Philipp Stangor & Lars Kürzinger

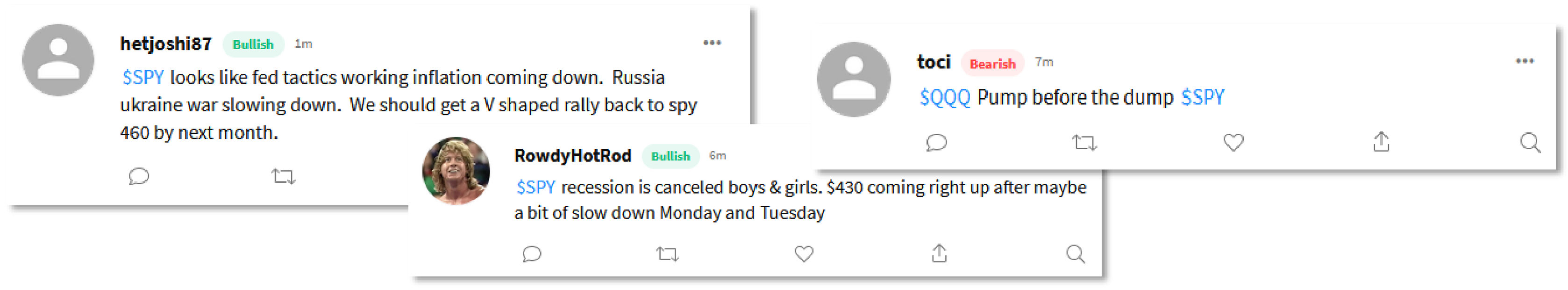
Measuring Investor Sentiment from Social Media Data – An Emotional Approach

I. THEORY & DATA

- Financial markets tend to be not fully information-efficient due to the existence of **cost for information**, transactions, etc. (see economic research starting from FAMA (1970))
- Following GROSSMANN/STIGLITZ (1980) little excess returns are possible as compensation for **continous information collection**
- Use of **information cost reducing institutions** can be efficient

→ **Social Media Data** centralizes, selects & verificates information and can be transformed to sentiment with **textual analysis**

- In 2019 more than 15,000 active users per day share their ‚ideas' about financial topics on the microblogging platform **StockTwits** like this:



DATA FACTS

- $N = 173{,}600{,}190$ ideas between 01/12 and 12/19
- Users can tag their ideas ‚bullish' or ‚bearish'
  - $N_{Bu} = 52{,}174{,}591$ bullish tagged ideas (30.05%)
  - $N_{Be} = 37{,}894{,}760$ bearish tagged ideas (8.22%)
- Creation time of ideas implies users mainly talk about US stock markets
- Collected via developer API

LONG STORY SHORT

In our analysis we employ a multidimensional approach extracting investor sentiment from social media data using the NRC-Emotion Association Lexicon. Considering a vast number of short text messages from the financial microblogging platform StockTwits **(I)**, we analyze eight different emotions contained in each message **(II)**. Subsequently, we classify these posts as ‘bullish’ or ‘bearish’ signals on basis of their emotional profile using machine learning techniques to develop aggregated investor sentiment **(III)**. Further, we use this to forecast intraday returns of the NASDAQ 100 **(IV)**.

II. EMOTION SCORING

- Starting with HENRY (2008) and LOUGHRAN/MCDONALD (2011) **positive-negative** dictionaries further have been implemented and improved in economic literature
- Meanwhile, linguistic research introduced **multidimensional approaches**
  - **Reflects complexity** of language better
  - For example via emotions as ‚**EmoLex**' by MOHAMMAD/TURNEY (2013)
- In this regard we define two hypotheses...

H1 The classification accuracy and economic relevance of **emotion-based** dictionaries are higher than the accuracy of positive-negative based dictionaries in text with an economic background.

H2 The classification accuracy and economic relevance of **economic-related** dictionaries are higher than the accuracy of non economic related dictionaries in text with an economic background.

