SOCCER COMMENTARY MINING

Natural Language Processing
Project 1 proposal

Szymon Maksymiuk
Adam Narożniak
Władysław Olejnik
Patryk Świątek

WHY IS IT INTERESTING?

WHY IS IT INTERESTING?

Unexplored – we have not found previous work about analysing boradcast's transcriptions.

Accessible – everyone at least one has seen a football match and knows basic rules.

Challenging – data from three different sources – audio, text and tabular events.

Part of the bigger picture – chance to contribiute in a field where recent years were marked by rapid introduction of various statistical approches.

WHY IS IT INTERESTING?

WHY IS IT INTERESTING?

WHY IS IT INTERESTING?

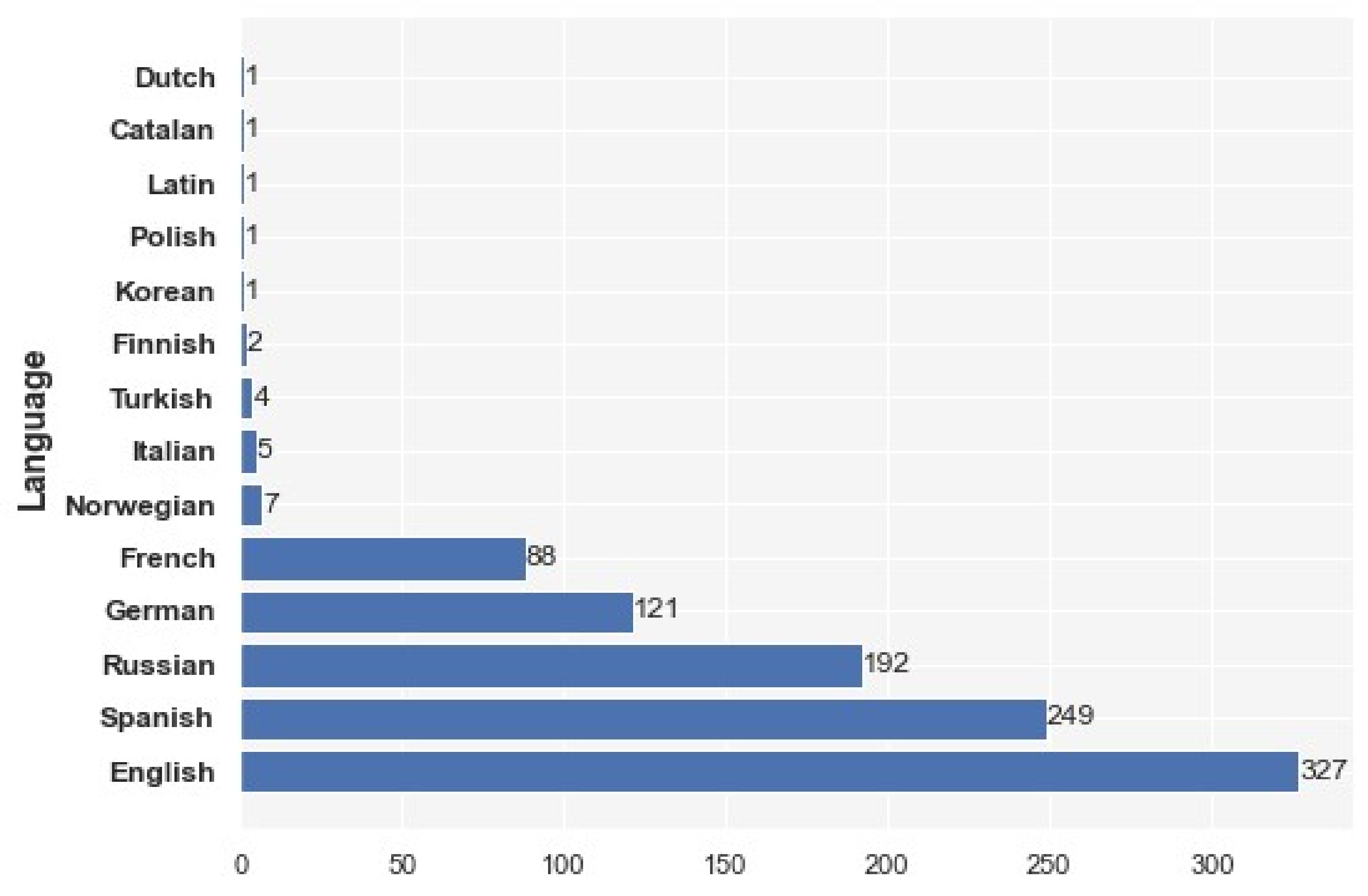
DATASET

SOCCERNET

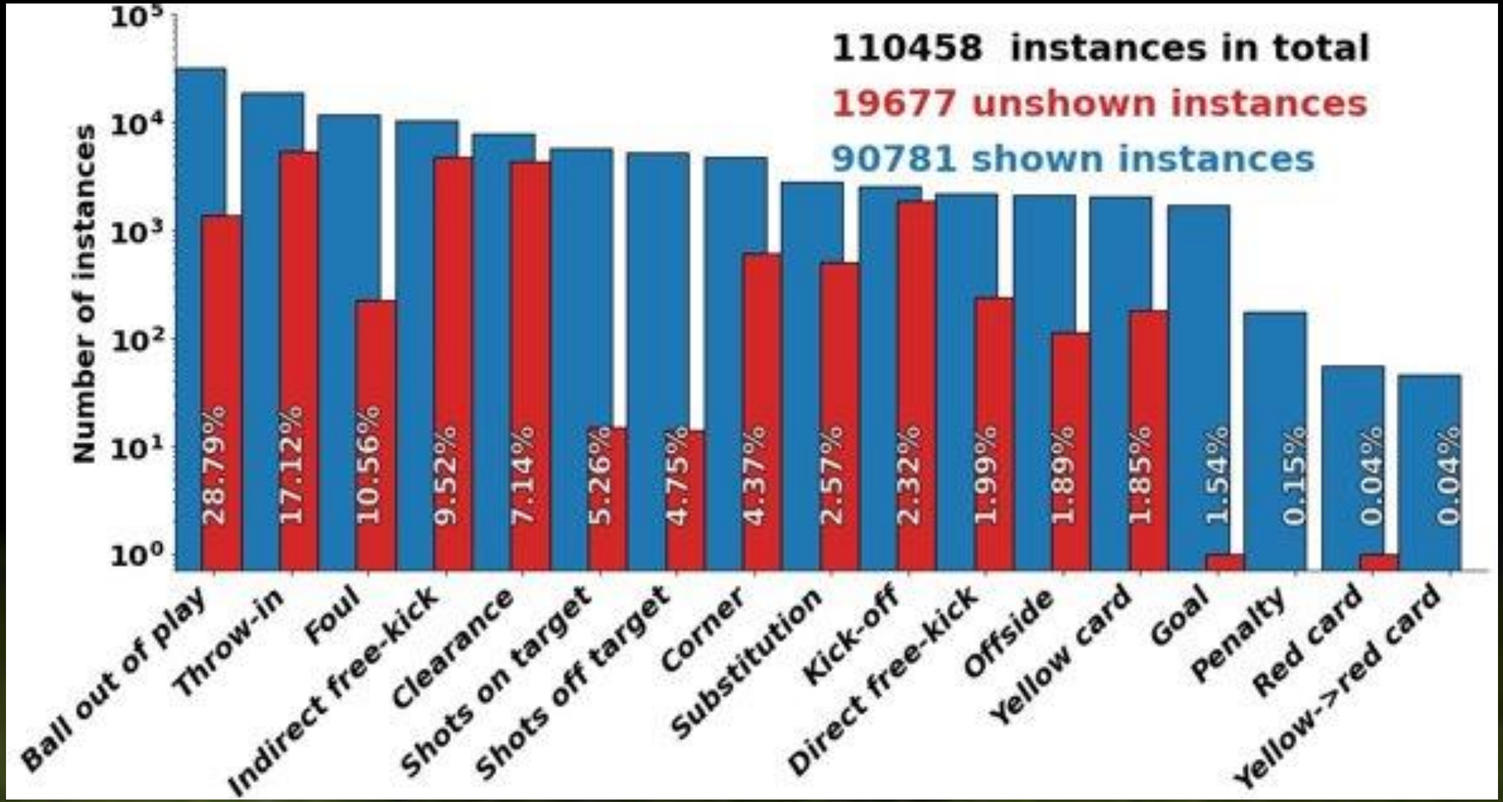
League	Seasons 14/15	15/16	16/17	Total
EN - EPL	6	49	40	**95**
ES - LaLiga	18	36	63	**117**
FR - Ligue 1	1	3	34	**38**
DE - BundesLiga	8	18	27	**53**
IT - Serie A	11	9	76	**96**
EU - Champions	37	45	19	**101**
Total	**81**	**160**	**259**	**500**

