Proceedings Paper:
Qu, Haizhou, Sardelich Nascimento, Marcelo, Qomariyah, Nunung Nurul et al. (1 more author) (2016) Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events. In: Khan, Fahad, Vintar, Špela, Araúz, Pilar León, Faber, Pamela, Frontini, Francesca, Parvizi, Artemis, Simeunović, Larisa Grčić and Unger, Christina, (eds.) LREC 2016 Proceedings. International Conference on Language Resources and Evaluation, 23-28 May 2016 European Language Resources Association (ELRA) , SVN , pp. 57-63.

Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events

Haizhou Qu, Marcelo Sardelich, Nunung Nurul Qomariyah and Dimitar Kazakov
Artificial Intelligence Group, Department of Computer Science, University of York, UK
hq524, msn511, nq516, dimitar.kazakov@york.ac.uk

Abstract

This article describes a novel dataset aiming to provide insight on the relationship between stock market prices and news on social media, such as Twitter. While several financial companies advertise that they use Twitter data in their decision process, it has been hard to demonstrate whether online postings can genuinely affect market prices. By focussing on an extreme financial event that unfolded over several days and had dramatic and lasting consequences we have aimed to provide data for a case study that could address this question. The dataset contains the stock market price of Volkswagen, Ford and the S&P500 index for the period immediately preceding and following the discovery that Volkswagen had found a way to manipulate in its favour the results of pollution tests for their diesel engines. We also include a large number of relevant tweets from this period alongside key phrases extracted from each message with the intention of providing material for subsequent sentiment analysis. All data is represented as a ontology in order to facilitate its handling, and to allow the integration of other relevant information, such as the link between a subsidiary company and its holding or the names of senior management and their links to other companies.

Keywords: Financial forecasting, stock prices, Twitter, ontology

1. Introduction

On 18 Sep 2015, Volkswagen, one of the world's largest and best known automakers, was named by the US Environment Protection Agency (EPA) as being in breach of its regulations concerning the amount of pollution from diesel engines. Volkswagen had manipulated the outcomes of a vehicle emission test by detecting the specific conditions under which the test took place, and adjusting the performance of its diesel engines in other to meet the required pollution targets, while the same vehicle might fail those targets by a vast margin in actual driving conditions.

Several recent models, including Golf, Polo and the Passat equipped with certain diesel engines were confirmed to contain cheating software that would reduce harmful emissions. The revelation led to a fall of more than 30% of the VW stock price in a single day, which continued to fall in the following weeks, as news was gradually released about the number and seniority levels of people who had knowledge of the deception, until the company CEO himself decided to resign and apologise. There was a prolongued period of uncertainty regarding the spread of this deception across the different continents, and the prices continued to tumble as it became clear that it was not limited to the US market. In addition, subsidiary brands, such as Audi, Seat and Škoda soon revealed the existence of similar practices, with a corresponding effect on their own sales figures and share prices.

We observed these events and collected relevant tweets for the period 15–30 Sep 2015, as well as the minute by minute intra-day stock market prices for Volkwagen (stock symbol $VW, as traded on the Frankfurt Stock Exchange), Ford ($F, NYSE) as an example of an automotive company with no links to the scandal, and the same type of data for the S&P500 stock market index (SPY), providing a baseline for comparison with the US economy as a whole. The Twitter data was then enhanced with the addition of extracted key phrases suitable for sentiment analysis, and the entire dataset was stored as an ontology[1].

2. Financial Forecasting

Since the advent of the stock markets, studying and predicting the future of companies and their share price have been the main tasks facing all market participants. It is extremely difficult to achieve an accurate model that remains reliable over time. There is a very famous yet controversial Efficient Market Hypothesis (EMH) (Fama, 1965), which comes in three forms: weak, semi-strong and strong. If the weak form holds true, stock price cannot be predicted using history prices. The semi-strong form of EMH suggests that stock price reveals all publicly available information. The strong form implies that stock prices will always reflect all information including any hidden information, including even insider's information, if the hypothesis holds. Numerous studies show that EMH does not always hold true (Grossman and Stiglitz, 1980; Haugen, 1995; Shleifer, 2000; Shiller, 2003; Butler and Kazakov, 2012). In all cases, attempts to model and forecast the market are based on time series containing the prices of relevant stock along with other relevant information, which often includes indicators of the general state of the market to allow the evaluation of the relative performance of a given company with respect to the general market trends.