...and check them by using following dictionaries:

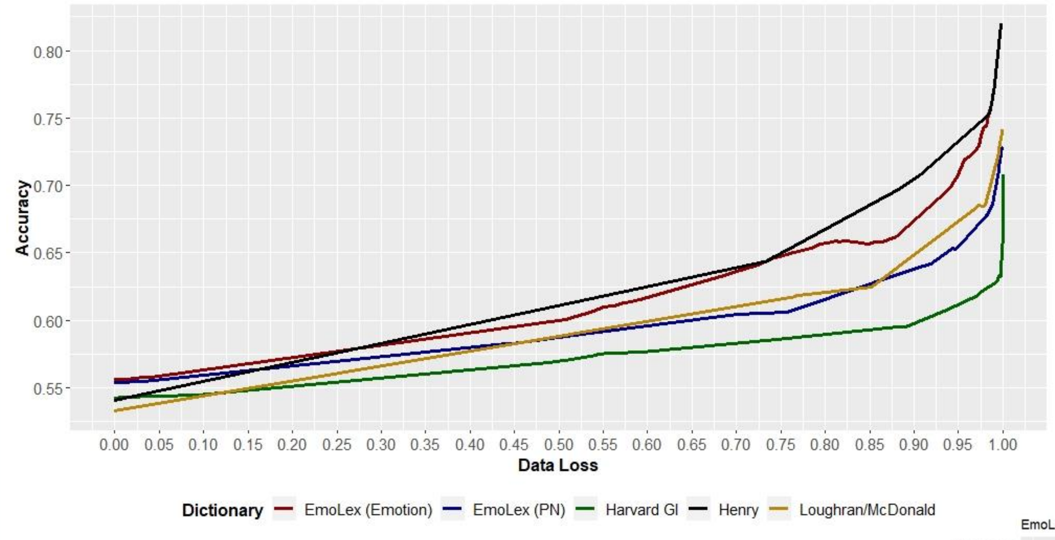
Name	Symbol	Emotions	Pos.-Neg.	Economic?
EmoLex	EM_{EL}/PN_{EL}	X	X	
Harvard GI	PN_{GI}		X	
Loughran/McDonald	PN_{LM}		X	X
Henry	PN_{HE}		X	X

lemmatization → EmoLex →

Origin	Edited	Emotion Scores	
Finally $TZA I am in green. I am off to enjoy my weekend. Signing off early. Lot of stress and anxiety. Need a break. Good luck to all.	Finally I green I enjoy weekend sign early Lot stress anxiety Need break Good luck	1	Anger
		5	Anticipation
		1	Disgust
		2	Fear
		5	Surprise
		1	Sadness
		4	Joy
		4	Trust

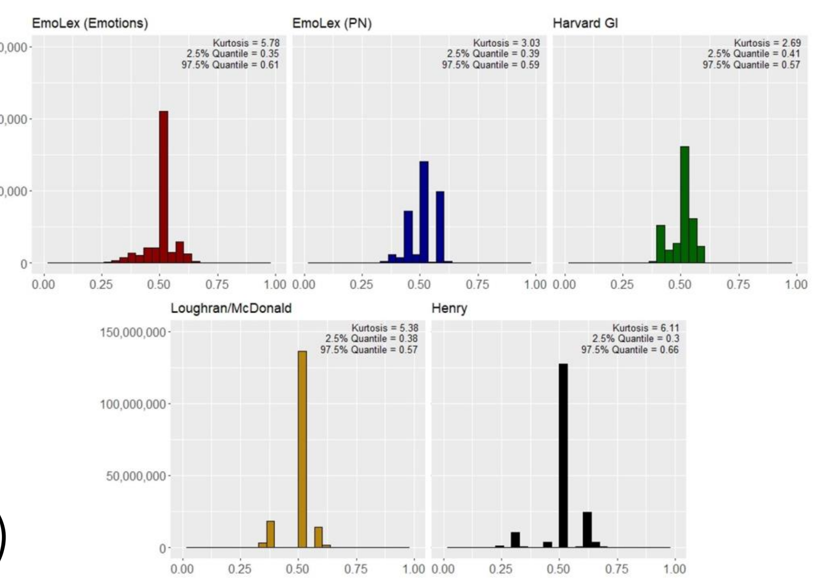
III. CLASSIFICATION

- We use the scores from all dictionaries now to **tag the 62% untagged ideas** as ‚bullish' or ‚bearish'
- So we train a **ML algorithm (sigmoid function)** with the same amount of ‚bullish' (1) and ‚bearish' (0) tagged ideas and their scores
  - We assume: predicted values > 0.5 → ‚bullish'
    predicted values ≤ 0.5 → ‚bearish'
- As many ideas can't be classified by the dictionaries we also exclude unsafe predictions as ‚hold' and observe the **development of classification accuracy**