- Opensource dataset.
- 500 games from the main European Championships.
- Two untrimmed videos for one match (each for one half period).
- An average of 13 labels per match:
  - Goal
  - Yellow/Red Card
  - Substitution
- Exact minute of the match for each label.

LANGUAGE


Language
Dutch 1
Catalan 1
Latin 1
Polish 1
Korean 1
Finnish 2
Turkish 4
Italian 5
Norwegian 7
French 88
German 121
Russian 192
Spanish 249
English 327
0 50 100 150 200 250 300

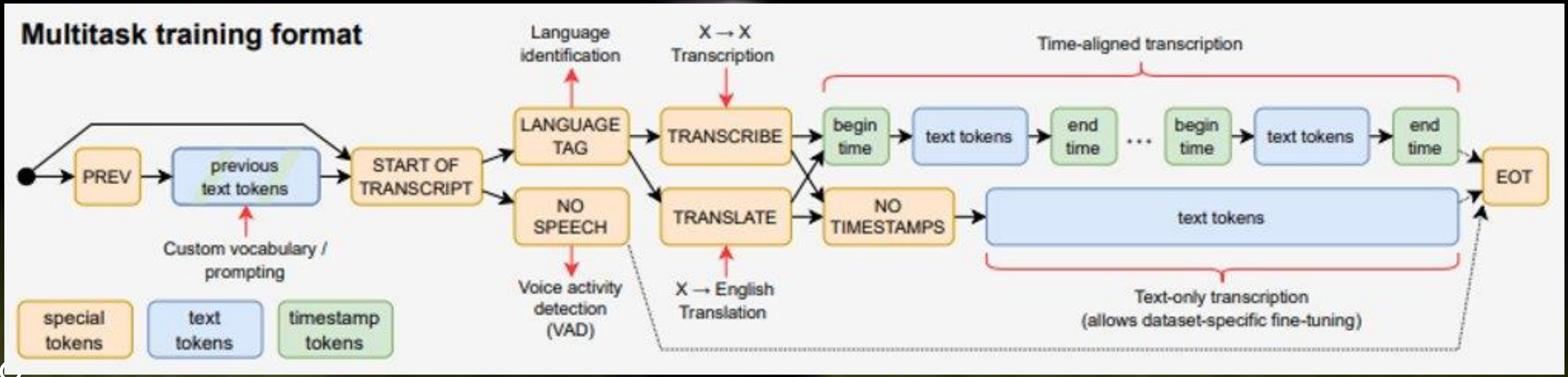
SOCCERNET-V2



- Updated version of SoccerNet.
- 17 types of labels instead of 3.
- An average of 221 labels per match (17 times more than for the previous version).

TRANSCRIPTS GENERATION: WHISPER

- ASR model developed by OpenAI.
- Transformer encoder-decoder framework.
- Trained on different speech processing tasks:
  - multilingual speech recognition,
  - speech translation,
  - spoken language identification,
  - voice activity detection.