3. Mining Twitter

Along with the development of Social Networking, Twitter has became one of the most popular ways for people to publish, share and acquire information. The two characteristics of this service, instantaneity and publicity, make it a good resource for studying the behaviour of large groups of people. Makinf predictions using tweets has proved a popular research topic. Asur and Huberman (2010) used tweet rate

[1] See the data available at `http://j.mp/FinancialEventsOntology`.

time series to forecast movie sales, with the result outperforming the baseline market-based predictor using HSX,[2] the gold standard of this industry. O'Connor et al. (2010) presented a way to use tweets to predict the US presidential polls. The authors concluded that evolution of tweet sentiment is correlated with the results of presidental elections and also with presidential job approval. Tumasjan et al. (2010) used a much smaller dataset of tweets to forecast the result of the 2009 German elections. Eichstaedt et al. (2015) studied the use of sentiment keywords to predict country level heart disease mortality.

Information extraction from social media can be rather challenging, due to the fact that the texts are very short (up to 140 characters only), noisy and written in an informal style, which often contains bad spelling and non-standard abbreviations (Piskorski and Yangarber, 2013).

4. Ontologies For Financial Data

Ontologies are powerful Artificial Intelligence approach to representing structured knowledge. Their use can also facilitate knowledge sharing between software agents or human users (Gruber, 1993). They are often used in text mining to represent domain knowledge, but their use to describe dynamic processes like time series has been much more limited. The use of ontologies has already been considered in the context of Twitter, as well as in the domain of financial news. For instance, Kontopoulos et al. (2013) discuss the benefits of their use when calculating a sentiment score for Twitter data. Mellouli et al. (2010) describe a proposal for an ontology with 31 concepts and 201 attributes for financial headline news. Lupiani-Ruiz et al. (2011) present an ontology based search engine for financial news. Cotfas et al. (2015) have used ontologies to model Twitter sentiments, such as happiness, sadness or affection. Lee and Wu (2015) developed a framework to extract key words from online social messages and update related event ontologies for fast response to unfolding events.

5. The VW Pollution Scandal Dataset

Despite the substantial amout of research on Twitter data in recent years (Bollen et al., 2011; Wolfram, 2010; Zhang et al., 2011; Si et al., 2013), there are very few publicly available datasets for academic research, with some of the previously published datasets becoming unavailable for various reasons. Yang and Leskovec (2011) provide a large Twitter dataset which has 467 million tweets from 20 million users from 1 June to 31 Dec 2009, or 7 months in total, representing an estimated 20–30% of all tweets published during this period. Go et al. (2009) provide a Twitter dataset labelled with sentiment polarity (positive, neutral or negative), and also split into a training set of 1.6 million tweets (0.8 million positive and 0.8 million negative), and a manually selected test set with 182 positive tweets, and 177 negative tweets.

So far, there has not been a publicly available Twitter dataset, which is aligned with company stock prices. We aim to address this gap, with a focus on an extreme financial event, which could prove helpful in revealing the interplay between financial data and news on social media.

[2]Hollywood Stock Exchange

We collected tweets and retweets from 00:00h EDT on 15 Sep 2015 until 23:59h EDT on 30 Sep 2015.[3] In order to retrieve only relevant tweets, we queried the Twitter API using the tags and keywords listed in Table 1.

Table 1: Tags and keywords for the selection of tweets

Tag/keywords	
@vw	#volkswagen
$vow	#volkswagengate
$vlkay	#volkswagencheat
#vw	#volkswagendiesel
#vwgate	#volkswagenscandal
#vwcheat	#dieselgate
#vwdiesel	emission fraud
#vwscandal	emission crisis

One encouraging observation about this dataset is that it contained tweets with relevant information that predated the official EPA announcement that started the VW diesel engine pollution scandal, as shown below.

Published at 2015, September 18, 10:56:35 EDT

> EPA[4] set to make announcement on major automaker $GM $F $TM $FCAU $HMC $NSANY $TSLA $VLKAY $DDAIF $HYMLF http://t.co/02hNHKq9cx

Published at 2015, September 18, 11:47:58 EDT

> .@EPA to make announcement regarding a "major automaker" at 12 noon today. Source says it will involve @VW. No details yet. Stay tuned.

