Predicting equity markets with social media and online news: using sentiment-driven Markov switching models.

Steven Nooijen
University of Amsterdam

In close cooperation with:
The Faculty of Economics and Business
ING Investment Management

December 2013
Abstract

This paper examines the predictive capabilities of online investor sentiment on both returns and volatility of various equity markets. For this purpose, exogenous variables are added to the mean and EGARCH volatility specifications of a Markov Switching model. We find that the Thomson Reuters MarketPsych Indices (TRMI) derived from equity specific digital news are overall better indicators of future market returns and volatility than similar sentiment from social media. In the two regime context, there is only weak evidence supporting the hypothesis of emotions playing a more important role during stressed markets compared to calm periods. However, we do find differences in sentiment sensitivity between different industries. The TRMI were the most predictive for Financials, whereas the Energy and Information Technology sectors were hardly affected by sentiment. Industry-wide we find that volatility is better predictable than returns. This is confirmed by out-of-sample Value at Risk (VaR) statistics that improve when adding the TRMI as regressors.

**Key words:** investor sentiment, social media, digital news, Markov Switching models, EGARCH, Maximum Likelihood, volatility modeling.

**Special thanks:** for making this thesis possible, I would like to thank:

Valentijn van Nieuwenhuijzen  
Principal  
ING Investment Management

dr. Simon A. Broda  
First supervisor  
University of Amsterdam

prof.dr. Peter H. Boswijk  
Second supervisor  
University of Amsterdam
Contents

Abstract i

1 Introduction 1

2 Literature review 3
   2.1 Behavioral finance: why sentiment matters ......................... 3
   2.2 Previous works ................................................. 4
   2.3 Internet based sentiment ....................................... 6
   2.4 Prior beliefs .................................................... 7

3 The model 10
   3.1 Single regime ARX-EGARCHX .................................... 10
   3.2 Two regime Markov switching .................................... 12
   3.3 Maximum Likelihood estimation ................................. 15

4 Data collection and selection 17
   4.1 Four types of web content ..................................... 17
   4.2 What information to extract ................................... 19
   4.3 The Thomson Reuters MarketPsych Indices (TRMI) ............... 19
   4.4 MarketPsych linguistics and scoring ........................... 21

5 Data description 23
   5.1 Market data ...................................................... 23
   5.2 Sentiment variables ............................................. 26
      5.2.1 Structural breaks ........................................... 27
      5.2.2 Daily seasonality .......................................... 29
      5.2.3 Unavailable data dummies .................................. 30
      5.2.4 Buzz weighting and outliers ............................... 30
      5.2.5 Weekend dummies .......................................... 31
      5.2.6 Final remarks ............................................... 32
CONTENTS

6 Results 33
  6.1 Preliminary analysis .............................................. 33
  6.2 News: a dual regime constant variance model ...................... 35
  6.3 News: the regime-switching EGARCHX model .................... 37
  6.4 Social media: model results ..................................... 42

7 Discussion 46

8 Conclusion 48

Bibliography 50

Appendix 53
Chapter 1

Introduction

Well-known scientific assumptions on efficient markets and rational agents have been the base of economic models for years. However, an increasing domain of science nowadays focuses on market irrationality, behavioral biases, and crowd psychology as explanations for what Keynes called ‘animal spirits’. His understanding that markets are largely driven by spontaneous human behavior rather than mathematical expectations is affirmed by emotions of fear and trust underlying the credit crisis. This thesis attempts to capture such emotions that influence investor decisions by extracting them from digital media sources like online newswires and social media platforms.

The existence of a relationship between digital media and financial markets has proven itself in a series of events. In April 2013, a false Twitter message caused a flash crash of one percent on the Dow Jones Industrial Average. Shortly thereafter, a live Twitter feed was incorporated in the Bloomberg accounts of institutional traders. Furthermore, several start-ups are founded to capitalize on the possibilities of social media data, one of them being a hedge fund. Another example is the stock trading platform eToro that enables the users to share and discuss their trades publicly. From an academic standpoint there is quite a substantial literature confirming the significance of social media sentiment on financial markets too. An overview is provided in Chapter 2.

This paper distinguishes itself from that existing literature on several fronts. First, the Thomson Reuters MarketPsych Indices (TRMI) we use as our data, are superior both coverage- and time-wise to data seen in other academic work. Coverage-wise, as the TRMI scan 50,000 professional news and 2 million social media sites for content every day. Plus for each of these two types, they monitor 24 different emotions rather than just bipolar positive or negative sentiment. Time-wise, as the history includes 15 years of professional news sentiment. Social media content has been collected before the launch of Twitter in 2006, but we decide only to use its sentiment from that point in time onward.

A second notable contribution to the sentiment literature is that we examine its forecast abil-
CHAPTER 1. INTRODUCTION

ity on both returns and volatility in a two regime context. For this purpose, sentiment variables are added to the mean and EGARCH volatility specifications of a general Markov switching model as described in Hamilton (1994). This allows us to observe whether sentiment is more predictive in either calm or stressed markets. Eventually, we compare the performance of this all-embracing MS-EGARCHX model with simpler models through likelihood ratio tests, information criteria, and out-of-sample test statistics. The results will indicate whether sentiment helps in predicting financial markets.

Assuming efficient markets, our main hypotheses are that sentiment does not help predict returns nor volatility. However, in case of significance, we expect to find relations mostly for volatility forecasting. A simple reasoning is that events trigger news, more news triggers more trades, and more volume could eventually lead to higher market volatility. Furthermore, we test the hypothesis that sentiment plays a larger role during stressed high volatility markets compared to tranquil times with low price swings. We suspect this as emotions tend to play a larger role during crisis periods. The sample includes data on the internet bubble, and the Lehman and Euro crises that can be used to this end.

We investigate the validity of the above hypotheses for ten different equity sectors. Earlier work has mostly focused on forecasting an industry-wide index like the DJIA or S&P500, but that does not take into account that one sector might be driven more by sentiment than another. The TRMI employ entity reference algorithms to categorize web content according to the industry it concerns. Our emotion variables are thus sector specific. As retail investors typically do not own stocks they do not read about, this could help us identify those industries that are less sensitive to for example herd behavior.

As a last point I would like to emphasize that the scope of this thesis is mainly about testing a specific model’s fit and finding out its limitations rather than achieving the best investment strategy on social media sentiment data. There are probably more practical models in existence that produce better hit ratios and Value at Risk backtest statistics than the academic MS-EGARCHX model of this research. Having said that, we begin our work with a literature review in the next chapter. Thereafter, the model is discussed in Chapter 3. Different types of internet content and the TRMI specifically follow in Chapter 4. Then we analyze the data more deeply including graphics and descriptive statistics, so that we gradually build towards the results in Chapter 6.
Chapter 2

Literature review

This chapter provides a theoretical background on the relationship between sentiment and financial markets. We start with a short introduction to behavioral finance and see how sentiment can be used as a means or attempt to quantify investor behavior. Several of these attempts are discussed subsequently. In a third chapter, the pros of minutely derived sentiment from the internet are compared to the cons of more traditional sentiment surveys. Throughout the chapter, we will refer to earlier results, to finish off with a table containing prior beliefs on the to-be-estimated signs of the sentiment variables under review in this thesis.

2.1 Behavioral finance: why sentiment matters

In the established efficient market hypothesis (EMH), classic theory assumes that arbitrage plays a critical role in driving prices back to their fundamental values. Early work by De Long et al. (1990) already questions this assumption by dividing investors into two different groups: rational arbitrageurs and irrational noise traders. Arbitrageurs bet against the beliefs of noise traders, but their unpredictable and long-lasting beliefs can cause prices to significantly diverge from fundamental values even in the absence of fundamental risk. The authors call this ‘noise trader risk’ and demonstrate the risk with the following example. An arbitrageur buys an asset whose price has been driven down significantly by pessimistic noise traders’ opinions. If the arbitrageur does not recognize that the price could be driven down even more in the near future due to increasingly pessimistic beliefs, he might have to liquidate and suffer a loss before the price recovers. Shleifer and Vishny (1997) show that this is particularly likely to happen in extreme mispricing circumstances when arbitrageurs are fully invested. This crucial result is aptly labeled ‘the limits of arbitrage’.

The essential assumption in De Long et al. (1990) is that the opinions of noise traders are unpredictable. Being able to capture this behavior of irrational investors would therefore allow an arbitrageur to successfully bet against it. Shleifer and Vishny (1997) confirm this thought and
stress that it is essential to understand the source of noise trading that caused the mispricing in the first place. Signals examined for this purpose include among others volume, price patterns, institutional restrictions, expert opinions, and sentiment. This last possible driver of noise trading, sentiment, is what this paper will focus on.

Over time, the importance of capturing sentiment, emotions, or any other form of irrational behavior has increasingly been recognized among economists. Even the most experienced professional traders are now shown to demonstrate different physiological reactions and thus different emotions in periods of heightened market volatility (Lo and Repin, 2002). This, together with behavioral biases proven by Nobel prize winners Daniel Kahneman and Robert Shiller, challenge the rationality assumptions underlying important finance models like the capital asset pricing model (CAPM) or the EMH. Loss aversion, anchoring, overconfidence, and mental accounting are just some of the examples that question the validity of these models (Kahneman, 2011). In an attempt to model this irrational inconsistent human judgment, economists today join forces with psychologists, sociologists, and neuroscientists. Their relatively new approach to economic theory is also called behavioral finance.

### 2.2 Previous works

Having acknowledged that emotional factors play an important role in financial decision making, the best way to measure these mood swings of investors remains debatable. In this section, we discuss papers that derive sentiment from macro-economic figures, traditional consumer confidence surveys, news, internet weblogs, search queries, and social media. Section 2.3 then argues why internet sentiment specifically is of our main interest. Although some of these works also examine commodities and Forex markets, we focus on summarizing equity related findings here. Table 2.1 provides an overview of the works discussed.

We start with Baker et al. (2012), who construct investor sentiment indices using a variety of macro series like volatility premiums, IPO volumes, and turnover data. They find that sentiment is a contrarian predictor of returns. When global and local sentiment are high, future local stock returns are low. Their approach is followed by Finter et al. (2012) for the German stock market. Their similarly constructed sentiment index on the contrary has only weak power in predicting future stock returns. This could suggest that information derived from macro-economic variables is priced in or does not capture the irrationality or emotions we are after. The latter is confirmed by Da et al. (2011) who argue that, for example, turnover data can also be driven by factors unrelated to investor attention.

Traditional investor sentiment survey data is compared with a variety of online sentiment sources in Mao et al. (2011). The Investor Intelligence survey and Daily Sentiment Index are respectively in existence since 1964 and 1987, but are proven to lag financial markets in their
research. This is in sharp contrast to information from Twitter and Google search queries which in the same paper do possess predictive power on the DJIA and market volatility VIX when forecasting daily. To a lesser extent, Mao et al. (2011) lastly find that stock prices also react to news headlines from major media outlets like Bloomberg, CNBC, and BusinessWeek. The bodies of texts from ‘regular’ media are analyzed in Tetlock (2007). He scans 16 years of Wall Street Journal columns and finds that high media pessimism leads lower stock market returns.

A more personal kind of column is analysed by Gilbert and Karahalios (2010): weblogs. They construct an Anxiety index based on metrics of anxiety, worry, and fear derived from LiveJournal weblog posts. A negative coefficient is found when predicting S&P stock returns, meaning that anxiety slows a market climb and accelerates a drop. Weblogs together with other online platforms such as communities, blogs, product reviews, and wikis are together termed user-generated content (UGC) by Tirunillai and Tellis (2012). In their paper they investigate how product reviews and product ratings from Amazon.com, Yahoo! shopping, and Ebay affect stock market performance across 6 markets over a 4-year period. A multivariate time series model on daily data reveals that volumes of chatter significantly lead abnormal returns. A second finding is that negative sentiment has a larger impact than positive chatter, i.e. the relation is asymmetric.

Volumes of chatter are remarked as a proxy for investor attention by, among others, Rao and Srivastava (2013). Another measure of attention could be Google search volumes. The authors combine and compare both in an effort to model equity, commodity, and Forex markets. A Granger causality analysis rejects the null hypothesis of sentiment measures not affecting financial market returns. Second, their model has an accuracy of over 90% in forecasting the direction of weekly movements of the DJIA and NASDAQ-100 during a 16 week testing period.

The relation between Google search volumes and financial markets however is not that clear-cut. Da et al. (2011) find that an increase in search volumes predicts higher stock prices in the next 2 weeks for Russell 3000 stocks. Contrary to this, Preis et al. (2013) detect an increase in Google search volumes before a stock market drop. The latter is explained by search queries reflecting the information gathering process of concerned investors preceding a sell off. The point that both agree on is that volumes are a way to reveal and quantify the interests of investors.

The effect of more specific emotions than general bipolar sentiment is the object of study in a paper of Bollen et al. (2011). Text content of daily Twitter feeds are scored along 6 different emotions: calm, alert, sure, vital, kind, and happy. Some specific mood dimensions are found to contribute to DJIA forecasting, but not others. Sign prediction on daily up and down changes is significantly improved when the emotion calm is included, up to an accuracy of 87.6%.

There are also examples in which sentiment does not lead financial returns. Antweiler and Frank (2004) scan messages in internet stock chat rooms for ‘buy’, ‘hold’ and ‘sell’ recommendations and find that message activity does not predict returns, but rather return volatility.
CHAPTER 2. LITERATURE REVIEW

<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Source</th>
<th>Return pred.</th>
<th>Volatility pred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker et al. (2012)</td>
<td>Offline</td>
<td>Macro data</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>Finter et al. (2012)</td>
<td>Offline</td>
<td>Macro data</td>
<td>No</td>
<td>–</td>
</tr>
<tr>
<td>Mao et al. (2011)</td>
<td>Both</td>
<td>Twitter, News, Google search</td>
<td>Mixed</td>
<td>VIX</td>
</tr>
<tr>
<td>Gilbert and Karahalios (2010)</td>
<td>Online</td>
<td>LiveJournal weblogs</td>
<td>‘Anxiety’</td>
<td>VIX*</td>
</tr>
<tr>
<td>Tirunillai and Tellis (2012)</td>
<td>Online</td>
<td>Amazon, Yahoo!, Ebay</td>
<td>Yes</td>
<td>Trade volumes</td>
</tr>
<tr>
<td>Rao and Srivastava (2013)</td>
<td>Online</td>
<td>Twitter, Google search</td>
<td>Yes</td>
<td>VIX</td>
</tr>
<tr>
<td>Da et al. (2011)</td>
<td>Online</td>
<td>Google search</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>Preis et al. (2013)</td>
<td>Online</td>
<td>Google search</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>Bollen et al. (2011)</td>
<td>Online</td>
<td>Twitter</td>
<td>‘Calmness’</td>
<td>–</td>
</tr>
<tr>
<td>Karabulut (2013)</td>
<td>Online</td>
<td>Facebook</td>
<td>Yes</td>
<td>Trade volumes</td>
</tr>
<tr>
<td>Antweiler and Frank (2004)</td>
<td>Online</td>
<td>Yahoo! Finance, Raging Bull</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Gilbert and Karahalios (2010) find that their Anxiety index is an alternative to the VIX ‘fear gauge’.