SOCCER COMMENTARY MINING

RESEARCH QUESTIONS

RESEARCH QUESTIONS

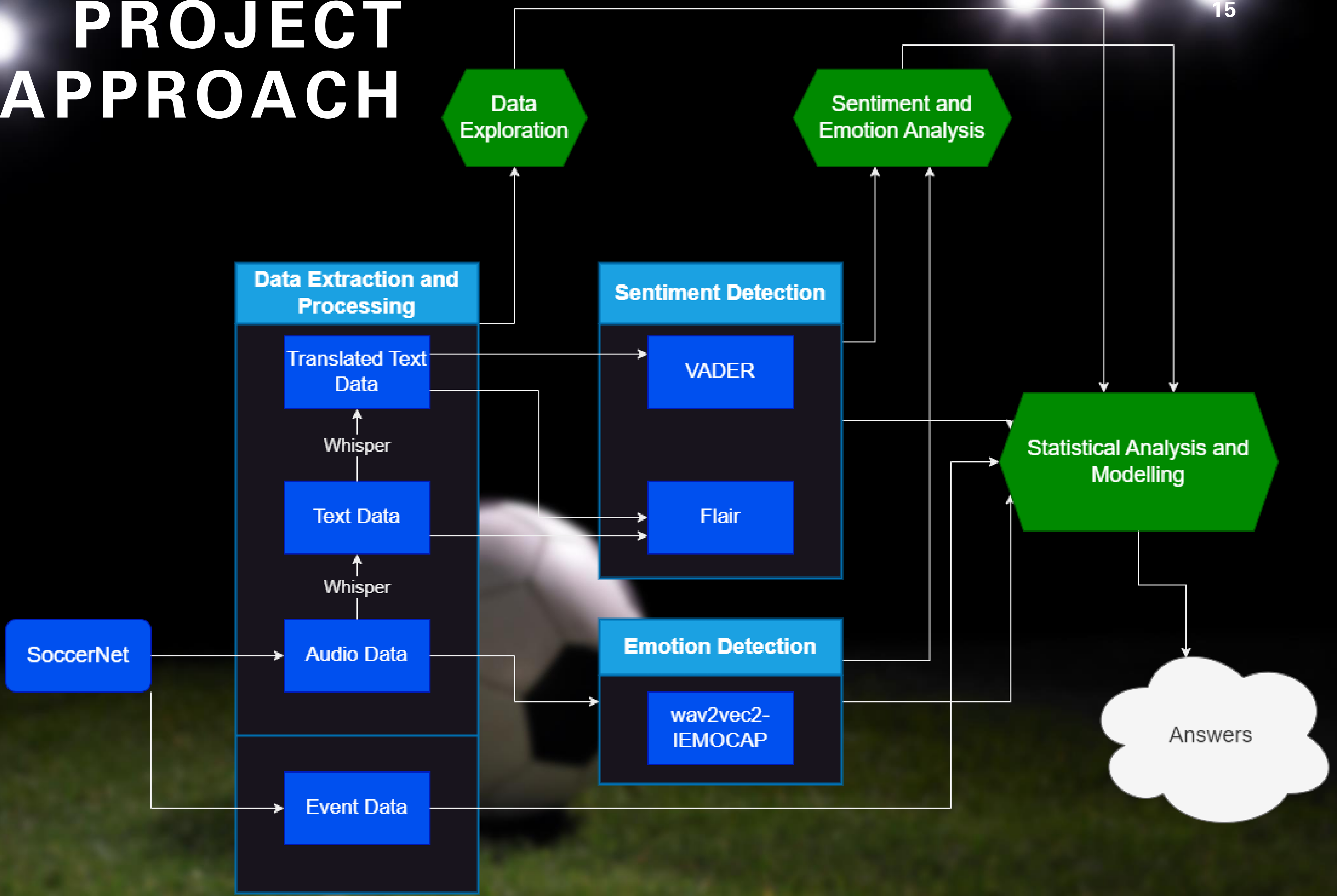
- **How** well does the tabulated data we have, explain the estimated sentiment?
- **Statistical** differences between lexical and ML approaches?
- **Can** emotion intensification serve as an indicator of an important event?
- **What** are the differences between sentiment from text and emotion from audio?
- **Do** we really need audio data to spot interesting relationships? Is text alone not enough?
- **Do** commentators of different nationalities react differently to the same events during a match?
- **How** much and for how long do emotions intensify after important events in a match?
- **Can** emotion intensification serve as an indicator of a future important event?