Published at 2015, September 18, 11:51:42 EDT

> Inbox: EPA, California Notify Volkswagen of Clean Air Act Violations

The first and second tweet did not clearly state that Volkswagen was exactly the automaker, the third tweet is the first one with a clear statement which is ahead of EPA official announcement.[5]

A total of 536,705 tweets were extracted. We have chosen the third tweet as a point in time to split the data into the period 'before the news was out', and the one that followed, resulting in 51,921 tweets before11:51:42 on 18 Sep 2015, and 484,784 after that time. Figure 2 shows a histogram of the number of tweets over each 12h period. A brief timeline of relevant events of the Volkswagen scandal according to Kollewe (2015) is listed below:

18 Sep EPA announces that Volkswagen cheated on the vehicle polluiton test. 482,000 VW diesel cars are required to be recalled in the US.

[3]Earlier tweets were also included if they were retweeted during the indicated time interval.

[4]US Environmental Protection Agency

[5]The attentive reader will find it interesting to compare the timing of the EPA announcement with the closing for the weekend of the Frankfurt stock exchange on that Friday.

20 Sep VW orders an external investigation and CEO apologizes to public.

21 Sep Share price drops by 15 billion Euros in minutes after the Frankfurt stock exchange opens.

22 Sep VW admits 11 million cars worldwide fitted with cheating devices. The CEO says he is "endlessly sorry" but will not resign. The US chief, Michael Horn, says the company "totally screwed up".

23 Sep The CEO quits but insists he is "not aware of any wrongdoing on his part". Class-action lawsuits are filed in the US and Canada and criminal investigations are launched by the US Justice Department.

24 Sep Officical confirms that VW vehicles with cheating software were sold across Europe as well. The UK Department for Transport says it will start its own inquiry into car emissions, as VW faces a barrage of legal claims from British car owners.

26 Sep Switzerland bans sales of VW diesel cars.

28 Sep German prosecutors launch an investigation of VW ex-CEO Winterkorn.

30 Sep Almost 1.2 million VW diesel vehicles in the UK are affected by the scandal, more than one in ten diesel cars on Britain's roads.

We have extended the Twitter dataset with a set of key phrases of length 2 that are potentially relevant to sentiment analysis. In this, we followed the approach discussed by Turney (2002). The main idea is to identify syntactic patterns that are considered suitable to matching subjective opinions (as opposed to objective facts). The resulting candidates for such *polarity_keywords* are linked in the database to the tweet from which they were extracted. This approach can be compared to another related approach to opinion extraction from financial news (Ruiz et al., 2012), in which sentiment gazetteers were also used to indicate the news polarity. Here the decision about polarity has not been made, but is left to future users of the data.

To extract the keywords in question, we employed the Stanford Part-Of-Speech (POS) tagger and Tgrep2 tool to extract the tag patterns proposed by Turney (2002), as listed in Table 2. About a third of all messages were annotated with pairs of key words as a result of the above mentioned procedure. In Table 3 we list the 20 most common pairs: on the whole, they appear quite specific and well correlated with the corpus topic.

In addition to the Twitter data, our dataset includes price information on the per-minute basis for Volkswagen (symbol: VOW.DE) shares and those of Ford (symbol: F) as an example of an automaker unaffected by the scandal. In addition, we have included S&P500 data (American Stock Market Index, symbol: SPY) as an indication of the state of the markets as a whole during the period in question. The data, as available from a number of public websites, includes time stamps, the 'open' and 'close' price, as well as the 'high' and 'low' price for the given one minute interval. Figure 1 shows a comparison of Buy-and-hold[6] cumulative returns of those three securities during 15-30 Sep. 2015.

6. Ontology Representation and Sample Queries

The hierarchy of classes representing the dataset is shown in Figures 3. The **Event** class has three properties: *datetime*, *epoch* and *duration*. The *epoch* property is the number of seconds elapsed from 1st January 1970 00:00 UTC, which provides a common timeline between individuals. The *duration* property describes how long an event lasts and in our dataset, we use second as the timing unit. The **Event** class has two subclasses: **Tweet** and **OHLC**[7]. **Tweet** contains all the individuals storing tweets with their properties: *id*, *username*, *url*, *sourceUrl*, *numberOfRetweet* and *polarityKeyword*. **OHLC** contains all the individuals of stock price of specific company or market index. Each of them has the following properties: *high*, *low*, *open*, *close*, *symbol* and *isin*[8] (See Listing 1).