Table 2.1: Existing literature on sentiment and stock market activity. Most studies that use online data do find a predictive relation for either bipolar sentiment or message volumes (buzz, Table 2.2). However, some prove the significance of more specific emotions like ‘pessimism’. Concerning volatility forecasts, many find a predictive connection with trade volumes or the implied S&P 500 Volatility Index (VIX).

Robustness of this result is ensured by using several volatility models, including a GARCH specification. Of the works discussed so far, three confirm sentiment’s leading relation to volatility by investigating the implied Volatility Index VIX (Table 2.1). Another three found a significant correlation between message posting and trading volumes. Transaction volumes are proven to positively correlate with volatility (Jones et al., 1994). Therefore, we would like to conclude this section with the notion that digital media seem to help predict both returns and volatility of equity markets. However, as the previous literature contains some conflicting claims on the direction of the relationships, we like to form prior beliefs on our variables in Section 2.4.

2.3 Internet based sentiment

The works discussed in the previous paragraphs show that investor sentiment is something to take into account when modeling financial markets; whatever its source may be. In this paragraph we make our case for using internet derived sentiment instead of traditional offline sentiment indicators or surveys.

First, traditional surveys are cost intensive and time consuming (Bollen et al., 2011). Mood analysis from large-scale online data on the other hand is more rapid and cost effective due to automated language processing (Mao et al., 2011). Furthermore, one could question the
truthfulness or validity of respondents’ self-reported emotional states in surveys (Brener et al., 2003). Sophisticated language tool kits allow for the extraction of a variety of emotions from an author’s text, while not explicitly asking the writer about his or her emotion.

Second, the World Wide Web contains massive amounts of inter-consumer communication. Word of mouth advertisement is now more important than ever, with user product reviews instantly and freely available to everyone with internet access (Tirunillai and Tellis, 2012). The same counts for stock forums or chat rooms where retail investors inform themselves before trading. Decisions to buy or sell a certain stock are equally based on digital media as decisions to buy a consumer good. The massive volumes and easy accessibility of this kind of information are unmatched by the sources underlying offline sentiment indices or surveys.

Third, internet sentiment is always up-to-date. Markets still react to monthly consumer confidence indices, but sentiment from the internet could theoretically be derived every second. With 800 million Twitter users generating over 250 million Tweets every day, their aggregate mood could be captured in a real-time public sentiment index (Bollen et al., 2011). Similarly, Karabulut (2013) show how an increase in the real-time Facebook Gross National Happiness index predicts an 11 basis point increase for the next day’s return. Social media thus allow us to track the mood of millions in a more timely fashion.

Lastly, with some examples already provided in the previous section, a growing amount of literature demonstrates that these computational web-based sentiment indicators are genuine predictors of financial market movements. News events themselves may be unpredictable, but social media often provide a first indication of what is about to happen. Social media’s forecast capabilities have been demonstrated on a variety of socio-economic phenomena, including presidential elections (Tumasjan et al., 2010) and influenza epidemics (Culotta, 2010).

2.4 Prior beliefs

So far, this chapter has explained why behavioral aspects should be included in financial models, how sentiment could fulfill that role, and why we prefer to use sentiment derived from the internet specifically. Whereas the process of extracting sentiment from online media is described in Chapter 4, the last section of this chapter focuses on forming prior beliefs on our sentiment covariates. We keep the academic works discussed so far in mind and combine them with ING Investment Management (IM) best practice insights.

It is important to form a prior belief on the sign of each coefficient to be estimated in order to avoid data mining. For example, based on Lerner et al. (2004), we expect that emotions like gloom, fear, and anger are negatively correlated with equity returns. The heavier these emotions become, the likelier people are to sell their assets. This is confirmed by the negative slope found for the Anxiety index constructed by Gilbert and Karahalios (2010). If instead, our
estimated coefficient turns out to be positive, it would not match our expectation based on the above literature. Trying to explain the counter intuitive sign would then introduce hindsight bias.

An overview of all variables investigated and their expected relationships with both returns and volatility is shown in Table 2.2. The table also contains a brief description of what each variable measures. As mentioned before, the variables’ assembly is comprehensively discussed in Chapter 4. For the moment, it should be understood that each sentiment series has a separate reading per sector. For example, optimism could be higher in the Consumer Staples sector than in the Utilities sector at any given time. For most variables our beliefs are consistent across all equity sectors, except for violence and conflict. Instead of having a negative impact on Energy sector returns, these variables are thought of as reflecting turmoil or unrest, which could drive energy prices upwards. Higher energy prices are good for energy companies, leading to higher returns. We look at the different sectors in more detail in Section 5.1.

The prior beliefs reported in Table 2.2 are based on academic work presented throughout this chapter and are supplemented by the strategic insights of ING IM’s Global Strategy Team. This team develops, formulates and communicates the macro and asset class (fixed income, equities, real estate and commodities) views of ING IM. Next to these top-down views on macroeconomics and asset classes, they also develop the sectoral or intra-asset class calls like equity sectors and regional attractiveness. We consult their extensive investor experience to overrule any contradictory theoretical findings with practical insights where necessary.
# CHAPTER 2. LITERATURE REVIEW

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
<th>Type</th>
<th>Prior return</th>
<th>Prior volatility</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUZZ</td>
<td>Sum of entity-specific words and phrases used in computations</td>
<td>Buzz metric</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SNTMENT</td>
<td>Overall positive references, net of negative references</td>
<td>Emotion</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>OPTIMSM</td>
<td>Optimism, net of references to pessimism</td>
<td>Emotion</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>FEAR</td>
<td>Fear and anxiety</td>
<td>Emotion</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>JOY</td>
<td>Happiness and affection</td>
<td>Emotion</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>TRUST</td>
<td>Trustworthiness, net of references connoting corruption</td>
<td>Emotion</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>VIOLENC</td>
<td>Violence and war</td>
<td>Emotion</td>
<td>−*</td>
<td>+</td>
</tr>
<tr>
<td>CONFLICT</td>
<td>Disagreement and swearing, net of agreement and conciliation</td>
<td>Emotion</td>
<td>−*</td>
<td>+</td>
</tr>
<tr>
<td>URGENCY</td>
<td>Urgency and timeliness, net of references to tardiness and delays</td>
<td>Emotion</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>UNCERTN</td>
<td>Uncertainty and confusion</td>
<td>Emotion</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>PRICEUP</td>
<td>Price increases, net of references to price decreases</td>
<td>Buzz metric</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>MKTCST</td>
<td>Predictions of asset price rises, net of references to predictions of asset price drops</td>
<td>Buzz metric</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>MKTRISK</td>
<td>Positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles - higher values indicate greater bubble risk. Also known as the ‘Bubbleometer’</td>
<td>Buzz metric</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GLOOM</td>
<td>Gloom and negative future outlook</td>
<td>Emotion</td>
<td>−</td>
<td>0</td>
</tr>
<tr>
<td>ANGER</td>
<td>Anger and disgust</td>
<td>Emotion</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>INNOVAT</td>
<td>Innovativeness</td>
<td>Buzz metric</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>STRESS</td>
<td>Distress and danger</td>
<td>Emotion</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>FUNDSTR</td>
<td>Positivity about accounting fundamentals, net of references to negativity about accounting fundamentals</td>
<td>Buzz metric</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>EARNEXP</td>
<td>Expectations about improving earnings, less those of worsening earnings</td>
<td>Buzz metric</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>MRGBUZZ</td>
<td>Merger or acquisition activity</td>
<td>Buzz metric</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LAYOFFS</td>
<td>Staff reductions and layoffs</td>
<td>Buzz metric</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>LITIG</td>
<td>Litigation and legal activity</td>
<td>Buzz metric</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>UPGDWNGD</td>
<td>Upgrade activity, net of references to downgrade activity</td>
<td>Buzz metric</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>VOLATIL</td>
<td>Volatility in market prices or business conditions net of stability</td>
<td>Buzz metric</td>
<td>−</td>
<td>+</td>
</tr>
</tbody>
</table>

*We expect the effect of violence and conflict on Energy sector returns to be positive (+).

Table 2.2: An overview of the Thomson Reuters MarketPsych Indices (TRMI) specific to equity markets. It includes descriptions, types (Section 4.3), and our prior beliefs concerning the relationship with returns and volatility. A 0 means that we do not expect to find any relation.
Chapter 3

The model

The model that tests the prior beliefs of Section 2.4 is described in this chapter. In order to determine the influence of sentiment variables on both mean returns and variances, we extend the standard econometric AR-EGARCH model with sentiment variables in both the mean and volatility specifications, naming the new model an ARX-EGARCHX. To investigate whether sentiment plays a more important role during troubled or tranquil markets, we then extend this model with a Markov switching part. This additional flexibility allows us to estimate high and low volatility regimes, as will be described in Section 3.2. Documentation-wise, we base ourselves on three main publications: Tsay (2005), Hamilton (1994) and Gray (1996).

3.1 Single regime ARX-EGARCHX

Let $r_t$ be the log return of an equity sector at time $t$. Basic financial time series modeling is about specifying the conditional mean and variance of a log return series $r_t$ at time $t$:

$$ r_t = \mu_t + a_t $$

$$ \mu_t = E(r_t|F_{t-1}) $$

$$ \sigma_t^2 = Var(r_t|F_{t-1}) = E[(r_t - \mu_t)^2|F_{t-1}] = Var(a_t|F_{t-1}), $$

where $F_{t-1}$ denotes the information set available at time $t - 1$. For the moment, we focus on what we call the mean equation $\mu_t$ for the return series $\{r_t\}$. However, we quickly switch to discussing conditional heteroscedastic models as a tool for modeling the volatility $\sigma_t^2$ of $r_t$. The mean equation residual $a_t = r_t - \mu_t$ is referred to as the innovation or shock to an asset return at time $t$. Their squares $\{a_t^2\}$ are tested for serial correlation as an indicator of conditional heteroscedasticity.

The mean equation removes any linear dependence from the asset return series $\{r_t\}$. Serial correlation is accounted for by a simple stationary ARMA$(p,q)$ time series model, with the order
(p, q) depending on ACF and PACF plots of returns and squared returns. Tsay (2005) notes however that for most return series the serial correlations are weak, if any. For daily series, the more basic AR(p) specification might already be sufficient. The ARMA(p, q) model can easily be extended with exogenous (sentiment) drivers \( x_{i,t-1} \) so that

\[
\mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + \sum_{i=1}^{k} \delta_i x_{i,t-1} \left( -\sum_{i=1}^{q} \theta_i a_{i,t-1} \right),
\]

(3.1)

where \( p, k \) and \( q \) are non-negative integers. As we deal with daily observations, we follow Tsay (2005) and exclude the MA(q) terms which are now displayed between parentheses. Leaving them out furthermore improves estimation time. Because of the exogenous regressors in the mean equation, we will refer to this particular specification as the ARX(p) model from here on. There are no restrictions on the coefficients in Equation (3.1); multiple exogenous variables can be added.

Like the well-known ARMA model for the mean equation, there is also a very widely practiced time series model for the volatility equation. One of the earliest is the GARCH(m, s) specification proposed by Bollerslev (1986). It models the evolution of \( \sigma_t^2 \), which influences the conditional variance of \( r_t \) through

\[
\begin{align*}
\sigma_t^2 &= \omega + \sum_{i=1}^{m} \alpha_i a_{i,t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \\
&= \omega + \sum_{i=1}^{m} \alpha_i a_{i,t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2,
\end{align*}
\]

(3.2)

in which \( \epsilon_t \) is an independent and identically distributed (iid) random variable with mean 0 and variance 1. Although different distributions can be chosen for \( \{\epsilon_t\} \) to allow for fat tails (Student-t, GED), we stick with the normal distribution as it greatly simplifies the likelihood function of our more complicated Markov switching model discussed in the next Section 3.2. The resulting Quasi-Maximum Likelihood estimates remain valid even if the true distribution is non-Gaussian.

The coefficients of the GARCH(m, s) model in 3.2 are subject to some constraints. A positive volatility is ensured by \( \omega > 0, \alpha_i \geq 0 \) and \( \beta_j \geq 0 \), whereas \( \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_j) < 1 \) implies a finite unconditional variance of \( a_t \). With the constraints satisfied, the GARCH model basically allows the conditional variance \( \sigma_t^2 \) to be dependent on past squared shocks and its own recent history. This means that a large shock is now likely to be followed by another large shock, and high volatility at time \( t \) continues to persist in the near future through its lagged values.

To investigate the effect of sentiment variables on volatility, we extend the basic GARCH(1,1) by including exogenous volatility regressors \( v_{i,t-1} \):

\[
\sigma_t^2 = \omega + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \sum_{i=1}^{k} \theta_i v_{i,t-1}.
\]

(3.3)
However, the standard GARCH(1, 1) conditions $\omega > 0, \alpha_1 \geq 0, \beta_1 \geq 0$ and $(\alpha_1 + \beta_1) < 1$ no longer guarantee positivity as no restrictions are placed on the domain nor estimated coefficients of $v_{i,t-1}$. Although it does not seem to bother the results of Chebbi et al. (2013), Han and Kristensen (2012) solve this problem by squaring their exogenous regressors and adding a non-negativity constraint for the estimable parameters. The drawback of this approach is that negative values influence volatility in the same way as their positive equivalents; i.e. any asymmetry is removed.

An easy way of overcoming this weakness is by using the exponential GARCH (EGARCH) model proposed by Nelson (1991). The model has two major advantages. One, by modeling the logged conditional variance, the positivity constrains on the coefficients are removed. We therefore no longer need to take squares or the absolute value of our external sentiment regressors. Second, the EGARCH specification enables the model to react differently to positive and negative past shocks of $a_t$, capturing the so called leverage effect. A large negative shock is expected to increase volatility more than a positive shock similar in absolute magnitude.

Although the model allows for different orders of the ARCH and GARCH parameters, we stick to the EGARCH(1, 1) model for ease of estimation. We do add some sentiment variables $v_{i,t-1}$ which we expect to influence volatility. The single regime heteroscedasticity model becomes

$$
\log_e \sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1} + \gamma_1 (|\epsilon_{t-1}| - E[|\epsilon_{t-1}|]) + \beta_1 \log_e \sigma_{t-1}^2 + k \sum_{i=1}^{k} \theta_i v_{i,t-1}
$$

where $\epsilon_t = a_t/\sigma_t$ is the standardized innovation. If $\epsilon_t \sim$ i.i.d. $N(0, 1)$, then $E(|\epsilon_t|) = \sqrt{2/\pi}$. The series is stationary if $|\beta_1| < 1$ because the specification is basically an AR(1) process for $\log_e \sigma_t^2$. Lastly, the second and third term together form the weighted innovation $g(\epsilon_t)$ which captures the leverage effect if $a_{t-1} < 0$:

$$
g(\epsilon_t) = \alpha_1 \epsilon_{t-1} + \gamma_1 (|\epsilon_{t-1}| - E[|\epsilon_{t-1}|]).
$$

Taking the exponential like in Equation (3.4) thus provides a relatively easy way of incorporating external regressors in the volatility equation, while keeping the volatility positive and allowing for the estimation of important characteristics like the leverage effect. Liu et al. (2012) use the same EGARCHX(1, 1) model on equity return volatility and find significance for their external covariate trading volume. The model is estimated using Maximum Likelihood. A derivation of the log-likelihood function is given in Section 3.3.

