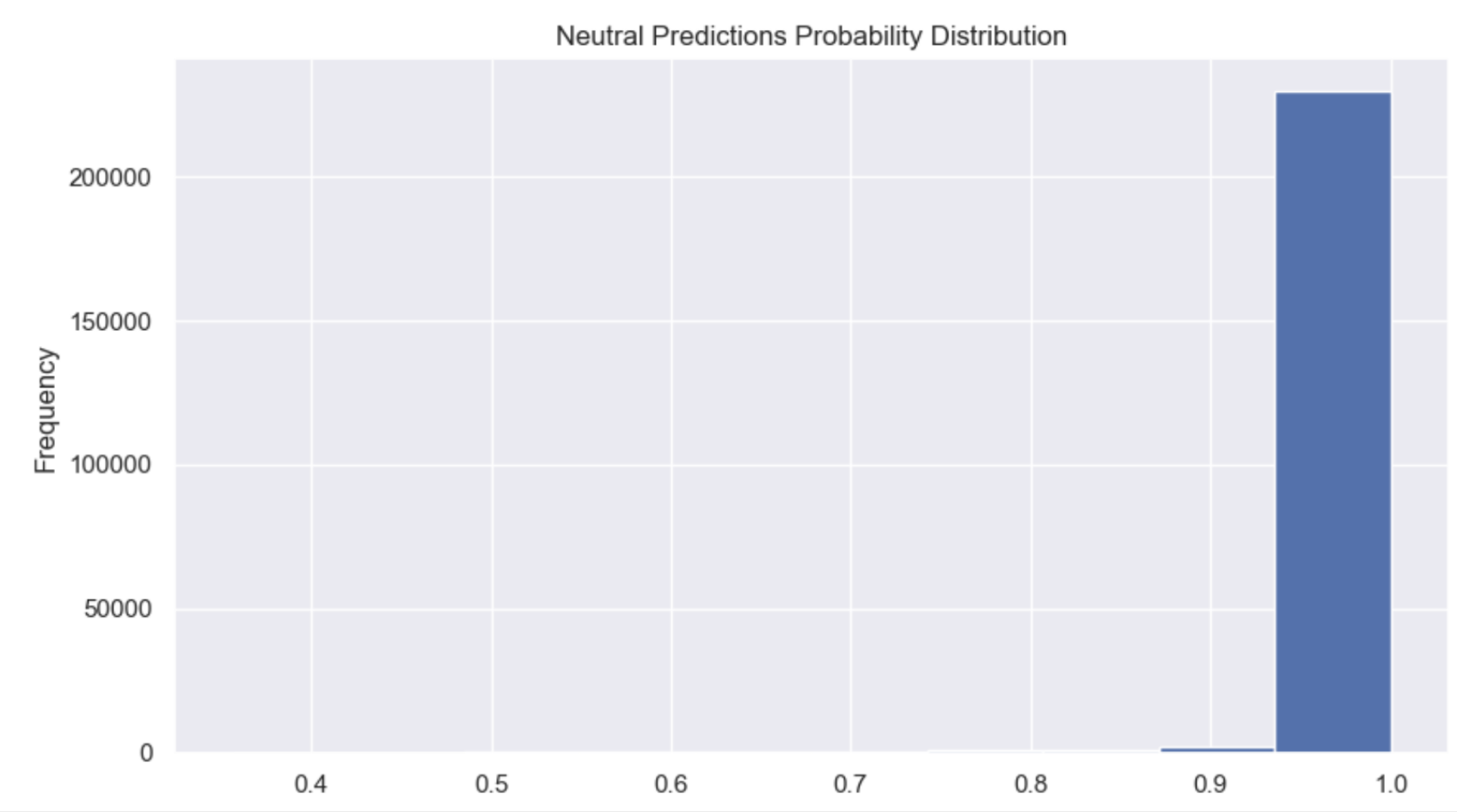
Happy Predictions Probability Distribution



Angry Predictions Probability Distributions



Neutral Predictions Probability Distributions



Answers to audio related research questions

- Can audio emotions serve as an indicator of important match events?
  Answer: NO
- Do we really need audio data to spot interesting relationships? Is text alone not enough?
  Answer: AUDIO ALONE (IN THE CURRENT FORM OF DATA PREPARATION) IS NOT ENOUGH