- Generally all dictionaries profit from excluding unsafe predictions
- **But:** Multidimensional and economic-related dictionaries profit more

- As we want to know more about the ‚why?', we observe the statistical properties of the predicted values
- Multidimensional and economic-related dictionaries **offer more safe predicitions**
  - Economic: better language ‚fit'
  - Multidimensional: **larger variety of unique expressions** (286.043 vs. 624)



IV. MAIN RESULTS

- As economists we are naturally interested in the economic relevance of our results
- Following ANTWEILER/FRANK (2005) we define market sentiment at day t for dictionary i as:

$$Sentiment_{i,t} = \frac{N_{Bu,i,t} - N_{Be,i,t}}{N_{Bu,i,t} + N_{Be,i,t}}.$$

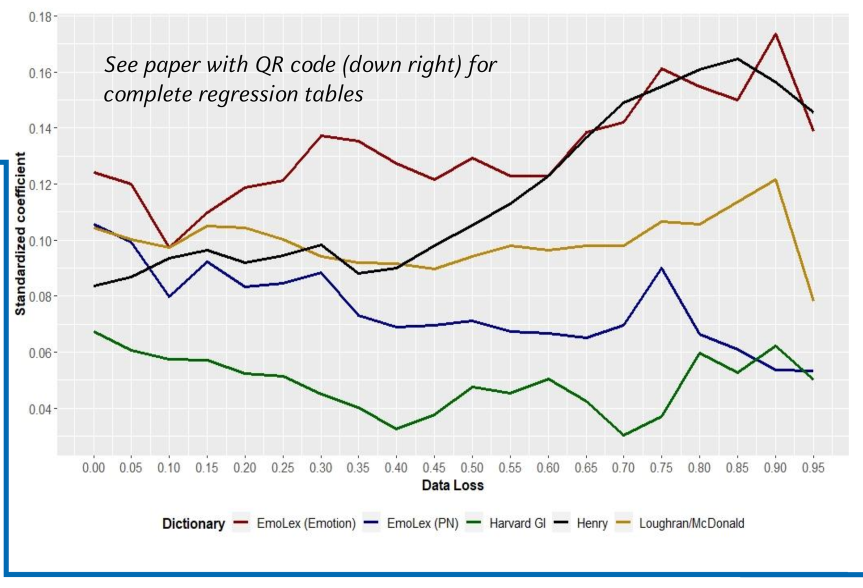
- As users mostly talk about the US stock market and are technological affine we try to **forecast NASDAQ intraday returns**
- As returns state the shift of market prices, we also need to observe the **shift in market sentiment** at market closing the day before ($t-1$) and market opening the day observed (t):

$$Intraday_t = \beta_0 + \beta_1 * Intraday_{t-1} + \beta_i * \Delta Sentiment_{i,t} + \varepsilon_t$$

$$\text{with} \quad \Delta Sentiment_{i,t} = Sentiment_{i,t,1} - Sentiment_{i,t,2}$$

$$\text{and} \quad Sentiment_{i,t,m} = \begin{cases} m = 1 \; for \; 09.30 \; am \; to \; 10.30 \; am \\ m = 2 \; for \; 03.00 \; pm \; to \; 04.00 \; pm \end{cases}$$

Results for β_i for different stages of excluded data



- Multidimensional and economic-related dictionaries **profit strongly from using safe predictions**
- Others lose their forecasting power
- Results differ between trader groups and their language

→ **Emering task:** create *multidimensional* field-specific dictionaries for economic analysis

Chair of Financial Services @ HeiCAD Lightning Talks 2022
Universitätsstraße 1, D-40225 Düsseldorf

Philipp.stangor@hhu.de | +49 211 81-10264
Lars.kuerzinger@hhu.de | +49 211 81-10261

Scan us for more about us & our research

Scan me for the paper