Listing 1: Individuals of **OHLC** and **Tweet**, shown in turtle format.

```
@prefix nsp: <http://example.org/vwevent2015/property/> .
@prefix nss: <http://example.org/vwevent2015/ontology/OHLC/> .
@prefix nst: <http://example.org/vwevent2015/ontology/Tweet/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# An individual of OHLC
nss:f-1442323800 nsp:close "13.84"^^xsd:float ;
    nsp:datetime "2015-09-15T09:30:00-04:00" ;
    nsp:duration "60"^^xsd:xsd:unsignedLong ;
    nsp:epoch "1442323800"^^xsd:unsignedLong ;
    nsp:high "13.86"^^xsd:float ;
    nsp:low "13.79"^^xsd:float ;
    nsp:symbol "F" ;
    nsp:isin "US3453708600" ;
    nsp:open "13.8"^^xsd:float ;
    nsp:return "0.00289855072464"^^xsd:float .

# An individual of Tweet
<http://example.org/vwevent2015/ontology/Tweet/646575192907644928> nsp:
      datetime "2015-09-23T02:41:53-04:00" ;
    nsp:epoch "1442990513"^^xsd:unsignedLong ;
    nsp:id 646575192907644928 ;
    nsp:numberOfRetweet "0"^^xsd:unsignedLong ;
    nsp:polarityKeyword "criminal charges" ;
    nsp:sourceUrl <http://twitter.com/brian_poncelet/status
          /646575192907644928> ;
    nsp:url <http://twitter.com/Brian_Poncelet/status/646575192907644928> ;
    nsp:username "brian_poncelet" .
```

Representing our data as an ontology makes it possible to be queried in a flexible and powerful fashion, allowing its users to link the textual and time series data in a seamless way. Here are some examples of SPARQL queries seeking to extract useful features through the use of both polarity keywords and stock price movements.

Query 1 This SPARQL query will extract the tweets whose time stamp coincides with a drop in the Volkswagen stock price by more than 1%, ranked by *numberOfRetweets*.

The results of this query 1 are shown in listing 3. In order to improve readability, returns only show three decimal places, and *datetime*s are reformatted not to show the year.

[6]Buy-and-hold is a trading strategy, typically for benchmarking purposes, that considers the performance of buying the security and holding it for the whole period of analysis. Cumulative return on day i: $r_i = (price_i - price_{buy})/price_{buy}$.

[7]OHLC stands for open, high, low and close price of stock price during a period of time.

[8]ISIN refers to International Securities Identification Numbers, which provides a unique identification for each security.

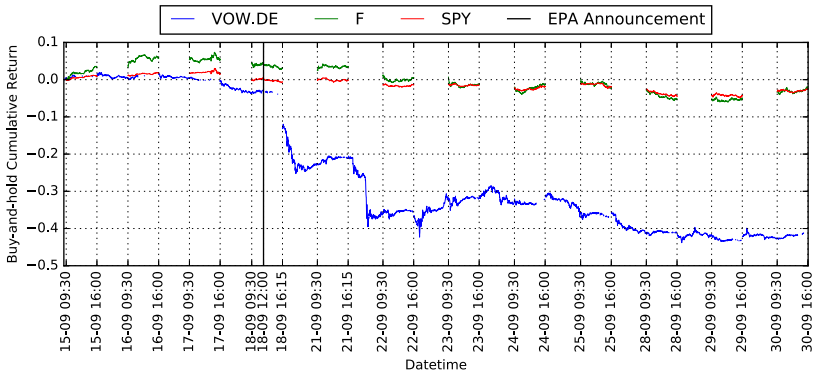

Figure 1: Buy-and-hold cumulative returns of Volkswagen stock, Ford stock and S&P500 during 15-30 September 2015.

