Financial Text Mining in Twitterland

S.D. Nikolopoulos, I. Santouridis and T. Lazaridis

Introduction

We are living in the "information age", where information is a valuable asset. Information is created by models utilizing data, which most of them are in textual form, and the amount of data in our world has been exploding. The International Data Corporation (IDC), estimated that the total amount of data created and replicated in 2009 was 800 exabytes. Further, they projected that data volume is growing 40% per year, and will grow 44 times between 2009 and 2020 (McKinsey and Company 2011).

Text mining is an emerging research area in accounting and finance that it has many similarities with traditional qualitative analysis (Loughran and Mcdonald 2016). The purpose of text mining is to process, by means of appropriate hardware infrastructure and algorithms, textual data in order to extract meaningful information from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Common terms shared between qualitative analysis and text mining are document summarization, topic modeling, sentiment analysis, etc.

Microblogging is an increasingly popular form of "professional" communication on the web. Twitter is currently one of the best-known microblogging platforms and online social networking service that enables users to send and read short 140 character messages called "tweets". The significant impact Twitter may have on financial markets become apparent on April 23, 2013 when a fake tweet on a

S.D. Nikolopoulos (✉) · I. Santouridis
Department of Accounting and Finance, TEI Thessaly, Larissa, Greece
e-mail: s.nikolopoulos@teilar.gr

T. Lazaridis
Department of Business Administration, TEI of Western Macedonia,
Grevena, Greece

A. Kavoura et al. (eds.), *Strategic Innovative Marketing*, Springer Proceedings
in Business and Economics, DOI 10.1007/978-3-319-56288-9_16

hacked official Twitter account of the Associated Press news agency (@AP account), sent out at about 1:07 p.m. ET, saying "Breaking: Two Explosions in the White House and Barack Obama is Injured." The AP quickly announced it was hacked. However, the market impact was already intense. The Dow Jones Industrial Average plunged more than 140 points and bond yields fell. Within 6 min, the Dow recovered its losses and was trading with triple-digit gains. Reuters estimated that the temporary loss of market capitalization in the S&P 500 alone totaled $136.5 billion (Karppi and Crawford 2015).

Mining microblogging data to model stock market is a very active research topic (Bollen et al. 2011; Groß-Klußmann and Hautsch 2011; Rao and Srivastava 2013; Sprenger et al. 2014; Zhang et al. 2012). In the relevant literature, it is argued that investor sentiment can be used to forecast stock market variables such as prices, returns, volume, trends, etc. For example, several studies have shown that individual's financial decisions are significantly affected by their emotions and mood (Nofsinger 2005; Peterson 2007; Ranco et al. 2015). These findings are in line with recent advances in behavioral and Emotional Finance which provide plausible explanations for market inefficiencies (Fairchild 2012; Malkiel 2003; Raines and Leathers 2011; Tuckett 2011).

Twitter Mining Methodology

The first step in text mining analysis of tweets is to search for relevant tweets and then create a corpus or text database that is a collection of documents/tweets. We search for tweets utilizing the application programming interface (API) provider by the Twitter to obtain a collection of public tweets (Makice 2009). There are various ways or keywords one may use to search information on Twitter. For example, by performing a search by a simple text term (i.e., INTC), or by using hashtag (i.e., #INTC), username (i.e., @INTC), or cashtags (i.e., $INTC). The selection of the search term greatly affect the results. From our experiments in financial text mining, the cashtag is the most appropriate way to search relevant financial information since it can reduce the information noise and the size of the corpus.