### 3.2 Two regime Markov switching

The joint ARX-EGARCHX model as described in the previous section is a well accepted tool for capturing three important characteristics of volatility data: its rather continuous evolution
over time, the leverage effect and its persistence. Some argue however that the GARCH terms installed for capturing this persistence are too inflexible. Structural relationships between variables are assumed to be constant over the entire sample, while in fact relations might be quite different under certain circumstances. Gray (1996) for example observes that the short-term interest rate behaves differently in high and low volatility environments. He develops the generalized regime-switching (GRS) model that enables the parameters of the conditional mean and variance processes to take two different values, depending on a latent regime indicator \( S_t \). In this section we introduce regime switching to our model too, as equity markets clearly display periods of tranquility alternating with increased market turmoil (internet bubble, Lehman, Euro crisis). For the construction of our model, we largely base ourselves on Gray’s GRS model and the chapter on regime switching models in Hamilton (1994).

To understand why a single regime model could lead to misspecification, think of a particular GARCH parameter taking a high value in one regime and a low value in the other. The parameter estimate of a single regime model would then average over the two regimes, producing a false indication of the actual relationship. This could in turn lead to a wrong economic interpretation. By making the parameters of Equation (3.1) and (3.4) regime-dependent, we can learn whether sentiment plays a more important role in stressed markets compared to calm periods. Besides, we can investigate which regime is more persistent to shocks, which regime is most prone to the leverage effect, and so forth.

Suppose there are two unobserved regimes \( S_t \) at time \( t \), \( S_t = \{1, 2\} \). Each regime has its own conditional mean \( \mu_{it} \), variance \( \sigma_{it}^2 \), and assumed normal distribution. The state means and variances have the form of respectively Equation (3.1) and (3.4), but subscripts are added to each coefficient to allow for different regime behavior. The distribution for a return series \( r_t \) then combines both regimes in

\[
\begin{align*}
r_t | F_{t-1} \sim & \begin{cases} 
N(\mu_{1t}, \sigma_{1t}^2) & \text{w.p. } \pi_{1t} = P(S_t = 1|F_{t-1}), \\
N(\mu_{2t}, \sigma_{2t}^2) & \text{w.p. } \pi_{2t} = 1 - \pi_{1t},
\end{cases} 
\end{align*}
\tag{3.5}
\]

with \( \pi_{1t} \) as the mixing parameter or probability of being in State 1 at time \( t \). The regimes are never actually observed, but conditional on the information set we can distinguish three different probabilities that indicate the likelihood of a regime occurring/having occurred. First, the ex ante probability \( P(S_t = 1|F_{t-1}) \) only depends on the information of time \( t - 1 \) and is used for forecasting. Using the realized return at time \( t \) we can calculate the filtered regime probability \( P(S_t = 1|F_t) \). This is used for calculation of the new ex ante probabilities in the next period \( t + 1 \). The third probability, the smoothed \( P(S_t = 1|F_T) \), is calculated recursively afterwards and uses the entire data set. It shows which regime was dominant at each point in time.

To calculate the ex ante state probabilities we need to know the transition or switching
probabilities of moving from one regime to the other,
\[
P(S_t = i | S_{t-1} = j) = \begin{pmatrix} P & 1 - Q \\ 1 - P & Q \end{pmatrix}. \tag{3.6}
\]
For example, \(1 - Q\) is the chance of moving from regime 2 to regime 1. The transition probabilities are assumed constant over time, but can be made time varying and external variable dependent through a normal cumulative distribution function like in Gray (1996) or Ozoguz (2009), or through a logit function (Liu et al., 2012). Any other function that maps input to a zero-to-one range will also work for that matter. As the regime probabilities \(\pi_{it}\) are not observed at time \(t\), we use the transition matrix to predict the (ex ante) regime probabilities at time \(t\) as follows:
\[
\pi_{1t} = P \cdot P(S_{t-1} = 1 | F_{t-1}) + (1 - Q) \cdot P(S_{t-1} = 2 | F_{t-1}). \tag{3.7}
\]
This is a Hamilton first-order Markov model, which assumes that all information up to \(F_{t-1}\) is encapsulated in the last state \(S_{t-1}\). The ex ante probabilities are used for calculating the log-likelihood function, as will be explained in the next subsection. The last part of this paragraph focuses on a problem caused by the GARCH coefficient \(\beta_i\) in the volatility equation. Because it can take on two different values depending on the regime \(i\), the volatility at time \(t\) is, among other things, determined by the regime at time \(t - 1\). Clearly the volatility at \(t - 1\) in turn depends on the regime at \(t - 2\) and so on. This problem of full path dependence is graphically illustrated in Gray (1996), but it is clear that the dependence of the conditional variance on its entire history makes the estimation of Markov switching models impossible. The solution proposed by Gray (1996) is to compute, at each time step \(t\), the conditional variance summed over both states:
\[
\sigma^2_{it} = \mathbb{E}[r^2_t | F_{t-1}] - \mathbb{E}[r_t | F_{t-1}]^2 = \pi_{1t}(\mu_{1t}^2 + \sigma_{1t}^2) + \pi_{2t}(\mu_{2t}^2 + \sigma_{2t}^2) - (\pi_{1t}\mu_{1t} + \pi_{2t}\mu_{2t})^2, \tag{3.8}
\]
in which \(\pi_{it}\) represent the filtered regime probabilities \(P(S_t = i | F_t)\). The conditional regime variances \(\sigma^2_{it}\) will depend on it through
\[
\log_e \sigma^2_{it} = \omega_i + \alpha_i \epsilon_{t-1} + \gamma_i (|\epsilon_{t-1}| - \mathbb{E}(|\epsilon_{t-1}|)) + \beta_i \log_e \sigma^2_{t-1} + \sum_{j=1}^k \theta_{ij} v_{j,t-1} \tag{3.9}
\]
and will thus no longer be fully path dependent, while preserving the persistence effect of a GARCH parameter. The residuals \(a_t\) are also aggregated over two states by \(a_t = r_t - (\pi_{1t}\mu_{1t} + \pi_{2t}\mu_{2t})\), so that the standardized residuals become \(\epsilon_t = a_t / \sigma_t\). In a normal EGARCH process we require \(\beta < 1\) for stationarity, but it is not sure how the stationarity conditions are affected by Gray’s proposed variance aggregation. To see this, ignore the mean equation by setting \(\mu_{1t} = \mu_{2t} = 0\) and substitute Equation (3.8) in (3.9). Here \(\sigma^2_{it}\) depends on the stationarity of both
\( \sigma^2_{1t} \) and \( \sigma^2_{2t} \) through a combination of logarithms and regime probabilities \( \pi_{it} \). However, we are unable to take unconditional expectations and determine the exact stationarity conditions as the expectation of a logarithm does not equal the logarithm of an expectation \( \log_e(E[]) \neq \log_e(E[]) \).

Although we expect the condition \( \beta_i < 1 \) to be too strong of an assumption, we rarely observe an estimated \( \beta_i \) larger than one (Chapter 6). Just to be on the safe side, we restrict \( \beta_i < 1 \) in such cases as a higher beta does seem to cause very low transition probabilities \( P \) and \( Q \).

Note that by making all parameters in both the mean and volatility equations state dependent, we significantly increase the number of parameters to be estimated. To prevent overparameterization we therefore decide not to make the transition probabilities depend on external sentiment covariates too. As the Markov switching model in Equation (3.5) has the volatility specification of (3.9), we refer to this final model from here on as the MS-EGARCHX model.

### 3.3 Maximum Likelihood estimation

To estimate the single regime EGARCHX and dual regime MS-EGARCHX we use Maximum Likelihood. For both models we assume \( \{\epsilon_t\} \) to be i.i.d. normally distributed for ease of estimation. This means that the density function and accompanying log-likelihood are defined as

\[
f(r_t|F_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2_t}} \exp \left\{ -\frac{(r_t - \mu_t)^2}{2\sigma^2_t} \right\}
\]

\[
\mathcal{L} = \sum_{t=1}^{T} \log \left[ \frac{1}{\sqrt{2\pi\sigma^2_i}} \exp \left\{ -\frac{(r_t - \mu_i)^2}{2\sigma^2_i} \right\} \right]
\]

with \( \mu_t \) the ARX process of Equation (3.1), and \( \sigma^2_t \) the EGARCHX specification in Formula (3.4).

This is a standard likelihood function which can easily be estimated with the R \texttt{rugarch} package (Ghalanos, 2013). For Markov switching models, the density depends on the current regime. In State 1, an observed return \( r_t \) is supposed to be drawn from the \( N(\mu_{1t}, \sigma^2_{1t}) \) distribution, where in the second case this would be the \( N(\mu_{2t}, \sigma^2_{2t}) \) distribution. The density of \( r_t \) is thus conditional on the state or regime \( S_t \):

\[
f(r_t|S_t = i; F_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2_{it}}} \exp \left\{ -\frac{(r_t - \mu_t)^2}{2\sigma^2_{it}} \right\}.
\]

Following Hamilton (1994) we can next calculate the joint density distribution function of \( r_t \) and \( S_t \) with Bayes' theorem, so

\[
P(A \text{ and } B) = P(A|B) \cdot P(B)
\]

\[
f(r_t, S_t = i|F_{t-1}) = \frac{\pi_{it}}{\sqrt{2\pi\sigma^2_{it}}} \exp \left\{ -\frac{(r_t - \mu_t)^2}{2\sigma^2_{it}} \right\}.
\]
Here, $\pi_{it}$ is the ex ante state probability of regime $i$ as the actual state is unobserved at time $t$. The unconditional density of $r_t$ is then found by aggregating over the two states:

$$f(r_t|F_{t-1}) = \sum_{i=1}^{2} f(r_t, S_t = i|F_{t-1})$$

$$= \frac{\pi_{1t}}{\sqrt{2\pi} \sigma_{1t}} \exp \left\{ -\frac{(r_t - \mu_{1t})^2}{2\sigma_{1t}^2} \right\} + \frac{\pi_{2t}}{\sqrt{2\pi} \sigma_{2t}} \exp \left\{ -\frac{(r_t - \mu_{2t})^2}{2\sigma_{2t}^2} \right\},$$

so that the log-likelihood function becomes

$$\mathcal{L} = \sum_{t=1}^{T} \log f(r_t|F_{t-1}). \quad (3.12)$$

As no existing packages exist for maximizing the MS-EGARCHX log-likelihood in (3.12), we program the function ourselves and use the R `maxLik` package for the computation (Henningsen and Toomet, 2011). The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is used for the numerical optimization, where constraints can be set to make sure the estimated transition probabilities remain between 0 and 1. As the algorithm is not very robust to the starting values used due to a degeneracy problem described in Section 6.4, we try two sets for each sector and data source. First, generic values in which almost all parameters are assumed to have no significance and are set to zero. Exceptions are the transition probabilities $P$ and $Q$ which are both set to be quite persistent at 0.97. The GARCH parameter $\beta_i$ starts at 0.95 and $\gamma_i$ is set to 0.1. The second set of starting values is based on the estimates of the GARCH parameters in the simpler single regime EGARCHX model, combined with the estimated transition probabilities and mean parameters of a simple MS model with constant variance $\sigma_i^2$ in each state.

This latter model, the MS-constant variances, is also used as a reference or test to see whether our own code is (partly) correctly specified. We make use of an existing R package called `MSwM` (Sanchez-Espigares and Lopez-Moreno, 2013) to obtain estimates, which we compare to ours. The MS-constant variances model has the same log-likelihood as in Equation (3.12) but $\sigma_{it}^2$ is replaced by $\sigma_i^2$; the variance in each regime is no longer time varying.
Chapter 4

Data collection and selection

The enormous amount of data sources online vary from professional news articles and journalists’ blogs to stock discussion forums and social media. We categorize these sources into 4 different types based on their user base. We also evaluate what kind of information from these types could help us predict financial markets. In Section 4.3 we introduce the Thomson Reuters MarketPsych Indices (TRMI) as our data source and see how the former two points apply to it. In the final subsection the construction of the TRMI is discussed in more detail, including some remarks on the linguistic analysis behind it.

Throughout the chapter we will make references to earlier studies. We find that the TRMI are quite superior to data used in previous academic work for several reasons. One, its history is unequaled and makes thorough backtesting possible. Two, the TRMI are not composed of only one data source; data is gathered from a wide variety of both news and social media sources. Third, its advanced linguistic software is capable of extracting specific emotions rather than just bipolar sentiment.

4.1 Four types of web content

In assessing the influence of digital media on financial markets, we first have to think carefully what content to use. There are many different platforms through which internet users share their opinions and inform themselves. Different platforms serve different purposes and each purpose attracts a certain type of user and behavior. Compare a platform like Facebook, where people share their social activities and state of mood, with social media platform eToro which is specifically intended for sharing trading activity. Their user bases may overlap, but it is very unlikely that the same content is shared. One could therefore imagine that market related comments on eToro might have higher predictive power than general states of mood, which are likeable on Facebook. Differences between platform purposes become even more obvious when comparing Facebook content with for example a financial journalist’s blog on the Financial
CHAPTER 4. DATA COLLECTION AND SELECTION

also, some groups of users are more influential and listened to than others. Compare the individual retail investor who shares his thinking on buying Apple stock on Twitter, with a research analyst from an investment bank who issues a buy signal for Apple on for example the JP Morgan website. Based on the differences in skill, influence, and distribution channels, we think important characteristics of each data source are lost when aggregating them. We identify 4 different types of web content that have their own way of possibly influencing financial markets.

the first category is professional news. This content is generated by well-informed financial market participants who make their living from financial markets. Think of bankers, reporters, and financial columnists who express their market views through serious newswires like the Wall Street Journal, Bloomberg or the Financial Times. Digital news in this category comprises both articles that are quite objective, as well as columns or blogs that could be somewhat subjective. This news category addresses all financial markets on both a macro and micro level. Topics range from commodities to real estate to bonds but also to individual stock picking. This content is read by the entire financial industry; big institutions trade upon it.

second, we consider content generated by semi-professionals or what we like to call ‘financial hobbyists’: the retail investor. These participants do not necessarily work in the financial industry but do trade for their own account and have some sense or view of what is going to happen. They share and form their opinions through blogs and forums of investment websites (like the dutch IEX.nl), or specialized social media platforms like eToro, StockTwits, or TweetTrader.net. The latter two capture stock relevant tweets, which contain a dollar sign to mark stocks or financial assets. Financial hobbyists mostly participate on the easily accessible equity market and discuss stock performance. They also keep their eye on major financial newswires.

we call the third category content on retail consumption. This category comprises discussions, customer reviews, video blogs and basically everything that is not directly related to a company’s stock performance, but rather the company’s service or product line. The users do not necessarily trade on financial markets, but use social media channels to express their feelings on companies, products, promotions, advertisements and so forth. For example, someone who tweets that he dislikes the new iPhone 5 would fall in this category. Another starting a forum topic on how to repair a Dell laptop, is also included. The idea is that positive sentiment regarding product lines like the iPhone 5 or Dell laptop stimulate sales and eventually translate into better stock performance. This thought is also the danger that looms in this data type: there may not be a causal connection at all.

the fourth type of content that could affect financial markets is general well-being. The content is not company-related and concerns everyday life. By extracting its sentiment, we could have an indication of the mood state of millions, and investigate whether it is the masses that move the markets. In this light, Facebook’s Gross National Happiness index represents the
mood of US citizens and has been proven to help predict equity markets in Karabulut (2013). This kind of macro level sentiment might reveal information on welfare or consumer confidence. Those have traditionally been market drivers.