SOCCER COMMENTARY MINING

PROJECT APPROACH

PROJECT APPROACH

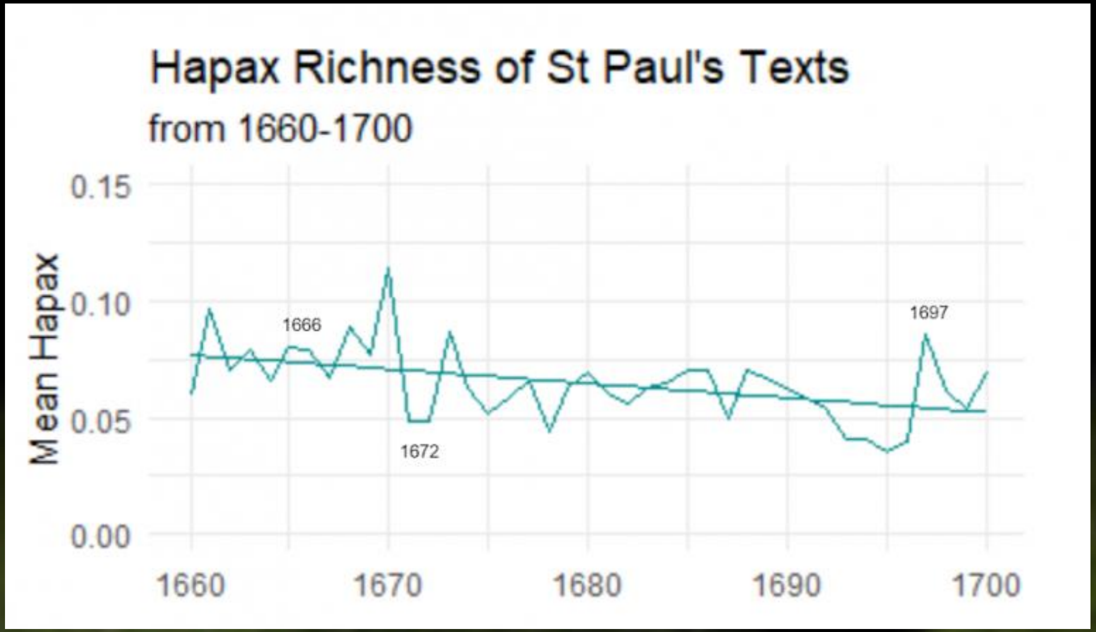
SOCCER COMMENTARY MINING



PRELIMINARY DATA EXPLORATION

PRELIMINARY DATA EXPLORATION

- Tokens vs types analysis
- Average statement length acorss languages
- Filtering proper names
- Words repetition and stopwords count
- Overall language complexity:
  - Lexical readability (Flesch-Kinacid)
  - Lexical richness (TTR, Hapax richness)




SOCCER
COMMENTARY
MINING
SENTIMENT
ANALYSIS

LEXICON-BASED "MODELS"

- LEXICON = DICTIONARY
- Pros:
  - Interpretable
- Cons:
  - Creation is time consuming (manual creation)
  - Dictionary size limited by the design
  - Based on expert knowledge

SEMANTIC ORIENTATION LEXICONS
(POLARITY-BASED)

Model	Example	Scores
POLARITY-BASED	This pass is **good**.	+1
POLARITY-BASED	This pass is **amazing**.	+1
POLARITY-BASED	This pass is **not bad**.	-1

LIWC Category	Examples	No. of Words
Positive Emotion	Love, nice, good, great	406
Negative Emotion	Hurt, ugly, sad, bad, worse	499

Table 1: Example words from two of LIWC's 76 categories. These two categories can be leveraged to construct a semantic orientation-based lexicon for sentiment analysis.