The next step is to preprocess the data. Text preprocessing is the process of making clear each language structure and to eliminate as much as possible the language dependent factors (Wang and Wang 2005). We applied standard data cleaning and preprocessing techniques for preparing the Twitter data for the subsequent analysis. That is, removing numbers; converting to lowercase; remove punctuation; removing common words that usually have no analytic value; removing common word endings (e.g., "ing", "es"); stripping white space, to remove white space left over by the previous prepossessing steps; etc. Thus, the tweet:

> RT: $IBM's cloud revenue grew 30% in 2Q, reached $11.6B for the last 12 mos. See what's driving this growth https://t.co/DFn3Tv6h

After the appropriate preprocessing, it is transformed into the following text:

ibm cloud revenue grew reached last what driving this growth

Then, we create mathematical matrices, called document term matrix (dtm) or term document matrix (tdm), describing the frequency of terms/words that exist in a corpus. In a dtm, rows correspond to documents in the collection and columns correspond to terms. These matrices can be used for information retrieval and plotting, clustering and PCA, thematic and sentiment analysis, etc.

Information retrieval is the activity of obtaining relevant information from a corpus. Searches for information on the corpus can be based on full-text or other content-based indexing (Rijsbergen 1979).

Clustering is an unsupervised learning paradigm. Clustering methods try to identify inherent groupings of the text documents so that a set of clusters are produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity (Shawkat Ali and Xiang 2010).

In machine learning and natural language processing, topic models represent a class of computer programs that automatically extracts topics from texts. Topic modeling is a frequently used text-mining tool for discovering hidden semantic structures in a text bod. In machine learning and natural language processing, topic models represent a class of computer programs that automatically extracts topics from texts. Topic modeling is a frequently used text-mining tool for discovering hidden semantic structures in a text body. It does that by exploiting the correlations among the words and latent semantic themes. Latent Dirichlet Allocation (LDA) and "Topic Modeling" are often used synonymously, but LDA is a special case of topic modeling. LDA represents documents as mixtures of topics that spit out words with certain probabilities (Bei et al. 2003).

Finally, one important application of text mining is text sentiment analysis. This technique tries to discover the sentiment or polarity of a written text. This can be used to categorize text documents into a set of predefined sentiment categories (e.g., positive or negative sentiment categories), or it can be used to give the text a grade on a given scale. In essence, it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions, and emotions expressed within an online mention. We use the positive and negative words dictionary created by Hu and Liu (2004) to create two sentiment indexes. The first sentiment index is created as the difference between positive and negative words in each document and the second by the polarity score of the qdap package (Rinker 2016).

Application

In this section, we apply the methodology described in the previous section for the analysis of mined tweets of the Intel Company. We discovered 1.920 relevant tweets for the Intel stock, searching the Twitter database using as a search string the stock quote for Intel corporation common stock $INTC. The 1.920 retrieved tweets used to create a corpus. Then we created a dtm and from that point, we are able to retrieve useful information from the corpus. For example, in Table 1 all terms that appear more than 50 times are presented. In Table 2 the correlation between the term bearish and other terms of the corpus are presented.

Thus, the terms bearish and doji coexist in all documents of the corpus as indicated by the correlation 1. Therefore, in our corpus the term bearish that is

Table 1 Terms of frequency higher than 50

aal	aapl	Alert	amd	amzn	Apple	bac	Bernstein	Big	Cat
cmg	cop	corp	csco	dis	Dividend	dow	Earnings	epd	ete
Global brand	goog	IBM	Intel	Internet	jnj	Market	msft	nasdaq	New
Next	nflx	nvda	Options	pfe	Poised	Price	pru	qcom	Read
Reporting	sbux	Shares	Stock	Stocks	Tech	Things	Top	Trend	Week

Table 2 Relationships/correlation between the term "Bearish" and other terms in corpus

Doji	Technical	ddd	itw	jns	pbi	pff	pfg	pgx	psec	swingtradebot
1	0.820	0.580	0.580	0.580	0.580	0.580	0.580	0.580	0.580	0.580