We will see in Section 4.3 that our data covers the first two types of content: professional news and social media content with an investment focus. The third and fourth being equally interesting, this could be a potential field for future research. The third category of retail consumption content could for example be captured by analyzing the data of official Twitter data vendors. They control what they call the full Twitter firehose and can provide enough history for proper backtesting. Another way to capture this type of data is by performing (Google) custom search queries and counting the number of mentions on for example the iPhone product line.

4.2 What information to extract

Having selected a source, text analysis is the next step. To score content quickly and objectively linguistic algorithms have to be used. Besides linguistics there are other characteristics of media data that could be interesting. This paragraph mentions some of the market relevant measures that could be retrieved from a message.

Volumes are simply the number of mentions of certain keywords or topics. First of all, this type of entity reference can help determine if an article concerns the topic of interest at all. Secondly, if it does, then the number of mentions within the article reveals information on the relevance of the article on that topic. So is an asset just mentioned once in comparison to another, or is the entire article about it?

Sentiment is derived with linguistic algorithms and indicates whether a text is positive or negative on a topic. Basically, an algorithm searches for keywords around the topic mentioned and looks for up and down words in a dictionary. ‘Terrible’ will yield a negative score, but ‘terribly good’ should yield a (double) positive score. The algorithms are often quite complex, therefore we treat them separately in Section 4.4.

Novelty is extracted by comparing words in the text with past news to see how often the topic has already come to attention. The idea is that more novel stories have more impact on markets. However scoring is not easy as one has to search for keywords in a history database. Our data does not include this potentially relevant statistic.

4.3 The Thomson Reuters MarketPsych Indices (TRMI)

The data chosen for this paper come from the company MarketPsych Data and is provided by Thomson Reuters. The Thomson Reuters MarketPsych Indices (TRMI) are minutely updated
sentiment indices that comprise time series of human emotions derived from online media sources. All web content crawled from the internet is screened for its financial relevance, and consequently emotions are extracted that are specific to several financial markets. The TRMI thus provide a way of quantifying the emotional pulse of the market. The data are quite superior to those used in earlier cited work with respect to the number and variety of sources included, as well as its long history starting in 1998.

The TRMI make a distinction between content derived from news and content derived from social media. This allows us to compare the impact of professionalized news with the retail investment type of content (Section 4.1). For the first category, over 50,000 internet news sites are scraped every day, including leading newswires like The New York Times, The Wall Street Journal, and Financial Times. Less influential news sources are captured through crawler content from Yahoo! and Google news aggregators. For the sake of brevity we will from here on simply refer to this category as ‘news’.

The TRMI social media content comes from over 2 million social media sites. Primary sources include StockTwits, Yahoo! Finance, Blogger and other common chat rooms, forums and blogs. The collection starts in 1998 with some small internet forums, but only really kicks off with the rise of big social media platforms in the second half of the previous decade. Although MarketPsych claims to capture the top 30 percent of blogs, microblogs, and other social media sources, a big portion concerning retail consumption mentions is excluded from the equity indices we analyze. The underlying thought is that a forum on how to repair a Dell laptop does not add value to the forecasting power of an emotion time series on technology stocks. Therefore the TRMI social media data perfectly capture the second group of content, the retail investment type. From here on we will call this category ‘social media’, abiding by the distinction in series made by the data supplier MarketPsych.

Scraping all sources minutely, the entire content set includes over 2 million new articles and posts every day. Within minutes of publication, any new content is processed into the TRMI feed, after which advanced linguistic software scores the content specific to companies, currencies, commodities and countries. In this way, 100,000 articles downloaded from blogs, chat rooms, and news feeds are analyzed and incorporated in the time it takes a human to read two such articles (Peterson, 2013).

The TRMI track a broad range of entities including 29 currencies, 34 commodities, and 119 countries. This research focuses on 10 out of 41 equity indices that correspond to the 10 MSCI US equity sectors. We realize that some of the macro sentiment indices related to the United States or the USD might be of influence when predicting equity market returns. However, we decide not to control for additional variables in order to keep the project feasible. Including them is recommended for future investigation. In addition, Stambaugh et al. (2012) find that the ability of sentiment to predict returns is robust to the addition of macro-related fluctuations.
like the real-interest or inflation rate. Therefore macro variables like GDP or unemployment rates are also excluded.

4.4 MarketPsych linguistics and scoring

After media texts are downloaded from the internet, whether of the news or social media type, the TRMI employ advanced linguistic algorithms. The process is broadly described in the following; a more detailed description can be found in the MarketPsych white paper (Peterson, 2013).

MarketPsych’s text analysis techniques were designed to score business-specific language for quantitative financial applications. The linguistic software starts by identifying explicit entities like the company IBM. This process is complicated by the wide variety of aliases on the internet. Consider for example ‘IBM’, ‘Big Blue’ and ‘International Business Machines’ all referring to the same company. A list with over 60,000 companies and entity names is used to ensure content is associated with the proper entity.

In a second step, the software utilizes classifier algorithms to identify sentiment in a text. Specific words and phrases are recognized in the body of text by comparing it to information from specific curated dictionaries, like modifier words (like ‘small/large’) and parts of speech such as verbs. Similar psycho-social categories based on the Harvard General Inquirer lexicon are used in Tetlock (2007) and Mao et al. (2011). A variety of approaches can then be used to score sentiment, the simplest being ‘bag of words’. In this technique, words are counted according to their frequency, no additional grammatical analysis is performed. For example, the count of the word ‘up’ versus ‘down’ could be translated into a one dimensional score on positivity for a certain topic.

A drawback of a simple sentiment score derived by counting word frequencies is that it is unable to capture other emotions implied by grammatical structures. The production of the TRMI therefore involve more advanced algorithms which employ not only grammar, but also machine learning to solve ambiguities in the text. Machine learning algorithms identify correlated words in the proximity of certain entity references to recognize ambiguities. For example, gold and silver are commonly spoken of as both commodities and constituents of jewelry, but every two years they are frequently mentioned as Olympic medals.

Furthermore, the lexicons are modified to account for the variations in data sources. Twitter language with its hash tags, abbreviations, and popular words is obviously quite different from a respected financial newswire’s article. The phrase ‘That trade was the bomb!’ is recognized as a reference to a successful trade, rather than warfare which would be picked up by simplistic linguistic software. As a last point, the linguistic software needs not only to be specific to its source but also to its time. Words that used to indicate a certain emotion in the past may not
be usable for that purpose anymore.

The breadth of coverage is thus much wider than traditional bipolar positive/negative sentiment. This allows the TRMI to score along a number of dimensions including specific emotions, expectations, uncertainty, and urgency. Next to that, the TRMI include an array of one- and two-directional scores on asset-specific topics. Textual characteristics that indicate speculative bubbles are for example translated into a variable called \textit{MarketRisk}. These so-called buzz metrics give an indication of the amount of discussion on macroeconomic topics such as \textit{litigations} and \textit{mergers}. \textit{Buzz} is the word MarketPsych uses for volumes. In total there are 24 variables available for each equity sector index, see Table 2.2. The series are extensively discussed in Chapter 5 and an example of the raw data is shown in Figure 5.2.

Although the complexity of the lexical algorithms allows for analysis of many different emotions, it also generates uncertainty for the researcher. We do not have exact insight into how the language algorithms and scoring principles work precisely, nor can we influence them. The black box becomes even less transparent due to the evolution of language vocabulary and automatic machine learning. Keeping this in mind, the TRMI are very structured data series. There is little room for the researcher to influence the way the series are composed, but the final product is easy to use. We therefore take the complexity and uncertainty for granted, and carry on with a financial econometric approach to analyzing these multidimensional emotion time series.
Chapter 5

Data description

Here the analysis starts by carefully examining the return and TRMI data. The first part covers
the dependent equity sector returns, taken from the MSCI US. The second part covers the TRMI
as sentiment covariates. Graphs are shown to exemplify the applied transformations.

5.1 Market data

As dependent variable we use the MSCI US Equity Sector Indices. Each of these indices repre-
sents companies specific to one of ten industries as classified by the Global Industry Classification
Standard (GICS). All 10 subsets together constitute the broader MSCI US Index. A list of all
sectors and their abbreviations can be found in Table 5.1.

Modeling returns of United States equity is preferred over modeling MSCI World returns,
because our sentiment regressors are specifically designed to capture emotions of the US market.
MarketPsych’s language algorithms analyze English web content only, meaning we do not cap-
ture local emotions on for example the Japanese stock market. A behavioral pattern called home
bias suggests investors preferably invest in their domestic market, suggesting we would indeed
miss sentiment of important local investors when modeling a foreign market like the Japanese
(Moskowitz, 1999). With the MSCI World consisting of approximately 45% non-US equity, this
portion is too large to be missed. On the other hand, English web content by foreigners about
American stocks is picked up by the sentiment indices. We therefore expect the TRMI to be a
good reflection of both domestic and foreign emotions on the US market.

The data set contains daily closing prices of the 10 MSCI US equity sectors and is obtained
from the Thomson Reuters database. We select the same period as is available for the sentiment
variables: January 1, 1998 till June 30, 2013. Without weekends, the total number of obser-
vations is 4042, of which one is lost by taking log returns. Public holidays like Independence
day are included but the reading is the same as the day before. We do not aggregate the data
on a weekly or monthly level as we need many observations for proper MS-EGARCHX model
CHAPTER 5. DATA DESCRIPTION

<table>
<thead>
<tr>
<th></th>
<th>CD</th>
<th>CS</th>
<th>E1</th>
<th>FN</th>
<th>HC</th>
<th>ID</th>
<th>IT</th>
<th>M1</th>
<th>T1</th>
<th>U1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.00</td>
<td>-0.02</td>
<td>-0.01</td>
</tr>
<tr>
<td>min</td>
<td>-5.66</td>
<td>-9.70</td>
<td>-8.90</td>
<td>-11.30</td>
<td>-8.87</td>
<td>-5.35</td>
<td>-6.46</td>
<td>-6.13</td>
<td>-7.59</td>
<td>-10.30</td>
</tr>
<tr>
<td>max</td>
<td>3.83</td>
<td>5.31</td>
<td>11.38</td>
<td>9.77</td>
<td>4.98</td>
<td>4.22</td>
<td>10.75</td>
<td>7.20</td>
<td>7.67</td>
<td>7.12</td>
</tr>
<tr>
<td>std. dev.</td>
<td>0.61</td>
<td>0.93</td>
<td>1.27</td>
<td>1.12</td>
<td>0.89</td>
<td>0.58</td>
<td>1.06</td>
<td>1.03</td>
<td>1.06</td>
<td>1.10</td>
</tr>
<tr>
<td>skewness</td>
<td>-0.22</td>
<td>-0.43</td>
<td>-0.12</td>
<td>0.58</td>
<td>-0.27</td>
<td>-0.23</td>
<td>0.25</td>
<td>0.09</td>
<td>0.01</td>
<td>-0.36</td>
</tr>
<tr>
<td>kurtosis</td>
<td>5.71</td>
<td>8.34</td>
<td>4.99</td>
<td>16.03</td>
<td>6.64</td>
<td>6.71</td>
<td>7.93</td>
<td>4.87</td>
<td>5.51</td>
<td>5.09</td>
</tr>
<tr>
<td>JB test</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ADF test</td>
<td>0.91</td>
<td>0.74</td>
<td>0.40</td>
<td>0.61</td>
<td>0.83</td>
<td>0.46</td>
<td>0.59</td>
<td>0.27</td>
<td>0.94</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 5.1: Descriptive statistics of the MSCI US sector excess returns. The excess returns are obtained by subtracting the return of the general MSCI US index from each sector’s return. P-values of JB and ADF tests are reported in the bottom two rows. From left to right: Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Information Technology, Materials, Telecom and Utilities.

The top panel of Figure 5.1 plots the data for the Financials sector. The Lehman crisis of 2008 is clearly visible, which was by far the most important shock for the financial sector in this time span. It can easily be seen that the index itself is not stationary; indeed augmented Dickey-Fuller (ADF) tests do not reject non-stationarity for all 10 indices. This is a very common finding in financial time series often solved by taking log returns. Also returns have more attractive statistical properties than prices, of which stationarity is one (Danielsson, 2011). To examine the effects of sector specific emotions on their particular sector’s outcome, excess returns are created by subtracting the general MSCI US log return from each individual sector $j$’s log return:

$$r_{jt} = (\log P_{j,t} - \log P_{j,t-1}) - (\log P_{US,t} - \log P_{US,t-1})$$

This way we make sure that our sentiment series do not simply predict general market moves, but sector specific movements. The bottom panel of Figure 5.1 shows these excess returns for the Financials sector. Two other big crises immediately become visible: the internet bubble around the year 2000 and the Euro crisis of 2012. Judging from the graph, there is obvious volatility clustering in play with three major high volatility regimes. The Markov switching model as described in Section 3.2 should capture at least these periods.

Table 5.1 reveals the descriptive statistics and unconditional moments of the return data. Note that for all equity sectors, the mean excess return over the entire sample of 15 years is close to zero. Furthermore, the sector returns suffer from a well-known peculiarity among financial time series: excess kurtosis. Skewness is also different from its normal distribution value of zero. Non-normality is confirmed by Jarque-Bera tests that firmly reject for every sector. Energy has
CHAPTER 5. DATA DESCRIPTION

Figure 5.1: The top panel displays the Financials sector MSCI index. Its excess return over the more general MSCI US index is shown in the bottom panel. We clearly identify three crisis periods: the internet bubble around 2000, the Lehman crisis of 2008 and the Euro crisis in the summer of 2012. Observation are daily, starting January 1998 and ending on June 30th 2013.

witnessed the highest above market return, Financials the lowest. Without doubt the latter is to a large extent due to the recent financial crisis. The crisis might also be reflected in the enormous kurtosis of Financials compared to other sectors. Note furthermore that Energy and Financials were the most volatile, whereas Consumer Discretionary and Industrials are very stable sectors.