SENTIMENT INTENSITY LEXICONS
(VALENCE-BASED)

Model	Example	Scores
VALENCE-BASED	This pass is **good**.	+0.5
VALENCE-BASED	This pass is **amazing**.	+1.5
VALENCE-BASED	This pass is **not bad**.	-1.0

CONTEXT AWARE LEXICONS

Model	Example	Scores
CONTEXT AWARE	This pass is **good**.	+0.5
CONTEXT AWARE	This pass is **amazing**.	+1.5
CONTEXT AWARE	This pass is **not bad**.	**+1.0**

VADER

Social Media Text (4,200 Tweets)

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score	Ordinal Rank (by F1)
Ind. Humans	**0.888**	0.95	0.76	0.84	2
VADER	0.881	**0.99**	**0.94**	**0.96**	1*
Hu-Liu04	0.756	0.94	0.66	0.77	3
SCN	0.568	0.81	0.75	0.75	4
GI	0.580	0.84	0.58	0.69	5
SWN	0.488	0.75	0.62	0.67	6
LIWC	0.622	0.94	0.48	0.63	7
ANEW	0.492	0.83	0.48	0.60	8
WSD	0.438	0.70	0.49	0.56	9

Movie Reviews (10,605 review snippets)

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score	Ordinal Rank (by F1)
Ind. Humans	0.899	0.95	0.90	0.92	1
VADER	0.451	0.70	0.55	0.61	2
Hu-Liu04	0.416	0.66	0.56	0.59	3
SCN	0.210	0.60	0.53	0.44	7
GI	0.343	0.66	0.50	0.55	5
SWN	0.251	0.60	0.55	0.57	4
LIWC	0.152	0.61	0.22	0.31	9
ANEW	0.156	0.57	0.36	0.40	8
WSD	0.349	0.58	0.50	0.52	6

Amazon.com Product Reviews (3,708 review snippets)

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score	Ordinal Rank (by F1)
Ind. Humans	0.911	0.94	0.80	0.85	1
VADER	0.565	0.78	0.55	0.63	2
Hu-Liu04	0.571	0.74	0.56	0.62	3
SCN	0.316	0.64	0.60	0.51	7
GI	0.385	0.67	0.49	0.55	5
SWN	0.325	0.61	0.54	0.57	4
LIWC	0.313	0.73	0.29	0.36	9
ANEW	0.257	0.69	0.33	0.39	8
WSD	0.324	0.60	0.51	0.55	6