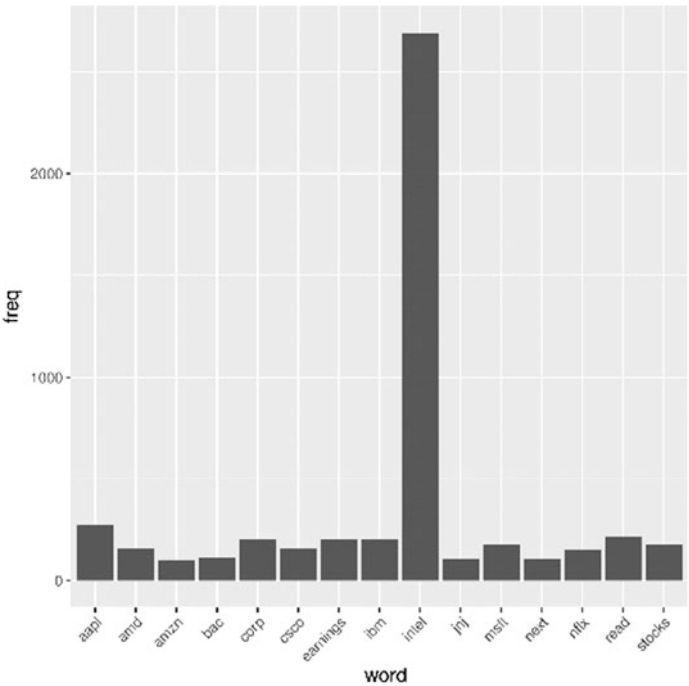


Fig. 1 Words that appear at least 100 times

investors who believe that a stock price will decline is related to companies of NYSE and NASDAQ such as ddd, itw, etc. Further, we can plot words that appear more than specific times or to create a cloud of words (see Figs. 1 and 2).

Then, we create a hierarchical cluster of terms using the ward metric (Fig. 3), and a bivariate cluster plot to visualize partitions of the terms in our corpus (Fig. 4). The Bivariate Clustering Plot (Fig. 4) gives a two-dimensional representation of the objects and the spanning ellipses of the clusters. Note that the boundary of a spanning ellipse always contains several objects. The distance between two clusters is represented as a line connecting the cluster centers. Objects belonging to different clusters are plotted with different characters. At the bottom of both plots, we see that 91.41% of the point variability is explained by the first two principal components.

The next step is to create the topics utilizing the LDA algorithm. The number of topics was identified by trying out different values of topics, to find which value produces the maximum log-likelihood, given the data (Graham and Ackland 2015). Twelve topics were identified but given space constraints of the paper only seven