It is important to stress that we are not actually predicting sector returns, rather we are looking at sector under-/outperformance compared to the market. A sector can outperform the market and have a positive excess return, even though that sector’s actual return is negative. For the sake of brevity we will from here on skip the words ‘excess’ and ‘log’ when talking about sector returns.

A correlation table does not reveal any significant correlations. Each sector’s returns compared to the others seem to be quite independent, which allows for diversification. The highest positive correlation is observed between Consumer Staples and Healthcare, namely 0.536. The
lowest of -0.516 is between Consumer Staples and Information Technology. Keep in mind that this table reports correlations between excess returns, meaning two sectors might move in the same direction, but if one moves more than the average US return and one moves less, the two sector excess returns have contradicting signs. Indeed when looking at correlations of the indices themselves, the correlation between Energy and Materials becomes as high as 0.917.

The statistics in this section reveal that there are quite some differences in sector specific returns. A couple of sectors are more volatile and suffer from more kurtosis than others. One of the aims of this research is to discover whether these differences are caused by differences in sentiment. We do expect some sectors to be more prone to digital media sentiment as retail investors typically only buy what they hear is in the news.

<table>
<thead>
<tr>
<th></th>
<th>CD</th>
<th>CS</th>
<th>E1</th>
<th>FN</th>
<th>HC</th>
<th>ID</th>
<th>IT</th>
<th>M1</th>
<th>T1</th>
<th>U1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>1.00</td>
<td>-0.06</td>
<td>-0.24</td>
<td>0.04</td>
<td>-0.16</td>
<td>0.16</td>
<td>-0.06</td>
<td>0.04</td>
<td>-0.07</td>
<td>-0.16</td>
</tr>
<tr>
<td>CS</td>
<td>-0.06</td>
<td>1.00</td>
<td>0.09</td>
<td>-0.26</td>
<td>0.54</td>
<td>0.00</td>
<td>-0.52</td>
<td>0.04</td>
<td>0.13</td>
<td>0.43</td>
</tr>
<tr>
<td>E1</td>
<td>-0.24</td>
<td>0.09</td>
<td>1.00</td>
<td>-0.14</td>
<td>0.04</td>
<td>-0.02</td>
<td>-0.38</td>
<td>0.32</td>
<td>-0.06</td>
<td>0.31</td>
</tr>
<tr>
<td>FN</td>
<td>0.04</td>
<td>-0.26</td>
<td>-0.14</td>
<td>1.00</td>
<td>-0.26</td>
<td>0.07</td>
<td>-0.27</td>
<td>0.02</td>
<td>-0.15</td>
<td>-0.22</td>
</tr>
<tr>
<td>HC</td>
<td>-0.16</td>
<td>0.54</td>
<td>0.04</td>
<td>-0.26</td>
<td>1.00</td>
<td>-0.10</td>
<td>-0.44</td>
<td>-0.09</td>
<td>0.04</td>
<td>0.29</td>
</tr>
<tr>
<td>ID</td>
<td>0.16</td>
<td>0.00</td>
<td>-0.02</td>
<td>0.07</td>
<td>-0.10</td>
<td>1.00</td>
<td>-0.23</td>
<td>0.33</td>
<td>-0.17</td>
<td>-0.05</td>
</tr>
<tr>
<td>IT</td>
<td>-0.06</td>
<td>-0.52</td>
<td>-0.38</td>
<td>-0.27</td>
<td>-0.44</td>
<td>-0.23</td>
<td>1.00</td>
<td>-0.30</td>
<td>-0.09</td>
<td>-0.41</td>
</tr>
<tr>
<td>M1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.32</td>
<td>0.02</td>
<td>-0.09</td>
<td>0.33</td>
<td>-0.30</td>
<td>1.00</td>
<td>-0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>T1</td>
<td>-0.07</td>
<td>0.13</td>
<td>-0.06</td>
<td>-0.15</td>
<td>0.04</td>
<td>-0.17</td>
<td>-0.09</td>
<td>-0.11</td>
<td>1.00</td>
<td>0.13</td>
</tr>
<tr>
<td>U1</td>
<td>-0.16</td>
<td>0.43</td>
<td>0.31</td>
<td>-0.22</td>
<td>0.29</td>
<td>-0.05</td>
<td>-0.41</td>
<td>0.07</td>
<td>0.13</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.2: Correlation table of sector returns. The highest positive correlation is observed between Consumer Staples (CS) and Healthcare (HC), namely 0.536. The lowest of -0.516 is the correlation between Consumer Staples (CS) and Information Technology (IT).

### 5.2 Sentiment variables

The exogenous variables in this investigation are the 24 TRMI on equity specific sentiment in Table 2.2. The indices are available on a minutely and daily basis, the latter simply being a 24 hour average of the first. The daily readings come in every day at 20:30, with the first observations on January 1, 1998 for both the news and social media data type. The last reading in our sample is on June 30, 2013 totaling 5,660 observations. There are 809 weekends in this 15.5 year time frame, leaving us 4,042 weekdays. The first day is excluded, as we take first differences of the MSCI US equity series.

Although available since 1998, we decide to start using the social media sentiment series
from August 2006 onward, when Twitter was launched. We find a steep increase in social media volumes from that point, affirming a trend break with the data before 2006. More on structural breaks will follow in the next subsection. Because of this we are left with only 1,804 readings. It remains questionable whether this is enough to feasibly estimate the MS-EGARCHX model with its at maximum 20 parameters.

In the remainder of this chapter we discuss the peculiarities or characteristics of the data. We correct for structural breaks, daily seasonality and missing values. Also each sentiment observation is weighted with its relative daily buzz (volume), so that more weight is attached to an emotion of many rather than a feeling among few. In a last step we take differences as markets often react to the direction rather than the absolute level of a statistic. We look at the 1 day, 1 week and 4 week change of each variable which will reveal whether information lies in daily mood swings (1 day change) or longer term sentiment momentum (1 week and 4 week change).

5.2.1 Structural breaks

When looking at plots of the news time series the first thing that catches the eye are two apparent structural breaks. The breaks coincide with two major additions of data sources in the MarketPsych news data feed. The first break is in 2003 adding Reuters and a couple of other major third-party newswires. Judging from the top panel in Figure 5.2, this news was structurally more positive than the content generated by the original sources. The second break occurs in 2005 when MarketPsych adds the aggregated news feed of Moreover Technologies to their data collection process. Again the level of sentiment structurally changes after this break.

For some series the breaks do not necessarily affect the level of an emotion, but in all cases they do impact a series’ variance. As volumes are a lot lower in the first two periods, their respective outcomes are more extreme. Volatility in period one is therefore higher than volatility in period two, which is in turn higher than the volatility in period three. This means that for some sentiment variables, both levels and volatility are structurally changed by the way data are gathered over time. To correct for this, we standardize each period by using its period mean and standard deviation. Means and standard deviations are calculated using the entire subsample, secretly introducing a forward looking bias. However, we cannot ignore this problem caused by the underlying data generating process as we want to compare observations from every period.

Another consequence of the addition of extra data sources lies in the volumes, buzz. During periods 1 and 2, buzz was more or less constant, however in period 3 the amount of news grows over time. This trend in news buzz can be removed by dividing buzz by its four week moving average. Thus we look at the relative weights,

\[
\text{buzzwgt}_t = \frac{\text{buzz}_t}{\frac{1}{28} \sum_{i=0}^{27} \text{buzz}_{t-i}}.
\]  

(5.1)
CHAPTER 5. DATA DESCRIPTION

Working with buzzweights (buzzwgt) instead of buzz has two advantages. First, by dividing a buzz observation by its rolling 4 week moving average, we remove the time trend of increasing volumes. This allows us to compare relative buzz values from for example 2013 with 1998. Second, the buzzweights can be multiplied with the sentiment observations. This causes emotions associated with large buzz to be weighted more heavily than the same emotion associated with a small buzz. This will be discussed in more detail in Subsection 5.2.4. For social media, we also calculate buzzweights to remove a trend in buzz. Although there are no structural breaks in the shorter subsample, we do standardize the data to make the estimates of news and social media comparable.

Figure 5.2: Both charts depict the news sentiment (SNTMENT) specific to the Industrials sector (ID). The top panel clearly shows the existence of two structural breaks in the raw data, which we compensate for by standardizing each period. It furthermore shows that weekday data (red) are apparently more positive than weekend data (blue). A boxplot of the standardized series in the bottom panel confirms this and provides an indication of the daily seasonality present in this particular series.
5.2.2 Daily seasonality

A second characteristic of the data becomes apparent when zooming in or considering daily averages in a boxplot: seasonality. Some of the sentiment series are highly subject to daily seasonals as can be observed for the variable sentiment in the Materials sector (Figure 5.2). In the top chart we already saw that sentiment is more positive during the week than on weekends, but from the boxplot we also find intra-week differences. Apparently sentiment on Industrials equity depresses over the week from Monday to Friday. Seasonality in regressors adds noise to the estimates as it is unknown whether a positive change from one day to another is caused by an actual increase in sentiment or a structural difference between the two days. We therefore correct for the daily seasonality with the simple additive decomposition model,

\[ Y_t = T_t + C_t + S_t + I_t. \]

The model assumes that the observed time series \( Y_t \) is composed of four independent components: a trend \( T_t \), a cyclical component \( C_t \), seasonals \( S_t \), and finally an irregular component \( I_t \). All sentiment variables look stationary based on their graphs which is confirmed by ADF tests that show no evidence of a trend of any kind; stochastic nor deterministic. Remember that the trended buzz series has already been transformed to the stationary buzzweight series, so \( T_t = 0 \) for all TRMI.

Concerning the cyclical component \( C_t \) one could easily imagine a series like joy to be of higher value during the summer, or gloom during the winter. In this research we decide not to correct the sentiment series for cyclicality for three reasons. One, changes in the sentiment values involve two readings which both possibly suffer from the same longer term cycicality. Subtracting the cyclical effects from each one would thus make no difference. This is especially the case for the day-to-day change covariates (Section 5.2.6). Second, we also do not remove cyclicals from the return data. Cyclical, if any, might overlap so that for example the well documented ‘tax-loss-selling’ and ‘January’ effect in monthly stock market returns are reflected in both dependent and explanatory variable. Third, we would introduce another forward looking information bias by estimating cyclical components on the entire data set and subtracting them afterwards. Therefore we set \( C_t = 0 \), leaving only the daily seasonal component \( S_t \) and the irregular factor \( I_t \). The deseasonalized time-series is obtained by subtracting the seasonal effects from the \( Y_t \) observations. The seasonal effects are estimated as the averages of each day over the entire time frame. As we are not using an expanding window for the averages, we again introduce a slight form of forward looking information bias (next to the one created by standardizing the data). Technically the standard errors of estimated coefficients should be corrected for this. However, asymptotically the estimation error is minimized due to the large number of observations.
5.2.3 Unavailable data dummies

If there were no conversations involving a specific emotion or topic in a certain 24 hour period no observation is reported for that particular day. A missing value thus differs from a zero reading in the sense that the latter did involve buzz, but people were neutrally disposed on the topic. There is quite a difference in the number of missing values for the emotion-type TRMI compared to the buzz metrics (see Section 4.3). Emotional TRMI like joy, fear and sentiment are almost always available, while buzz metrics like marketforecast, innovation and layoffs have the most missing values. There are also differences among industries. For example, there are more missing values on mergers in the Healthcare industry compared to the way larger and more competitive Financials sector.

In order to estimate our heteroscedasticity models we need to substitute all blank fields. This could be done by fixing them to zero, which is the mean of all variables after the standardization of structural breaks discussed in Subsection 5.2.1. However the variables are not normally distributed and the median often substantially differs from zero. Imputing a zero would therefore skew or bias the distribution of a variable and could consequently ruin the statistical properties of that variable. A better option is to assume that sentiment levels have remained the same in the absence of a new reading. Carrying the last observation forward implies that the 1 day difference regressor will be zero in case of a missing value. A missing Wednesday value is thus substituted by the Tuesday before. For Mondays this means that Friday’s value is inserted, as this is the last trading day prior to Monday. Weekends in between are treated differently as will be explained later in this chapter.

An important property that should not be ignored is that the missing values are not always random. Namely, for some variables with relatively many missing observations, the blanks are clustered. This means that periods with no talk on for example layoffs are alternated by periods with increased reporting on that topic. A period with no news on people being fired could correspond with a relatively tranquil period with no major events. To test this hypothesis of missing values predicting a lower volatility, we include the information by adding a dummy that is 1 when volatility regressor $v_t$ misses a value at time $t$:

$$D_{t}^{NA} = 1 \text{ if } v_t = NA.$$  \hfill (5.2)

5.2.4 Buzz weighting and outliers

At this stage, after correcting for structural breaks, seasonality and missing values, we apply our final transformation. Each observation of every series is multiplied by the buzzweight for that particular day (Equation (5.1)). As already briefly explained earlier this chapter, this transformation differentiates sentiment associated with large buzz from the same absolute sentiment level caused by just a little buzz. An emotion felt by many (high buzz) is thus more heavily
weighted than the same emotion experienced by only a few people (little buzz). Weighting each sentiment with its buzz has its pros and cons.

An advantage is that an extreme reading caused by only a few overly enthusiastic sources is downgraded, because there is less online content than usual. A daily buzz below its 4 week average implicates a buzzweight smaller than 1. It is important to correct for this kind of outliers as they are likely not big enough to move the market.

The drawback is that outliers caused by hyped press releases now become overweighted. Once in a while, a certain topic grabs major attention even though the actual impact of the story might not even be that significant. A low sentiment value could in that case be multiplied with a very high buzzweight, resulting in an outlier. In the best case, each outlier should therefore be investigated individually in order to determine its relevance in forecasting equity returns. However, it is practically impossible to check all outliers for their authenticity when having 24 variables, 10 sectors and 2 types of digital media content. Besides, it might just be these mass feelings that move the market; whether they are hyped or not.

5.2.5 Weekend dummies

A last piece of information hidden in the data is extracted from the weekends. Where weekday sentiment readings are based on trading activity, weekend sentiment are influenced by other ‘news’. In fact, the absence of trading might actually be a reason for news buzz. Consider a company that saves bad news for the weekend in order to avoid heavy stock trading. Another example would be the release of important Chinese financial data over the weekend. A third example concerns the substantial news reporting on weekend EU summits during the Euro crisis not too long ago. For social media, weekend data might be a little less relevant. This is because people discuss their private life when there is no trading going on, according to the data vendor MarketPsych.