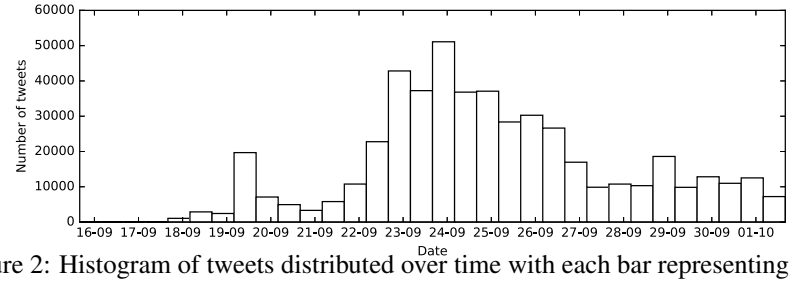

Figure 2: Histogram of tweets distributed over time with each bar representing 12 hours.

Listing 2: Query 1

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT
  ?username
  (?id AS ?tweet_id)
  ?return
  (?numberOfRetweet AS ?nbRt)
  ?datetime
  (group_concat(distinct ?pk;separator=", ") as ?polarityKeywords)
WHERE{
  ?ohlc nsp:epoch ?ohlc_epoch .
  ?ohlc nsp:return ?return .
  ?ohlc nsp:symbol "VOW.DE" .
  FILTER(?return < -0.01)
  ?tweet nsp:epoch ?tweet_epoch .
  ?tweet nsp:datetime ?datetime .
  ?tweet nsp:numberOfRetweet ?numberOfRetweet.
  ?tweet nsp:url ?url .
  ?tweet nsp:sourceUrl ?sourceUrl .
  ?tweet nsp:username ?username .
  ?tweet nsp:id ?id .
  ?tweet nsp:polarityKeyword ?pk .
  FILTER EXISTS{?tweet nsp:polarityKeyword ?pk}
  FILTER(
    ?url = ?sourceUrl
    && xsd:integer(?numberOfRetweet) >= 5
    && xsd:integer(?tweet_epoch) <= xsd:integer(?ohlc_epoch) + 60
    && xsd:integer(?tweet_epoch) >= xsd:integer(?ohlc_epoch)
  )
}
GROUP BY ?username ?id ?return ?numberOfRetweet ?datetime
ORDER BY DESC(xsd:integer(?numberOfRetweet)) ?return
LIMIT 10
```

Listing 3: Result of Query 1

	username	tweet_id	return	nbRt	datetime	polarityKeywords
1	business	646586797636644864	-0.023	113	09-23 03:28:00	as much
2	newswaala	646580334616645633	-0.011	30	09-23 03:02:19	high emissions, first detected
3	twistools_en	646586860173688832	-0.023	8	09-23 03:28:15	national embarrassment
4	nytimesbusiness	646260916129005568	-0.022	6	09-22 05:53:04	diesel cars, little effect
5	speedmonkeycouk	648435351476957184	-0.011	6	09-28 05:53:29	now being

Listing 4: Query 2

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT
  ?pk
  (COUNT(?pk) AS ?count)
WHERE{
  {
    SELECT
      (xsd:unsignedLong(xsd:float(?ohlc_epoch)/60.0) AS ?ohlc_minute)
      (xsd:unsignedLong(xsd:float(?tweet_epoch)/60.0+1.0) AS ?tweet_minute)
      ?pk
      ?return
    WHERE{
      ?ohlc nsp:epoch ?ohlc_epoch ;
        nsp:return ?return ;
        nsp:symbol "VOW.DE" .
      FILTER(?return <= -0.02)
      ?tweet nsp:epoch ?tweet_epoch ;
        nsp:polarityKeyword ?pk ;
    }
    HAVING(?ohlc_minute=?tweet_minute)
  }
}
GROUP BY ?pk
ORDER BY DESC(?count)
LIMIT 20
```

Listing 5: Result of Query 2

	pk	count
1	worldwide fitted	41
2	as much	23
3	entire auto	14
4	sure people	11
5	first detected	10
6	high emissions	10
7	multiple probes	7
8	totally screwed	7
9	chief executive	5
10	diesel cars	5
11	early trading	5
12	here come	5
13	national embarrassment	5
14	little effect	4
15	not sure	4
16	also installed	3
17	false emission	3
18	internal investigations	3
19	just lost	3
20	absolutely foolish	2