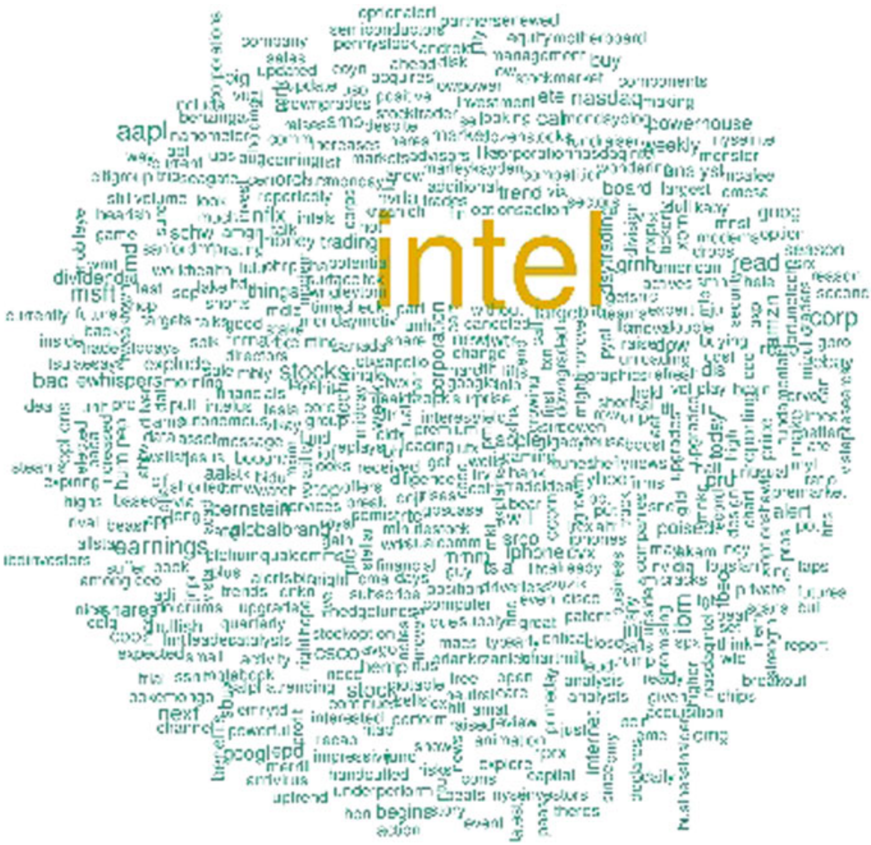


Fig. 2 Cloud plot of 600 most frequent words for Intel stock

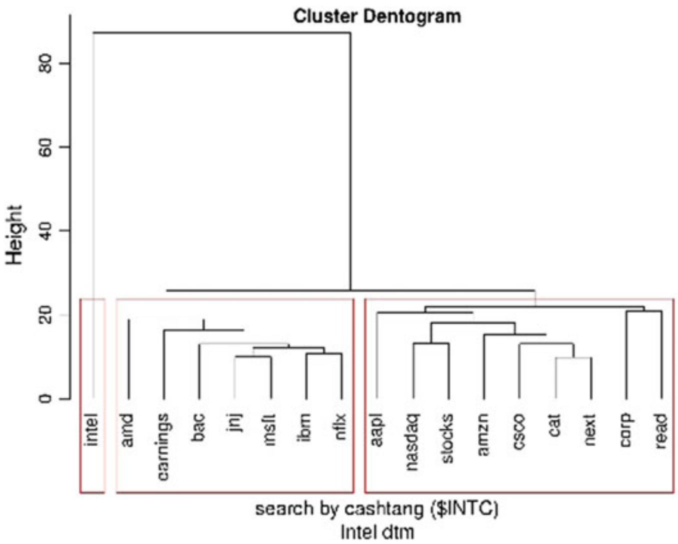


Fig. 3 Hierarchical clustering

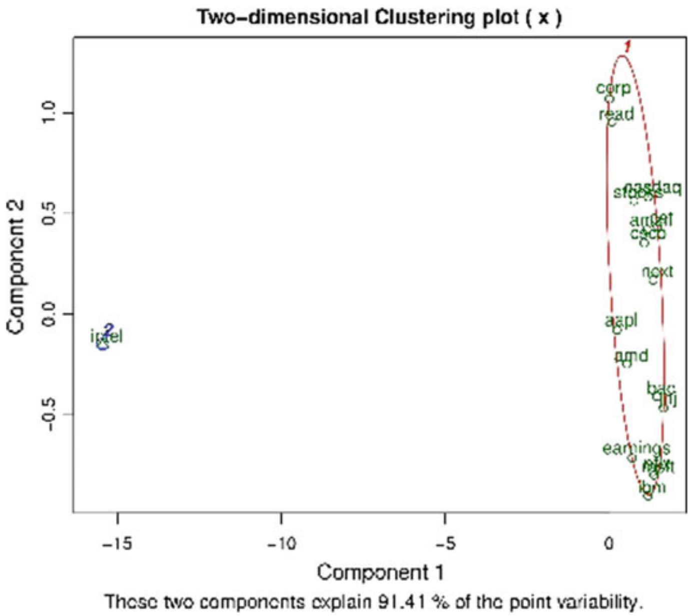


Fig. 4 Bivariate clustering plot

topics are presented in Table 2. Based on those terms, one may perform qualitative analysis but this experiment is left for a future paper (Table 3).