The relevance of weekend data has been investigated in other academic work. For example, the tendency of stock markets to decrease on Mondays is documented as a well known anomaly called the ‘Monday effect’. Pettengill (2003) mention how a large amount of bad news published over the weekend could cause this drop. To control for this weekend effect, we include a dummy that is 1 or -1 on Mondays whenever the buzzweighted sentiment of the weekend before is significantly different compared to the prior twelve weekends (a 3 month period):

\[ D_{\text{mon}}^{\text{wkdnd}} = \begin{cases} 
1, & \text{if } x_{\text{wkdnd}} > \mu_{\text{prior}12} + 2\sigma_{\text{prior}12}, \\
-1, & \text{if } x_{\text{wkdnd}} < \mu_{\text{prior}12} - 2\sigma_{\text{prior}12}, \\
0, & \text{otherwise.} 
\end{cases} \tag{5.3} \]

In Formula (5.3), \( \mu_{\text{prior}12} \) and \( \sigma_{\text{prior}12} \) are the 12 week rolling mean and standard deviation. A weekend’s sentiment value \( x_{\text{wkdnd}} \) is calculated as the Sunday minus the Friday observation, both
published at 20.30 hours. The dummy is added in both the return and volatility equation of our models.

Apart from weekends, there is also no trading activity on public holidays. For the MSCI US indices, two of them are January 1 and Independence day, July 4. The indices carry the last observation forward in case of a holiday, leading to a zero return. One could potentially investigate the returns after such a non-trading day and see how sentiment on the public holiday affects it. However, this research question is too specific for this paper.

5.2.6 Final remarks

With the corrections discussed in this chapter, the TRMI are now ready for regression. In a last step we calculate the 1 day, 1 week, and 4 week changes, which, together with the two dummies in Equations (5.2) and (5.3) form our set of covariates for each variable:

\[
\begin{align*}
    x^{1D}_t &= x_t - x_{t-1} \\
    x^{1W}_t &= x_t - x_{t-7} \\
    x^{4W}_t &= x_t - x_{t-28}.
\end{align*}
\]

(5.4)

Observing the three different changes tells us which sentiment trends to look at; short, medium, or longer term trends. The next chapter describes the estimation results.

The remainder of this subsection is a critical side note to the data used and the transformations applied. We have to acknowledge that there is a lot of uncertainty in the data. Not only is the way the data are constructed quite a black box to us (as noted in Section 4.4); we also do not exactly know how its curiosities come to be. I.e., some series seem to be extremely noisy for no immediate reason, which is especially true for the social media data type. We cannot tell whether this is caused by unstable linguistic machine learning algorithms or whether sentiment on financial markets is actually that wiggly. Furthermore we find an explosion in volumes over the last couple of years indicating the increased importance of the role that digital media play in our everyday lives. Relationships are likely not to have stabilized yet. In analyzing and drawing conclusions, we therefore keep in mind that this topic is still very much of a relatively new field of research.
Chapter 6

Results

This chapter contains the results of the Markov switching heteroscedasticity models for both the news and social media data type. Examples are given for one industry specifically but findings are similar across other sectors. We furthermore employ two series of out-of-sample tests to assess the quality of our return and Value at Risk predictions.

6.1 Preliminary analysis

As we are dealing with 24 sentiment variables, from 2 different data sources, possibly affecting both the returns and the volatility of 10 different industries, it is clear that we have to start with a careful preliminary analysis to filter out the most important relations. A specific-to-general two-pass estimation approach is applied to assess each variables’ predictive power individually.

In the first step, the 3 lags of a variable and its dummies are inserted in a basic ARX estimation (Equation 3.1) per sector. This reveals whether that variable has any predictive capabilities, and if so, what information to look at (1D, 1W, 4W lag, weekend or NA dummy, Chapter 5). If more than 2 variables were significant in predicting one sector’s returns, we select the best two in a combined ARX estimation. We find that some variables lose their significance when included with other sentiment covariates.

In the second step we take the residuals of the combined first step ARX estimation and fit an EGARCHX model to them (Equation 3.4). Each variable is again individually assessed regarding its relation with volatility, after which again the best two are selected through a combined EGARCHX estimation. Although two-step estimates often provide good approximations of the true parameters in relatively large samples like these, we do estimate the mean and volatility equation jointly to obtain the true parameters. These will be compared to the Markov switching models later. Note that the estimated signs of regression coefficients have to satisfy the prior beliefs formed in Section 2.4 in order to prevent hindsight bias.

The results of the two-step preliminary analysis are found in Table 6.1. There are four
CHAPTER 6. RESULTS

### Table 6.1

<table>
<thead>
<tr>
<th></th>
<th>News data</th>
<th></th>
<th>Social Media</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variables</td>
<td>Vol.</td>
<td>Variables</td>
</tr>
<tr>
<td>CD</td>
<td>3</td>
<td>TRUST 1D</td>
<td>2</td>
<td>URGENCY wknd</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MKTRISK 1D</td>
<td></td>
<td>UNCERTN 1D</td>
</tr>
<tr>
<td>CS</td>
<td>0</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>STRESS 1W</td>
<td></td>
<td>LAYOFFS 1D</td>
</tr>
<tr>
<td>E1</td>
<td>0</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LITIG 4W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td>2</td>
<td>GLOOM 1W</td>
<td>7</td>
<td>BUZZWGT 4W</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UPGDWNG 1W</td>
<td></td>
<td>FEAR 4W</td>
</tr>
<tr>
<td>HC</td>
<td>1</td>
<td>URGENCY 1D</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>2</td>
<td>TRUST wknd</td>
<td>2</td>
<td>BUZZWGT 1D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MKTFCST 4W</td>
<td></td>
<td>OPTIMSM 1D</td>
</tr>
<tr>
<td>IT</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>1</td>
<td>PRICEUP wknd</td>
<td>2</td>
<td>URGENCY 1D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MKTRISK 1D</td>
</tr>
<tr>
<td>T1</td>
<td>0</td>
<td></td>
<td>1</td>
<td>OPTIMSM 1D</td>
</tr>
<tr>
<td>U1</td>
<td>0</td>
<td></td>
<td>3</td>
<td>OPTIMSM wknd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LAYOFFS 1D</td>
</tr>
</tbody>
</table>

Table 6.1: The table displays the results of a 2-step preliminary regression analysis. The ‘mean’ column indicates the number of significant TRMI in the first step mean equation, ‘vol’ reports on the significant series for the second step EGARCHX estimation. We use the 1% significance level as a threshold for inclusion. The ‘variables’ columns show for each case the two most significant emotions that will be investigated further in the Markov switching context.

The immediate findings that help us in answering our main hypotheses. One, sentiment seems to play a more important role in predicting volatility than returns. More emotions were found to be significant in the second step estimation (columns 4-5) compared to the first step (columns 2-3). This finding is also true for the social media data (columns 6-9). Second, news sentiment seems to be more powerful than social media sentiment. Overall, fewer TRMI were found to be significant in both the mean and volatility equations compared to news.

Third, the Financials sector (row 4) seems to be the most sentiment-sensitive in this data set, especially when compared to for example Telecom or Energy, in which rarely any sentiment was significant. This shows that there are important differences between industries to take into account. This is confirmed by the fourth conclusion that themes vary per sector and per data type. News optimism and social media fear seem to be the only recurring emotions across sectors. Social media apparently picked up the fear of defaults during the Lehman and Euro crises, whereas professional news optimistically focused on the positive signs of a solution or recovery.

A last finding is that the ‘no talk’ dummies as discussed in Section 5.2.3 are rarely significant.
and inconsistent with our prior beliefs. For example we found a significant positive sign for no talk on layoffs, but it is hard to justify: no talks on people being fired leads to higher market volatility? Due to this inconsistency we reject our side hypothesis of silence on digital media corresponding to tranquil periods of low market volatility. We exclude the dummies from further analysis. The dummies indicating extraordinary weekend sentiment however do matter and are therefore kept.

The TRMI in Table 6.1 are the ones that will be investigated in the dual regime context. The remainder of this chapter reports on the Markov switching analysis for each sector, but demonstrates the findings for just one: Industrials. Although it does not have any significant sentiment variables for the social media mean equation, the Industrials sector is a stable example for which to display the MS-EGARCHX model capabilities. Sentiment has played a larger role in the Financials industry, but the huge industry specific shocks that this particular sector has witnessed also lead to exceptional estimation results. As these results are equally interesting, we will highlight some of them throughout the chapter. Output graphs of the Financials sector are added to the appendix.

6.2 News: a dual regime constant variance model

To demonstrate the flexibility advantages of regime switching models, over single regime models we start by comparing the single regime ARX-EGARCHX model with a Markov switching model that assumes constant variances. The log-likelihood function and its estimation technique are discussed in Section 3.3. The results of both models for Industrial specific news sentiment are shown in Table 6.2.

The parameters in the first column match the notation used in Chapter 3. An additional subscript indicates the regime of a parameter in the dual state scenario. From the preliminary analysis we add trust and marketforecast as mean equation regressors in both models, whereas buzzweight and optimism drive the EGARCHX process of the single regime model only. This is because we assume constant variances $\sigma^2_i$ for the Markov switching models for now. In the next section, the model is extended by making the volatility time-varying and dependent on external influences. $P$ and $1 - Q$ are the estimated transition probabilities.

The EGARCH parameters of the single regime model look usual. The process is stationary as $\beta < 1$, and the model captures a significant leverage effect as $\alpha$ is negative. Furthermore, we find that the news regressors marketforecast and buzzweight help predict respectively the returns and volatility of the Industrials sector. The weekend dummy of trust on the contrary does not seem to be significant any longer. This is in line with returns generally being unpredictable and mean equation terms losing their significance in the presence of a volatility specification.
## CHAPTER 6. RESULTS

### Table 6.2: Comparison of the single regime ARX-EGARCHX model with a simple Markov switching model that assumes constant variances $\sigma_i^2$. The sentiment variables come from Industrial specific news data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ARX-EGARCHX</th>
<th>MS const. variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{0,1}$</td>
<td>0.0020</td>
<td>0.0054</td>
</tr>
<tr>
<td>$\phi_{1,1}$</td>
<td>0.0388</td>
<td>0.0421</td>
</tr>
<tr>
<td>$\phi_{2,1}$</td>
<td>-0.0319</td>
<td>-0.0386</td>
</tr>
<tr>
<td>$\phi_{3,1}$</td>
<td>-0.0071</td>
<td>0.0088</td>
</tr>
<tr>
<td>TRUST wknd$_1$</td>
<td>0.0648</td>
<td>0.0480</td>
</tr>
<tr>
<td>MKTFCST 4W$_1$</td>
<td>0.0166</td>
<td>0.0120</td>
</tr>
<tr>
<td>$\phi_{0,2}$</td>
<td>-</td>
<td>-0.0037</td>
</tr>
<tr>
<td>$\phi_{1,2}$</td>
<td>-</td>
<td>0.0124</td>
</tr>
<tr>
<td>$\phi_{2,2}$</td>
<td>-</td>
<td>-0.0423</td>
</tr>
<tr>
<td>$\phi_{3,2}$</td>
<td>-</td>
<td>-0.0232</td>
</tr>
<tr>
<td>TRUST wknd$_2$</td>
<td>-</td>
<td>0.4397</td>
</tr>
<tr>
<td>MKTFCST 4W$_2$</td>
<td>-</td>
<td>0.0775</td>
</tr>
<tr>
<td>$\omega$</td>
<td>-0.0000</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.0201</td>
<td>-</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.9973</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.0848</td>
<td>-</td>
</tr>
<tr>
<td>BUZZWGT 1D</td>
<td>0.1526</td>
<td>-</td>
</tr>
<tr>
<td>OPTIMSM 1D</td>
<td>-0.0619</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>-</td>
<td>0.3575</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>-</td>
<td>0.8887</td>
</tr>
<tr>
<td>$P$</td>
<td>-</td>
<td>0.9933</td>
</tr>
<tr>
<td>$1 - Q$</td>
<td>-</td>
<td>0.0149</td>
</tr>
</tbody>
</table>

Log-likelihood:  -2799.10  -2817.93

*Signif. codes: ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ‘.’ 0.1 ‘ ’ 1

In the Markov switching context we have the same mean regressors, but volatility wise there are only two options: the volatility is either 0.3575 or 0.8887. This reflects the existence of a clear low and high volatility regime. Based on the estimated transition probabilities from Equation (3.6) we determine the duration of each regime $i$ as $1/(1 - P(S_t = i | S_{t-1} = i))$. With a duration of 148.85 days, we find that the low volatility regime is more persistent than the high volatility regime (67.28 days). This is consistent with the findings of Gray (1996).
persistence of regimes does seem to vary a lot across sectors; for Financials for example the low and high volatility regime only last 70.55 and 31.99 days respectively.

Considering the parameter estimates of the regime switching model, it now becomes clear how a single regime model potentially ‘averages out’ the effect of a variable. The weekend dummy for trust is not at all insignificant, but matters at the 1% level in high volatile environments. This is a finding we see across all industries and both data types: sentiment better predicts returns in periods of high volatility compared to low volatility. This confirms our hypothesis of emotions playing a larger role in stressed markets.

Remarkable is the fact that there are significant autoregressive terms in the mean equation of both models. Efficient market theory does not expect a market’s return to depend on its first and second lag. Although interesting, we exclude the AR terms from our more advanced models for two reasons. First they do not make sense economically. Second, adding these extra parameters substantially worsens estimation time of the Markov switching models.

Clearly modeling the volatility as a binary outcome is too restrictive. This is confirmed by the lower log-likelihood, standardized residuals that do not resemble white noise, and the conditional volatility in the middle panel of Figure 6.1. The conditional volatility over both states is not flexible enough and mimics the state probability of being in the high volatility regime \( \pi_{2t} \). Despite this simplicity, the smoothed probabilities do capture important volatile periods like the internet bubble and the 2008 Financial crisis as can be seen in the top graph of Figure 6.1.

6.3 News: the regime-switching EGARCHX model

This section relaxes the constant variances of the Markov switching model in the previous section and allows for a more flexible EGARCHX volatility specification per regime. Basically the same EGARCHX model is estimated for two states, meaning that the number of parameters more than doubles. One of the main questions of this paper is therefore whether this additional flexibility is worthwhile. We also estimate the MS-EGARCH model without exogenous sentiment regressors to compare it with its emotional equivalent. The outcomes of all three models for Industrials are displayed in Table 6.3.

Not surprisingly, the EGARCHX parameters without AR terms are almost exactly identical to those in Table 6.2. Concerning the Markov switching models, we again have a clear distinction between a low and high volatility regime, most notably characterized by the highly significant constants \( \omega_i \) in the volatility specifications. Furthermore the regimes have become a lot more persistent with transition probabilities \( P \) and \( Q \) approaching unity. However within regimes, volatility is less persistent for the low volatility case which is expressed by the GARCH parameter \( \beta_1 \) being smaller than \( \beta_2 \). As \( \beta_i < 1 \) for both states, the model overall is stationary.
CHAPTER 6. RESULTS

Figure 6.1: All three panels concern outcomes of the Markov switching model with constant variances for Industrial specific news sentiment. Due to its binary volatility options, the model’s conditional volatility in the middle panel mimics the regime probability of being in the high volatility regime. The smoothed version of this probability is displayed in the lower panel. Areas in which it exceeds 0.5 are shaded gray. They overlap with the internet and Lehman crises.
### CHAPTER 6. RESULTS

Table 6.3: Comparison of three models for Industrial specific news sentiment. The MS-EGARCH model does not include any sentiment covariates. The log-likelihood and AIC improve with each additional step of model sophistication.