Expression	Word1	Word2	followed by
(JJ.(NN|NNS))	JJ	NN or NS	no restrictions
(RB.(JJ!.(NN|NNS)))	RB	JJ	not NN nor NNS
(RBR.(JJ!.(NN|NNS)))	RBR	JJ	not NN nor NNS
(RBS.(JJ!.(NN|NNS)))	RBS	JJ	not NN nor NNS
(JJ.(JJ!.(NN|NNS)))	JJ	JJ	not NN nor NNS
(NN.(JJ!.(NN|NNS)))	NN	JJ	not NN nor NNS
(NS.(JJ!.(NN|NNS)))	NS	JJ	not NN nor NNS
(RB.(VB|VBD|VBN|VBG))	RB	VB, VBD, VBN or VBG	no restrictions
(RBR.(VB|VBD|VBN|VBG))	RBR	VB, VBD, VBN or VBG	no restrictions
(RBS.(VB|VBD|VBN|VBG))	RBS	VB, VBD, VBN or VBG	no restrictions

Table 2: Extracted Word1+Word2 keyphrases using *Tgrep2* expressions

keywords	count		
diesel scandal	3993	diesel deception	1294
chief executive	3835	multiple probes	1189
diesel emissions	3280	electric car	1166
diesel cars	2980	new tech	1110
sure people	2801	clean diesel	1059
new boss	2407	criminal probe	1037
totally screwed	2208	finally be	953
clean air	1919	fresh start	908
as many	1449	refit cars	898
criminal charges	1323	diesel vehicles	890

Table 3: 20 most common pairs of keywords extracted from the Twitter data.

Figure 3: VW Event Ontology Classes

The first tweet was published by Bloomberg (@business):

> CEO Martin Winterkorn faces a showdown with #Volkswagen's board later http://bloom.bg/1FdA4sA

Tweet No. 4 came from Business news of NY Times (@nytimesbusiness):

> Volkswagen's recall troubles may have little effect on China: It sells almost no diesel cars in the country. http://nyti.ms/1Jmipd8

Apart from main public media accounts, we found that among the authors of those tweets are also an indian media (No. 2), a marketing account (No. 3), a motor amateur (No. 5). This indicates our dataset contains information from a range of sources that provide potentially useful information on this event.

Query 2 We have also been able to check whether some of the keywords are associated with specific stock price movements by using the following SPARQL query, which aims to retrieve the keywords associated on drops in Volkswagen price greater than 2% within any one-minute-period.

The result of Query 2 shows that in most cases, the worst drops in VW price coincide with keywords expressing negative sentiment or referring to some of the specific facts of the scandal (e.g. "worldwide fitted", "diesel cars").

Query 3 For users with access to twitter contents (mapped to nsp:content), listing 6 shows the potential usage of connecting with other existing ontologies to combine domain knowledge with stock price time series: *get the average one minute return of stock the surname of a key person* (CEO for example) *appears in the tweets.*

Listing 6: Query 3

```
PREFIX nsp: <http://example.org/vwevent2015/property/>
PREFIX nst: <http://example.org/vwevent2015/ontology/Tweet>
PREFIX nss: <http://example.org/vwevent2015/ontology/OHLC>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX db: <http://dbpedia.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT
  ?sn (AVG(?return) AS ?avgReturn)
WHERE{
  SERVICE <http://dbpedia.org/sparql/>{
    ?company dbo:keyPerson ?person .
    ?person foaf:surname ?surname .
    BIND(LCASE(STR(?surname)) AS ?sn)
    FILTER(?company=<http://dbpedia.org/resource/Volkswagen>)
  }
  ?ohlc nsp:epoch ?ohlc_epoch .
  ?ohlc nsp:return ?return .
  ?ohlc nsp:symbol "VOW.DE" .
  ?tweet nsp:epoch ?tweet_epoch .
  ?tweet nsp:content ?content .
  ?tweet nsp:url ?url .
  ?tweet nsp:sourceUrl ?sourceUrl .
  ?tweet nsp:id ?id .
  ?tweet nsp:numberOfRetweet ?numberOfRetweet
  FILTER(?url=?sourceUrl && ?numberOfRetweet > 100)
  FILTER(CONTAINS(LCASE(?content), ?sn))
  FILTER(
    xsd:integer(?ohlc_epoch) >= xsd:integer(?tweet_epoch) &&
    xsd:integer(?ohlc_epoch) <= xsd:integer(?tweet_epoch) + 60
  )
}
GROUP BY ?sn
```

7. Conclusion and Future Works

With the advantages of ontology representation, discovering useful information in time-labelled text data (tweets) and numerical time series (stock prices) becomes an easier task. Both queries and dataset can be easily modified or extended. On the other hand, copyright issues with Twitter data put limits to displaying and sharing information in a more straightforward way, and restrict us to only displaying tweet IDs in our dataset.