Table 3 Identified topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Holding	Time	Reporting	Short	Reporting	Going	Explore
Shar	Book	Game	Buying	News	Overvalued	Qualcomms
Notebook	Breakout	Replays	Itunes	Optionsaction	Premarket	cons
Fousfan	Boost	Citigroup	Action	Expert	Another	Declares
Taps	Short	Neutral	Beast	Capital	Reiterated	Negative
mlb	Hits	Tweaktown	Thursday	Raised	slw	Reason
Put	Avastavg	Expiring	coyn	Announces	abc	Action
Semiconductor	Current	Highest	Gaming	Stockoption	adbe	Catalysts
Anavex	Fundamental	Nice	Second	Bought	Boosted	Holding
Division	Interested	Resources	Takes	Canceled	Channel	itus

Finally, we created a sentiment index using the QDAP polarity algorithm. We found strong relationship of the sentiment index and the daily returns of the intel

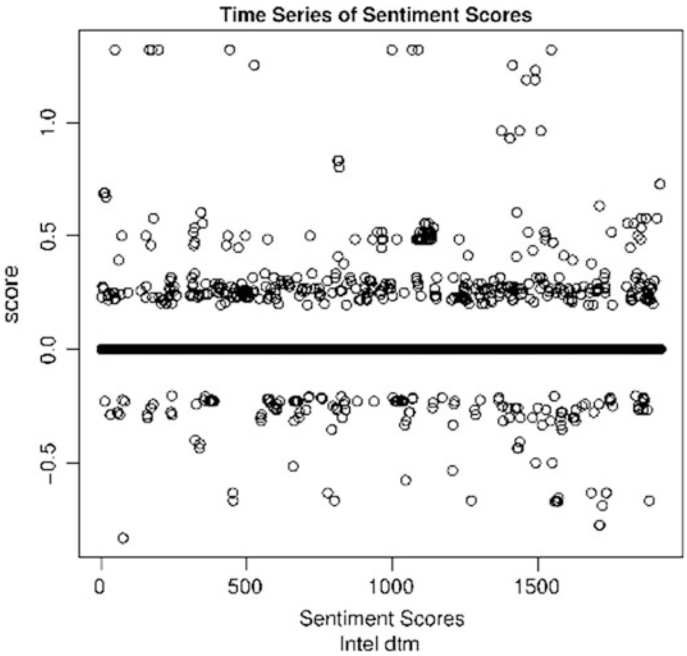
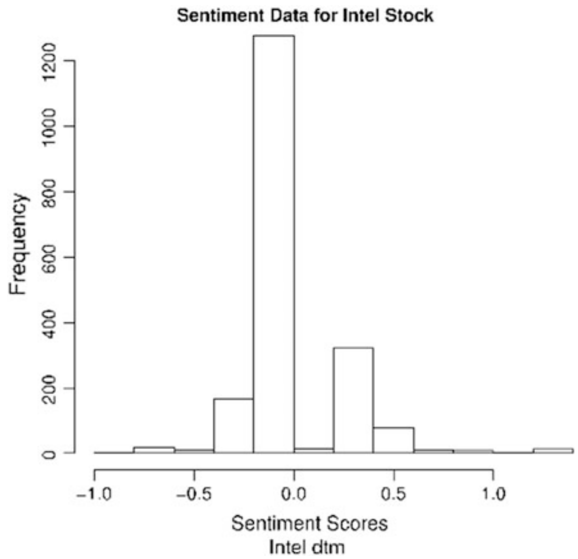


Fig. 5 Time series SI scores per tweet

Fig. 6 Histogram of SI scores



stock. However, the results are not shown here because the sample of the experiment is ten days which is very short to draw statistically firm conclusions. Nevertheless, we have shown how to explore relevant for a problem tweets in order to help the decision-making process. In future work, we will apply the techniques presented here in a number of interesting financial and accounting problems (Figs. 5 and 6).