Another interesting finding is that the leverage effect is very much present during high volatile environments, but is not even significant when volatility is low. This result could not have been captured by a single regime model. It turns out to be a common finding across many sectors.

The results imply that the TRMI can be helpful in both return and volatility forecasting. This is not only indicated by the t-statistics, but it is also confirmed by an increasing log-likelihood and a decreasing Akaike Information Criterion (AIC). The latter represents a trade-off between a higher log-likelihood and a penalty for extra parameters. The log-likelihood is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>EGARCHX</th>
<th>MS-EGARCH</th>
<th>MS-EGARCHX</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{0,1}$</td>
<td>0.0020 (0.0074)</td>
<td>0.0052 (0.0071)</td>
<td>0.0092 (0.0074)</td>
</tr>
<tr>
<td>TRUST wknd1</td>
<td>0.0648 (0.0524)</td>
<td>-</td>
<td>0.0340 (0.0565)</td>
</tr>
<tr>
<td>MKTFCST 4W1</td>
<td>0.0166 (0.0085)*</td>
<td>-</td>
<td>0.0020 (0.0089)</td>
</tr>
<tr>
<td>$\phi_{0,2}$</td>
<td>-</td>
<td>-0.0181 (0.0123)</td>
<td>-0.0139 (0.0081)</td>
</tr>
<tr>
<td>TRUST wknd2</td>
<td>-</td>
<td>-</td>
<td>0.1074 (0.0658)</td>
</tr>
<tr>
<td>MKTFCST 4W2</td>
<td>-</td>
<td>-</td>
<td>0.0263 (0.0096)**</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>-0.0000 (0.0001)</td>
<td>-0.4461 (0.1150)**</td>
<td>-0.4192 (0.0885)***</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.0201 (0.0087)*</td>
<td>-0.0078 (0.0187)</td>
<td>-0.0141 (0.0190)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.9973 (0.0012)***</td>
<td>0.8319 (0.0481)***</td>
<td>0.8442 (0.0370)***</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.0848 (0.0231)***</td>
<td>0.0778 (0.0314)*</td>
<td>0.0675 (0.0322)***</td>
</tr>
<tr>
<td>BUZZWGT 1D1</td>
<td>0.1526 (0.0488)**</td>
<td>-</td>
<td>0.1853 (0.0430)***</td>
</tr>
<tr>
<td>OPTIMSM 1D1</td>
<td>-0.0619 (0.0386)</td>
<td>-</td>
<td>0.0600 (0.0355)</td>
</tr>
<tr>
<td>$\omega_2$</td>
<td>-</td>
<td>-0.1154 (0.0195)***</td>
<td>-0.0856 (0.0141)***</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-</td>
<td>-0.0787 (0.0164)***</td>
<td>-0.0685 (0.0120)***</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-</td>
<td>0.9423 (0.0192)***</td>
<td>0.9752 (0.0101)***</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>-</td>
<td>0.1365 (0.0222)***</td>
<td>0.1065 (0.0165)***</td>
</tr>
<tr>
<td>BUZZWGT 1D2</td>
<td>-</td>
<td>-</td>
<td>0.0847 (0.0400)***</td>
</tr>
<tr>
<td>OPTIMSM 1D2</td>
<td>-</td>
<td>-</td>
<td>-0.0911 (0.0296)**</td>
</tr>
<tr>
<td>$P$</td>
<td>-</td>
<td>0.9990 (0.0006)***</td>
<td>0.9991 (0.0006)***</td>
</tr>
<tr>
<td>$1 - Q$</td>
<td>-</td>
<td>-</td>
<td>0.0001 (0.0001)</td>
</tr>
</tbody>
</table>

Log-likelihood: -2804.62 -2790.35 -2764.78

AIC: 1.400 1.394 1.385

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
demonstrated to be significantly higher by means of a likelihood ratio (LR) test. The degrees of freedom used for the $\chi^2$-distribution equal the difference in number of parameters between two models. We find that Markov switching is an improvement upon the single regime model (LR statistic of 72.85) and adding sentiment to it improves the model fit even more (LR statistic of 51.19). Sentiment does not add predictive power in every case. An overview of all test results is therefore given in Table 1 of the appendix.

Because of the preliminary selection, all significant signs of the TRMI are consistent with our prior beliefs (Table 2.2). Interpretation-wise, we find that predictions of Industrial asset prices made in professional news (marketforecast) actually drive prices upward during high volatile periods. Increasing optimism in professional news is a good indicator of volatility calming down again. More volume of news messages leads to higher volatility in both regimes. An overview of the significant sentiment variables across industries and data types can be found in appendix Table 2.

Looking at the smoothed state probability in the lower part of Figure 6.2, we immediately observe that the regimes have become more persistent. Where the constant variances model switched relatively often, small volatility spikes have now basically blurred into two main periods of heightened market stress. More interesting therefore are the ex ante probabilities shown by the black line in the same graph. Periods in which the ex ante probability exceeded 0.5 are shaded gray. We see that the probability of being in the high volatility regime gradually builds up in the years before the Financial crisis, but remains rather volatile itself. Therefore it is hard to say whether the statistic was a good predictor of 2008’s steep market decline. For the Financials sector a very sudden regime switch occurs at the end of 2007, before the long market decline. Again, it might not have predicted the crisis in time, but it was definitely indicating upcoming trouble. The graphs for Financials can be seen in Figure 2 in the appendix.

The middle panel of Figure 6.2 shows the MS-EGARCHX model’s conditional volatility. It is a more flexible representation of the unobserved market volatility than the MS model with constant variances of Figure 6.1.

To truly determine the additional predictive power induced by regime switching and external sentiment variables, we compare each model in a series of out-of-sample tests. So far all models were estimated full sample, but for forecasting purposes we exclude the last 2.5 years and re-estimate the models in-sample. The remaining data, starting on January 2011, contains 650 observations and includes the Euro zone debt problems as a possible high volatility environment. Surely that particular summer of 2011 affected some sectors more than others, with Financial stocks having suffered the most. A comparison of test statistics across all sectors and models is shown in Table 3 in the appendix.

We test each model’s forecast ability on two fronts. Return prediction is tested with the mean squared error (MSE), mean absolute error (MAE) and hit ratio (HR). The latter statistic
CHAPTER 6. RESULTS

Figure 6.2: All three panels contain outcomes of the MS-EGARCHX model driven by Industrial specific news sentiment. The shaded areas indicate time periods in which the ex ante probability of being in the high volatility regime was greater than 0.5. The ex ante probability itself is shown in the bottom panel, together with its smoothed counterpart (the thick red line). The model’s conditional volatility is shown in the middle. The returns of MSCI Industrials can be seen in the top panel of Figure 6.1.
CHAPTER 6. RESULTS

indicates how often our directional prediction matches the direction of the true market movement. Sadly the first two statistics are very comparable among each model and the hit ratio is below 50% for all five (see Table 6.4). This means that adding sentiment and regime switching capabilities does not help in forecasting returns, even though we found significant sentiment regressors in the mean equation of high volatility periods. This result however does not come as a surprise since most of the extra model complexity focuses on forecasting time-varying volatility rather than returns.

Therefore, the second set of backtests comprise Value at Risk (VaR) statistics. In computing the 1% and 5% forecast Value at Risk, we suppose that the standardized residuals of our models are Student-\(t\) distributed to allow for fat tails. As the variance of a Student-\(t\) distribution is not equal to one but given by \(\frac{\text{df}}{(1 - \text{df})}\), we need to scale the VaR quantiles with the estimated scale parameter \(s\) (Danielsson, 2011). The location parameter \(m\) is set to zero. Two hypotheses should not be violated. The unconditional coverage hypothesis firstly tests whether the probability of an ex-post loss exceeding the VaR forecast matches the theoretical coverage rate of 1% or 5%. If this is rejected, the model is either systematically under- or overestimating risk. Second, the independence hypothesis tests whether VaR exceedances are independently distributed rather than clustered. Both can be tested simultaneously by regressing the following hit series on a constant and its own lag:

\[
\text{Hit}_t(\alpha) = \begin{cases} 
1 - \alpha & \text{if } r_t < -VaR_{t\mid t-1}(\alpha), \\
-\alpha & \text{otherwise.}
\end{cases}
\]  

(6.1)

The test was proposed by Engle and Manganelli (2004). A graph of the exceedences and VaR forecast values is added to the appendix. Judging from the bottom rows of Table 6.4 the single regime models highly reject the independence hypothesis at the 1% VaR level. This hypothesis is not rejected at the 5% significance level for the three Markov switching models, indicating that they provide better risk forecasts. Note that there are some other minor rejections at the 10% significance level for the 1% VaR tests, but all seem to be resolved when considering 5% VaR forecasts. Across the other sectors the dual regime models also perform better. An overview is provided in Table 3 in the appendix.

6.4 Social media: model results

We repeat the analysis of the last section for the social media TRMI. Like we did with the news data, the preliminary test results have indicated which social media variables to include in the MS models. Again we only show outcomes for the Industrials sector to make it comparable to the professional news data. For results across all sectors, we refer to the appendix.

The results are shown in Table 6.5. No sentiment regressors were included in the mean
Table 6.4: Out-of-sample test results for news sentiment for the last 2.5 years of Industrial excess returns. All models have poor return forecast capabilities, but the Value at Risk predictions perform better for the Markov switching models compared to the single regime models. The unconditional coverage rate is tested by $b_0$, while the independence hypothesis is tested by $b_1$. The lack of data gives rise to issues we did not experience when analyzing the longer news sample. First, some estimates are quite odd. For example the $\beta_i$ values are on the low side, while the volatility constants $\omega_i$ have become more extreme. In the Information Technology (IT) sector the difference between $\omega_1$ and $\omega_2$ even grows so large that the MS-EGARCHX model basically behaves like a MS model with constant variances. Considering the large standard errors, we should be careful in interpreting these estimates. The insignificance of t-statistics makes it hard to infer the true parameters; many variables are not significantly different from zero.

Second, for some sectors the estimators converge to a different optimum depending on the starting values used. A couple of these (local) maximums have a very low transition probability $P$ or $Q$ essentially making one regime redundant. In these cases the model wants to fit one
regime to the data, and uses the second to capture an outlier. This is a well-known problem among mixture GARCH models caused by singularities in the likelihood function. Broda et al. (2013) note that this stability problem can make numerical estimation very challenging and propose a method to avoid these so-called degenerated mixture estimates. In this thesis we simply circumvent odd degenerated outcomes by plugging in different starting values until we find the global maximum.

Thus, the social media sample is too short to properly estimate the sophisticated Markov switching models. On top of that, the sample starting late 2006 is largely dominated by financial crises. We see this in the estimates for the Financials sector, which assumes the high volatility regime for the entire period between the housing bubble and early 2013 (appendix, Graph

<table>
<thead>
<tr>
<th>Parameter</th>
<th>EGARCHX</th>
<th>MS-EGARCHX</th>
<th>MS-EGARCHX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std. error</td>
<td>Estimate</td>
</tr>
<tr>
<td>$\phi_{0,1}$</td>
<td>0.0125 (0.0100)</td>
<td>0.0176 (0.0095)</td>
<td>0.0173 (0.0095)</td>
</tr>
<tr>
<td>$\phi_{0,2}$</td>
<td>-</td>
<td>-</td>
<td>-0.1601 (0.0514)**</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>-0.0136 (0.0100)</td>
<td>-0.7559 (0.2099)**</td>
<td>-0.6579 (0.2543)**</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.0302 (0.0166)</td>
<td>-0.0222 (0.0303)</td>
<td>-0.0221 (0.0284)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.9887 (0.0065)**</td>
<td>0.6895 (0.0943)**</td>
<td>0.7333 (0.1148)**</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.1121 (0.0375)**</td>
<td>0.1281 (0.0492)**</td>
<td>0.1207 (0.0471)*</td>
</tr>
<tr>
<td>FEAR 4W</td>
<td>0.0328 (0.0121)**</td>
<td>-</td>
<td>-0.0020 (0.0330)</td>
</tr>
<tr>
<td>LAYOFFS 1W</td>
<td>0.0210 (0.0157)</td>
<td>-</td>
<td>0.0154 (0.0261)</td>
</tr>
<tr>
<td>$\omega_2$</td>
<td>-</td>
<td>0.1382 (0.1098)</td>
<td>0.1366 (0.1146)</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-</td>
<td>0.0930 (0.0687)</td>
<td>0.0863 (0.0696)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-</td>
<td>0.6658 (0.1043)**</td>
<td>0.6877 (0.1081)**</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>-</td>
<td>-0.0358 (0.0919)</td>
<td>-0.0252 (0.0941)</td>
</tr>
<tr>
<td>FEAR 4W</td>
<td>-</td>
<td>-</td>
<td>-0.0113 (0.0437)</td>
</tr>
<tr>
<td>LAYOFFS 1W</td>
<td>-</td>
<td>-</td>
<td>0.0239 (0.0451)</td>
</tr>
<tr>
<td>$P$</td>
<td>-</td>
<td>0.9965 (0.0019)**</td>
<td>0.9970 (0.0018)**</td>
</tr>
<tr>
<td>$1 - Q$</td>
<td>-</td>
<td>0.0203 (0.0102)*</td>
<td>0.0202 (0.0103).</td>
</tr>
</tbody>
</table>

Table 6.5: Comparison of three models for Industrial specific social media sentiment. The MS-EGARCH model does not include any sentiment covariates. As the sample only contains 1,804 observations, the data sample is not long enough to accurately estimate all parameters of the Markov switching models. Note the high standard errors and for example the odd $\beta_i$ values.
3). Simply said there are too few observations to properly estimate the low volatility regime’s parameters as there was no such a regime for Financial equity.

Looking at the LR test statistics among models and sectors in Table 1 in the appendix, we often find that the null hypothesis of sentiment not having any added value is not rejected. The Financials sector is actually the only sector for which the forecast performance of both the single and dual regime model improves by adding social media sentiment. Of four industries we already knew that they are not social media sentiment driven from the preliminary analysis: Energy, Healthcare, Materials and Telecom. Because the Markov switching estimates are quite unreliable for the reasons mentioned before, we do not test the model forecast performance out-of-sample like we did for the news data. Namely, removing the last 2.5 years would leave even fewer observations for a convincing in-sample estimation. Based on the preliminary analysis and doubtful but insignificant parameter estimates of this section, we conclude cautiously that social media sentiment has less forecasting power than sentiment derived from news texts.
Chapter 7

Discussion

Throughout this paper we have occasionally suggested directions for future research. This chapter summarizes those and adds an extensive discussion on what more can be investigated. As this field of study is still relatively new, there is a lot we could not cover. The limitations we encountered, or the future research recommendations we would like to make, can be broken down into three categories: questions on the data, remarks on the model, and suggestions for research that was outside the scope of this thesis.