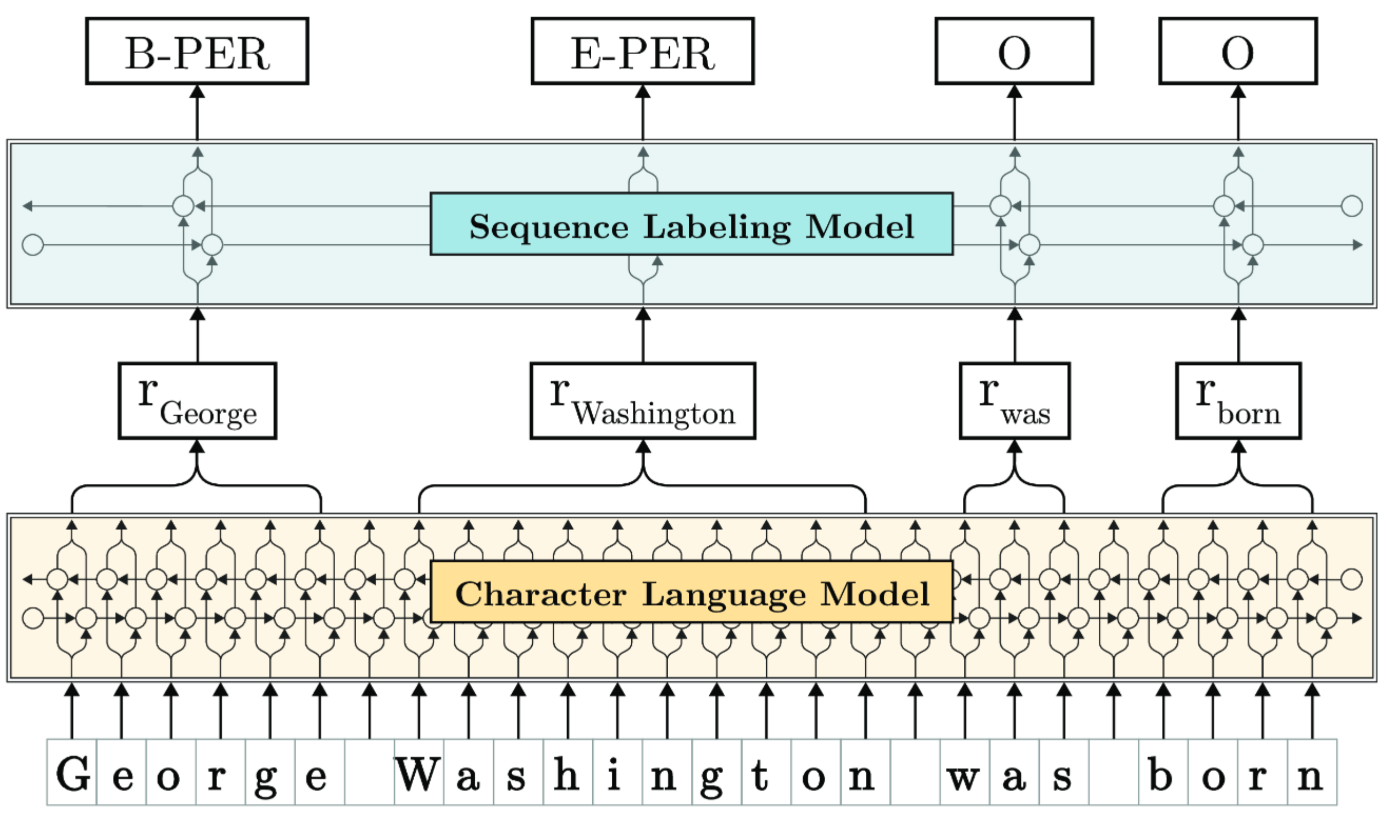
NY Times Editorials (5,190 article snippets)

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics: Overall Precision	Overall Recall	Overall F1 score	Ordinal Rank (by F1)
Ind. Humans	0.745	0.87	0.55	0.65	1
VADER	0.492	0.69	0.49	0.55	2
Hu-Liu04	0.487	0.70	0.45	0.52	3
SCN	0.252	0.62	0.47	0.38	7
GI	0.362	0.65	0.44	0.49	5
SWN	0.262	0.57	0.49	0.52	4
LIWC	0.220	0.66	0.17	0.21	9
ANEW	0.202	0.59	0.32	0.35	8
WSD	0.218	0.55	0.45	0.47	6