Conclusions

In this paper, we reviewed the current research of text data mining for financial applications. While the results of these applications are promising there are a lot yet to be asked and many issues should be addressed before this type of research becomes main stream. For example, in most of the studies the time period for forecasting stock variables are very short and the created variables (i.e., sentiment indexes) somewhere subjective. Further, while most of the studies make strong arguments against EMY, none of them estimates abnormal returns for a long period of time. Transaction cost is ignored as well as processing information cost which in this case could be quite high. On the other hand, we found promising use of text mining in event studies in order to better understand the factors that take off equilibrium the system (i.e., in structural brakes). We examined most of the available techniques for financial text mining and presented methodological guidelines for the implementation of those techniques through the application of text mining of tweets regarding IBM stock. The value of this methodology/approach is that it enables the researcher to apply a quantitative and objective methodology in handling unstructured or semi-structured data (e.g., annual reports, audit reports). By doing that, the bias of data selection is minimized and the data that derive from the methodology are more eligible by other researchers.

References

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Bollen, J., H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2 (1): 1–8.

Fairchild, R.J. 2012. From behavioural to emotional corporate finance: A new research direction. *International Journal of Behavioural Accounting and Finance* 3 (3–4): 221–243.

Graham, T., and R. Ackland. 2015. Topic modeling of tweets in R: A tutorial and methodology. https://www.academia.edu/19255535/.

Groß-Klußmann, A., and N. Hautsch. 2011. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance* 18 (2): 321–340.

Hu, M., and B. Liu. 2004. Mining opinion features in customer reviews. In 19th *National conference on artificial intelligence*, 755–760.

Karppi, T., and K. Crawford. 2015. Social media, financial algorithms and the hack crash. *Theory, Culture and Society*.

Loughran, T., and B. Mcdonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*.

Makice, K. 2009. *Twitter API: Up and running: learn how to build applications with the twitter API*. O'Reilly Media, Inc.

Malkiel, B.G. 2003. The efficient market hypothesis and its critics. *Journal of Economic Perspectives* 17 (1): 59–82.

McKinsey & Company. 2011. Big data: The next frontier for innovation, competition, and productivity. Technical Report.

Nofsinger, J.R. 2005. Social mood and financial economics. *The Journal of Behavioral Finance* 6 (3): 144–160.

Peterson, R.L. 2007. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioural Finance* 8 (2): 70–78.

Raines, J., and C. Leathers. 2011. Behavioral finance and post Keynesian—Institutionalist theories of financial markets. *Journal of Post Keynesian Economics* 33 (4): 539–554.

Ranco, G., D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič. 2015. The effects of twitter sentiment on stock price returns. *PloS One* 10 (9). doi:10.1371/journal.pone.0138441.

Rao, T., and S. Srivastava. 2013. Modelling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments. In *Proceedings of the 5th annual ACM web science*.

Rijsbergen, C.J.V. 1979. *Information retrieval. Butterworth-Heinemann*, 2nd ed.

Rinker, T.W. 2016. *qdap: Quantitative discourse analysis package*. http://github.com/trinker/qdap.

Shawkat Ali, A., and Y. Xiang. 2010. *Dynamic and advanced data mining for progressing technological development: Innovations and systemic approaches: Innovations and systemic*. IGI Global.

Sprenger, T.O., A. Tumasjan, P.G. Sandner, and I.M. Welpe. 2014. Tweets and trades: The information content of stock microblogs. *European Financial Management* 20 (5): 926–957.

Tuckett, D. 2011. *Minding the markets: An emotional finance view of financial instability*. Palgrave Macmillan.

Wang, Y.W.Y., and X.J.W. Wang. 2005. A new approach to feature selection in text classification. In *2005 International conference on machine learning and cybernetics*, 18–21, Aug 6.

Zhang, X., H. Fuehres, and P. Gloor. 2012. Predicting asset value through twitter buzz. In *Advances in Collective Intelligence 2011*, 23–34. Springer.