We start with a discussion of the data used. The sentiment data collection process is fully covered in Chapter 4. It describes how the TRMI as a source cover two of the four types of internet sentiment. We find that sentiment of both these types can have predictive power on return and volatility, but that professional news does a better job than semi-professional social media content. It would be interesting to see whether non-equity related content on retail consumption or general well-being correlates with sector returns. For example, are Information Technology returns predictable with online talks on new high-tech products? Or are web discussions on new drugs informative for future Healthcare stocks? Besides this, the TRMI do not retrieve novelty information from the data (Section 4.2). One could imagine that only the first messages matter, whereafter the ‘event’ is priced in.

Another key discussion point on the data concerns the demographics of web users and the scoring algorithms. We noted in Section 4.4 that the linguistics are quite a black box to us. We do not know how the algorithms are adapted to the demographics of internet users. What is the age, nationality, wealth, influence, or investor experience of an average news or social media user? Is this particular community representative for actual financial market participants? We have no insights into these points as we took the easy-to-use structured data for granted. A researcher might want to have more control by extracting information from the unstructured messages and underlying metadata himself. However as talks vary per platform, per user and over time, this could be a very time consuming process.

The dynamics of the data also make one doubt the appropriateness of models that assume
constant relations. We allowed for some flexibility by including a Markov switching component, however it is very unlikely that the influence of internet news and social media on financial markets has remained the same over a 15 year timespan. The ways digital media are written, distributed and read have undoubtedly changed over time. It would be surprising if a significant variable from this sample continues to contain the same predictive power in the future. Estimating the single regime EGARCHX model on subsamples of the data defined by the structural breaks (Section 5.2.1) already indicates that some variables are more significant in one period than another. We see that more recent periods are best predictable, which could be the consequence of either increased data quality or the large role of digital media during the financial crisis. Future research could apply subsample or rolling window estimation to assess how relations change over time.

A major drawback of the MS-EGARCHX model is its complexity. Many observations are needed for estimation of the up to 20 parameters. This turned out to be problematic for the shorter social media sample which contains only 1,804 data points (Section 5.2). The overparameterization restricts us in three more ways. In Section 3.2 we briefly discussed how we could make the regime switching dependent on sentiment through the transition probabilities like in Gray (1996) or Liu et al. (2012). Secondly, we have noted in Section 4.3 that we could control for macro-variables like GDP, unemployment rates, or United States country sentiment. Lastly, the complexity of the model requires us to simplify estimation by using a normal density in the log-likelihood, even though we know returns to have fatter tails in reality. Estimating the degrees of freedom of a Student-$t$ would yield another parameter. Aggregation to a weekly or monthly level is not possible as we would have even fewer observations. Other models might therefore suit the data better.

Lastly, we focused on the effect of sentiment on equity sectors. Future research could also investigate sentiment impact on regions, asset classes (real estate, commodities, credits) and individual companies. It would also be interesting to predict or replicate macro-economic variables like gross domestic products (GDP), purchase managers indices (PMI), or consumer confidence indices. These are now published on a weekly or monthly basis, but could theoretically be converted to daily or even minutely indices by using web sentiment. Or, one could use sentiment to strengthen well-known investment anomalies like momentum, winners/losers, book-to-market, and earnings strategies. Stambaugh et al. (2012) demonstrated for example that anomalies are more profitable following high levels of sentiment. Another practical application would be to construct a strategy that trades on the significant sentiment variables. We could buy or short the VIX volatility index based on significant news optimism in a high volatility environment.
Chapter 8

Conclusion

This paper adds exogenous variables to existing regime switching models to determine the predictive capabilities of investor sentiment on various equity markets. The Thomson Reuters MarketPsych Indices (TMRI) are 24 emotions and so-called buzz metrics which are derived from professional newswires as well as investment-related social media content. The data are unique considering their coverage, long history, and advanced linguistic algorithms. They furthermore allow us to compare the two data types resulting in the conclusion that news sentiment is an overall better indicator of future market returns and volatility than similar sentiment from social media.

This is backed up by the results of a simple single-regime preliminary analysis, but also by the outcomes of more advanced Markov switching models. The latter allow for the existence of low and high volatility regimes so that we can review whether sentiment plays a more important role in stressed markets compared to calm periods. The underlying thought is that a crisis is driven more by emotions rather than rationality. The hypothesis holds for the Industrials news data where we see that buzz on the direction of the market (marketforecast) is indeed a significant predictor of returns during high volatility periods. Across sectors there is only weak evidence for this theory. For volatility there is no apparent difference in sentiment’s predictive power between regimes. However, we do find that the EGARCH leverage effect is more significant during turbulent markets.

Aside from the differences between regimes, we demonstrate big differences in sentiment sensitivity across sectors. Among the ten industries, the TRMI were by far the most predictive for Financial equity. Probably this is caused by increased media reporting on recent financial crises like the Lehman and Euro crisis. On the other hand, there are sectors that are scarcely affected by sentiment, if at all. Likelihood ratio (LR) tests show that Energy and Information Technology are good examples hereof. Industry-wide we find that volatility is better predictable than returns. This is in line with our initial hypothesis.

A critical side note we have to make is that the social media estimates suffer from high
standard errors. The shorter sample that starts in 2006 rather than 1998 (as for the news data), makes the outcomes of the sophisticated dual regime models unreliable. Although two states can provide additional insights over a single-regime model, proper MS-EGARCHX model estimation requires a lot of observations. Future research could therefore focus on applying a different technique to maximize the forecast potential of social media sentiment.
Bibliography


Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. *1st Workshop on Social Media Analytics (SOMA)*.


Figure 1: Forecast Value at Risk results for the MS-EGARCHX model applied to the Industrial sector. Here sentiment is derived from news content. The dashed red line indicates the 5% VaR predictions, whereas the thicker solid line represents a 1% coverage ratio. According to Table 6.4 the unconditional coverage hypothesis is slightly rejected for the 1% VaR forecasts (at the 10% significance level).
### News data: parameters and model log-likelihoods

<table>
<thead>
<tr>
<th>nr. X mean</th>
<th>CD</th>
<th>CS</th>
<th>E1</th>
<th>FN</th>
<th>HC</th>
<th>ID</th>
<th>IT</th>
<th>M1</th>
<th>T1</th>
<th>U1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>EGARCH</td>
<td>-3005.32</td>
<td>-4308.37</td>
<td>-5965.11</td>
<td>-4191.95</td>
<td>-4249.45</td>
<td>-2826.78</td>
<td>-4608.69</td>
<td>-5122.62</td>
<td>-5415.92</td>
<td>-5339.65</td>
</tr>
<tr>
<td>MS-constant var.</td>
<td>-3069.58</td>
<td>-4507.80</td>
<td>-6109.35</td>
<td>-4555.42</td>
<td>-4316.52</td>
<td>-2823.53</td>
<td>-4847.88</td>
<td>-5141.30</td>
<td>-5377.31</td>
<td>-5390.82</td>
</tr>
<tr>
<td>MS-EGARCH</td>
<td>-2989.14</td>
<td>-4289.17</td>
<td>-5939.42</td>
<td>-4152.87</td>
<td>-4223.23</td>
<td>-2790.35</td>
<td>-4694.10</td>
<td>-5079.90</td>
<td>-5398.37</td>
<td>-5314.49</td>
</tr>
<tr>
<td>MS-EGARCHX</td>
<td>-2977.38</td>
<td>-4278.67</td>
<td>-5936.81</td>
<td>-4128.37</td>
<td>-4218.68</td>
<td>-2764.76</td>
<td>-4649.10</td>
<td>-5065.87</td>
<td>-5383.20</td>
<td>-5309.50</td>
</tr>
</tbody>
</table>

Likelihood ratio (LR) tests

<table>
<thead>
<tr>
<th></th>
<th>EGARCHX-EGARCH</th>
<th>MSEGARCH-EGARCHX</th>
<th>MSEGARCHX-MSEGARCHX</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGARCHX-EGARCH</td>
<td>29***</td>
<td>11.36**</td>
<td>10.95***</td>
</tr>
<tr>
<td>MSEGARCH-EGARCHX</td>
<td>32.38***</td>
<td>38.42***</td>
<td>51.38***</td>
</tr>
<tr>
<td>MSEGARCHX-MSEGARCHX</td>
<td>23.51**</td>
<td>20.99***</td>
<td>5.22</td>
</tr>
</tbody>
</table>

### Social media: parameters and model log-likelihoods

<table>
<thead>
<tr>
<th>nr. X mean</th>
<th>CD</th>
<th>CS</th>
<th>E1</th>
<th>FN</th>
<th>HC</th>
<th>ID</th>
<th>IT</th>
<th>M1</th>
<th>T1</th>
<th>U1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>EGARCH</td>
<td>-1131.06</td>
<td>-1708.42</td>
<td>-2219.83</td>
<td>-2127.77</td>
<td>-1538.32</td>
<td>-1063.45</td>
<td>-1382.68</td>
<td>-2029.75</td>
<td>-2156.35</td>
<td>-2086.08</td>
</tr>
<tr>
<td>EGARCHX</td>
<td>-1130.99</td>
<td>-1705.98</td>
<td>-2219.83</td>
<td>-2111.21</td>
<td>-1538.32</td>
<td>-1055.64</td>
<td>-1364.39</td>
<td>-2029.75</td>
<td>-2156.35</td>
<td>-2080.58</td>
</tr>
<tr>
<td>MS-EGARCH</td>
<td>-1116.39</td>
<td>-1691.57</td>
<td>-2204.36</td>
<td>-2094.42</td>
<td>-1519.81</td>
<td>-1041.28</td>
<td>-1339.08</td>
<td>-2005.46</td>
<td>-2140.78</td>
<td>-2078.33</td>
</tr>
<tr>
<td>MS-EGARCHX</td>
<td>-1114.85</td>
<td>-1686.70</td>
<td>-2204.36</td>
<td>-2080.52</td>
<td>-1519.81</td>
<td>-1040.92</td>
<td>-1340.24</td>
<td>-2005.46</td>
<td>-2140.78</td>
<td>-2077.38</td>
</tr>
</tbody>
</table>

Likelihood ratio (LR) tests

<table>
<thead>
<tr>
<th></th>
<th>EGARCHX-EGARCH</th>
<th>MSEGARCH-EGARCHX</th>
<th>MSEGARCHX-MSEGARCHX</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGARCHX-EGARCH</td>
<td>0.14</td>
<td>4.87</td>
<td>0</td>
</tr>
<tr>
<td>MSEGARCH-EGARCHX</td>
<td>29.33***</td>
<td>33.69***</td>
<td>30.94***</td>
</tr>
<tr>
<td>MSEGARCHX-MSEGARCHX</td>
<td>3.08</td>
<td>9.74*</td>
<td>0</td>
</tr>
</tbody>
</table>

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 1 ' ' 1

Table 1: Table containing the log-likelihood values of the five compared models across all sectors. The degrees of freedom used for a likelihood ratio test is the difference in number of regressors between two models. The number of external variables in the mean equation is given by ‘nr. X mean’, where the rows ‘nr. X vol.’ do the same for the volatility specification. In MS models these numbers are multiplied by 2. Judging from the likelihood model fits only, we find that a Markov switching model is always a significant improvement compared to a single regime model. Whether sentiment improves the model on top of that varies per sector and per data type (news being more predictive than social media). As concluded in Section 6.2, we furthermore find that the MS model with constant variances provides a poor fit. However the opposite seems true for the Telecom (T1) sector in which the MS-constant variances model actually does the best job.
Table 2: The table shows which Thomson Reuters MarketPsych Indices (TRMI) were significant in the MS-EGARCHX model for each sector and data source. Overall we find that sentiment is more significant in predicting volatility than returns. Within return forecasting, the sentiment variables perform slightly better in high volatility regimes (H) compared to low volatility periods (L). Due to large standard errors, many of the social media variables are not successfully identified as being different from zero (Section 6.4).

<table>
<thead>
<tr>
<th>Returns</th>
<th>Volatility</th>
<th>Returns</th>
<th>Volatility</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUST 1D</td>
<td>*</td>
<td>URGENCY wknd</td>
<td>-</td>
</tr>
<tr>
<td>MKTRISK 1D</td>
<td>*</td>
<td>UNCERTN 1D</td>
<td>*</td>
</tr>
<tr>
<td>STRESS 1W</td>
<td>**</td>
<td>LAYOFFS 1D</td>
<td>**</td>
</tr>
<tr>
<td>OPTIMSM wknd</td>
<td>-</td>
<td>***</td>
<td>VIOLENC 4W</td>
</tr>
<tr>
<td>GLOOM 1W</td>
<td>-</td>
<td>***</td>
<td>SNTMENT 4W</td>
</tr>
<tr>
<td>UPGDWNG 1W</td>
<td>-</td>
<td>OPTIMSM 4W</td>
<td>*</td>
</tr>
<tr>
<td>URGENCY 1W</td>
<td>*</td>
<td>**</td>
<td>PRICEUP 1W</td>
</tr>
<tr>
<td>TRUST wknd</td>
<td>-</td>
<td>BUZZWGT 1D</td>
<td>***</td>
</tr>
<tr>
<td>MKTFCST 4W</td>
<td>**</td>
<td>OPTIMSM 1D</td>
<td>-</td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td>FEAR 4W</td>
<td>-</td>
</tr>
<tr>
<td>M1</td>
<td>PRICEUP wknd</td>
<td>-</td>
<td>URGENCY 1D</td>
</tr>
<tr>
<td>MKTRISK 1D</td>
<td>**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>OPTIMSM 1D</td>
<td>***</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td>OPTIMSM wknd</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LAYOFFS 1D</td>
<td>**</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>OPTIMSM wknd</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

* Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Figure 2: The three panels contain outcomes of the MS-EGARCHX model driven by Financials specific news sentiment. The shaded areas indicate time periods in which the ex ante probability of being in the high volatility regime was greater than 0.5. The ex ante probability itself is shown in the bottom panel, together with its smoothed counterpart (the thick red line). The model’s conditional volatility is shown in the middle. The ex ante state probability clearly indicated a shift to a high volatility environment at the beginning of what would later be a very long decline in Financial equity prices (top panel).
Table 3: The table shows out-of-sample test statistics for 4 models based on news sentiment data. Return prediction is tested using the mean squared error (MSE) and hit ratio (HR). A second set of backtests are the Engle-Manganelli Value at Risk (VaR) statistics, which test the unconditional coverage and independence hypothesis of VaR violations at both the 1 and 5 percent level. For the VaR quantiles we fit a Student-t distribution to a model’s standardized residuals. The EGARCHX model is the only single-regime model and suffers from more VaR violations than the complicated regime switching models.
Figure 3: The three panels contain outcomes of the MS-EGARCHX model driven by Financials specific social media content. The shaded areas indicate time periods in which the ex ante probability of being in the high volatility regime was greater than 0.5. As almost the entire social media sample coincides with Financial crises, the model basically estimates just one regime. The insufficient number of observations for each period explains the odd estimates and large standard errors found in Section 6.4.