The polarity keywords are a useful feature, despite the unsupervised way in which they were extracted. Our future work will focus on adding to the range of features available in the dataset.

We also want to assess our work in connection with other related ontologies for stock markets[9] (Alonso et al., 2005) and companies[10] as described in DBpedia. Such integration for example should allow one to recognise Volkswagen Group as an entity of Public Company in DBpedia[11], where we can find information about their assets, revenue, owner, holding company, products and many more. This type of information would potentially allow one to automatically link one company affected by adverse events to, say, its subsidiary companies, which one may expect also to feel the repercussions of such events. Indeed, Audi, Seat and Škoda, all subsidiary companies of VW Group, were all eventually linked to the diesel engine cheating software scandal. More recent news from France has shown that any results from our data could also find use to handle other related news from the automotive industry. We hope that our work will encourage more interesting research in the financial domain as a whole.

8. References

Alonso, L., Bas, L., Bellido, S., Contreras, J., Benjamins, R., and Gomez, M. (2005). WP10: Case Study eBanking D10. 7 Financial Ontology. *Data, Information and Process Integration with Semantic Web Services, FP6-507483*.

Asur, S. and Huberman, B. A. (2010). Predicting the Future with Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, volume 1, pages 492–499. IEEE.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1 – 8.

Butler, M. and Kazakov, D. (2012). Testing Implications of the Adaptive Market Hypothesis via Computational Intelligence. In *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2012 IEEE Conference on*, pages 1–8. IEEE.

Cotfas, L.-A., Delcea, C., Roxin, I., and Paun, R., (2015). *New Trends in Intelligent Information and Database Systems*, chapter Twitter Ontology-Driven Sentiment Analysis, pages 131–139. Springer International Publishing, Cham.

[9] http://dbpedia.org/page/Stock_market

[10] http://dbpedia.org/ontology/company

[11] http://dbpedia.org/resource/Volkswagen

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., et al. (2015). Psychological Language on Twitter Predicts County-level Heart Disease Mortality. *Psychological Science*, 26(2):159–169.

Fama, E. F. (1965). The Behavior of Stock-market Prices. *Journal of Business*, 38(1):34–105.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, 1:12.

Grossman, S. J. and Stiglitz, J. E. (1980). On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, pages 393–408.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220.

Haugen, R. A. (1995). *The New Finance: the Case Against efficient markets*. Prentice Hall Englewood Cliffs, NJ.

Kollewe, J. (2015). Volkswagen Emissions Scandal Timeline. http://www.theguardian.com/business/2015/dec/10/volkswagen-emissions-scandal-timeline-events Accessed: Jan. 08, 2016.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based Sentiment Analysis of Twitter Posts. *Expert Systems with Applications*, 40(10):4065–4074.

Lee, C.-H. and Wu, C.-H. (2015). Extracting Entities of Emergent Events from Social Streams Based on a Data-Cluster Slicing Approach for Ontology Engineering. *International Journal of Information Retrieval Research*, 5(3):1–18, July.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., and Camón-Herrero, J. B. (2011). Financial News Semantic Search Engine. *Expert Systems with Applications*, 38(12):15565–15572.

Mellouli, S., Bouslama, F., and Akande, A. (2010). An Ontology for Representing Financial Headline News. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2–3):203–208.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, 11(122-129):1–2.

Piskorski, J. and Yangarber, R. (2013). Information Extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 23–49. Springer.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. (2012). Correlating Financial Time Series with Micro-blogging Activity. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12*, page 513.

Shiller, R. J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives*, pages 83–104.

Shleifer, A. (2000). *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford University Press.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. In *ACL (2)*, pages 24–29.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *ICWSM*, 10:178–185.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wolfram, M. S. A. (2010). *Modelling the Stock Market using Twitter*. Master thesis, The University of Edinburgh.

Yang, J. and Leskovec, J. (2011). Patterns of Temporal Variation in Online Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186. ACM.

Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, 26:55–62.