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

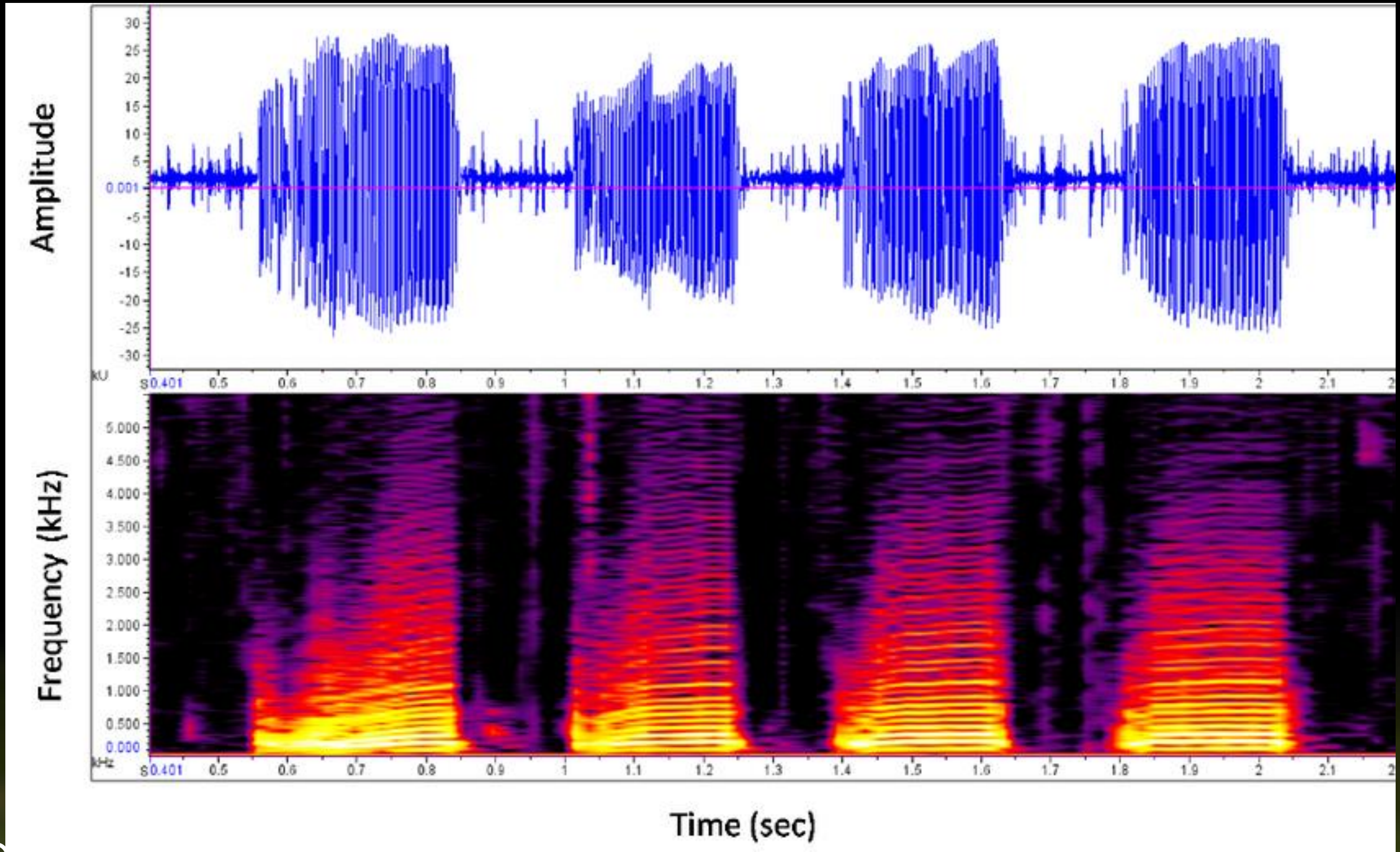
- Lexicon-based
- Rule-base

FLAIR



- Powerfull framework which utilizes different models and approaches
- Supports many types of text analysis, including sequence labeling, text classification, similarity learning and text regression.
- Supports many embeddings including GloVe, FastText, ELMo, BERT, XLM, Byte Pair Embeddings and of course our own Flair embeddings:
- Includes a „model zoo” of pre-trained models ready to use.
- Multilingual

SPEECH EMOTION DETECTION



- Typically deep neural networks
- Features might be:
  - Raw audio files
  - (mel)Spectrograms
  - MFCCs
  - Embeddings (e.g. created from self-supervised wav2vec2.0)
- |Emotions | > |Sentiments|
- Emotions involve anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. But might include also e.g. frustration. (Dataset dependent)

SOCCER COMMENTARY MINING

Q & A