Modelling

Sentiment classification (Flair – Negative, Positive)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.6263	0.5829	0.6263	0.6144	0.5320	0.0668	0.1115	187.1740
xgboost	Extreme Gradient Boosting	0.6237	0.5751	0.6237	0.6046	0.5312	0.0625	0.1017	19.9180
lightgbm	Light Gradient Boosting Machine	0.6221	0.5689	0.6221	0.6501	0.4948	0.0312	0.0903	29.2900
gbc	Gradient Boosting Classifier	0.6160	0.5502	0.6160	0.6090	0.4771	0.0089	0.0395	532.7720
ada	Ada Boost Classifier	0.6149	0.5424	0.6149	0.5771	0.4835	0.0116	0.0352	112.0540
lda	Linear Discriminant Analysis	0.6149	0.5293	0.6149	0.5947	0.4712	0.0030	0.0207	17.1580
ridge	Ridge Classifier	0.6148	0.0000	0.6148	0.5943	0.4711	0.0029	0.0203	6.3260
dummy	Dummy Classifier	0.6145	0.5000	0.6145	0.3776	0.4677	0.0000	0.0000	6.8100
lr	Logistic Regression	0.6144	0.5265	0.6144	0.5061	0.4677	0.0000	0.0006	19.5060
qda	Quadratic Discriminant Analysis	0.5984	0.5313	0.5984	0.5472	0.5263	0.0253	0.0335	14.9280
nb	Naive Bayes	0.5894	0.5205	0.5894	0.5418	0.5325	0.0226	0.0273	7.3180
rf	Random Forest Classifier	0.5881	0.5726	0.5881	0.5763	0.5800	0.1038	0.1048	507.8360
et	Extra Trees Classifier	0.5825	0.5673	0.5825	0.5751	0.5780	0.1026	0.1030	233.4040
knn	K Neighbors Classifier	0.5792	0.5750	0.5792	0.5722	0.5744	0.0959	0.0965	620.6680
dt	Decision Tree Classifier	0.5658	0.5362	0.5658	0.5608	0.5630	0.0728	0.0729	61.4480
svm	SVM - Linear Kernel	0.4559	0.0000	0.4559	0.4549	0.3515	0.0012	0.0025	151.1920

Sentiment classification (Flair – Negative, Neutral, Positive)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.5572	0.5714	0.5572	0.5812	0.4402	0.0530	0.0984	38.0800
catboost	CatBoost Classifier	0.5556	0.5705	0.5556	0.5982	0.4311	0.0437	0.0897	393.9500
lightgbm	Light Gradient Boosting Machine	0.5541	0.5645	0.5541	0.6209	0.4134	0.0290	0.0851	53.0200
gbc	Gradient Boosting Classifier	0.5481	0.5485	0.5481	0.5870	0.3962	0.0095	0.0417	1256.1200
ada	Ada Boost Classifier	0.5464	0.5334	0.5464	0.5032	0.3984	0.0086	0.0294	88.3367
dummy	Dummy Classifier	0.5460	0.5000	0.5460	0.2981	0.3857	0.0000	0.0000	7.7867

Events detection from audio emotions

	Classifier	Accuracy	AUC	Precision	Recall	F1 Score
0	Logistic Regression	0.706198	0.500683	0.000000	0.000000	0.000000
1	XGBoost	0.705885	0.502156	0.263158	0.000592	0.001181
2	CatBoost	0.706181	0.502027	0.333333	0.000059	0.000118
3	Dummy Classifier	0.706198	0.500000	0.000000	0.000000	0.000000
4	AdaBoost	0.706198	0.500010	0.000000	0.000000	0.000000

Plans for the future (Project 2)

- Analyse accordance beetween audio emotions and text sentiment.
- Add audio emotions as another feature in the statistical exploration.
- Compare text sentiment and audio emotions accorss different languages (or translations).
- Post events emotions intensification assesement.
- Employ Recurrent Neural Network for game events prediction.