INM430 COURSEWORK REPORT: PREDICTION OF STOCK PRICE VOLATILITY USING NEWS HEADLINES

By Niall Larkin

DECEMBER 16, 2018

CITY, UNIVERSITY LONDON
Northampton Square, Clerkenwell, London EC1V 0HB

1. Data source and domain description

In the finance sector being able to effectively predict stock index volatility or prices with respect to news or other sources of information is an area of constant research within industry [1]. This is due to the fact that being able to predict whether to buy, hold or sell an index when volatility is high provides a competitive advantage to managing equity portfolios. This report reviews how the stock indexes Dow Jones Industrial Average (DJIA), Australian Stock exchange (ASX-200) and Australian Finance sector (ASX-FIN) are correlated to news headlines topics from the Australian Broadcast Company (ABC).

Data set origin:

Data for each indexes are available from the S&P indices website [2, 3, 4]

Only daily values of closing prices can be obtained for long term data trawls. As a result daily values from the 31-Oct-18 to 02-Nov-2018 for each index was downloaded from this website. Each financial data set had the following columns:

1. Date of daily return
2. Indexes daily closing value

The news textual data are available from Kaggle [5]. The data consisted of approximately 200 headlines per day where the data is broken into two separate columns by date and news headline.

2. Analysis Strategy and Plan

1. Prepare the data: deal with missing data and load the data into dataframes.
2. Exploratory Data Analysis: look for auto-correlation and other intrinsic features of both the text and stock time series datasets
3. Add new features and transform the data where required for analysis
4. Perform topic modelling and topic assignment to each document in the text dataset.
5. Perform analytical analysis and visualise the correlation(if any) between the news headlines and stock volatility dataset.
6. Split the merged text and stock volatility dataset into Training/Validation and Test sets
7. Develop a Single Vector Machines (SVM) algorithm to predict stock price volatility based on news headlines.
8. Present the results

3. Problems and Analytical Question:

No.	**Analytical Question**	**Analytical tasks**
1	Why is the historical accuracy of stock prediction models using news headlines so low? [6]	1. Generate an XY plot to see the relationship between news topics and stock volatility 2. Perform X^2 testing between news topics frequency and discretized stock volatility signals
2a	What is the optimum topic model to utilise for short text documents (There are significant performance variability with these models and this text corpus type [7])	Perform topic modelling with text corpus with different topic models LDA and NMF
2b	What are effective evaluation metrics to determine the effectiveness of the topics model selected?	Assess performance of models manually and with coherence metrics and determine if there is correlation
3	What topics within the news have the highest impact on stock price volatility?	Reference task 1

Findings and Reflections

Why is the historical accuracy of these models so poor?

Chi squared tests between the frequencies of each news topic to stock volatility signals found statistical significance for each indices volatility to the different news topic models for each time frame. As a result a novel visualisation of a quadrant x-y plot was generated to better visualise this as explained in the associated jupyter notebook. An excerpt from the final visualisations used to correlate the text corpus with overall stock volatility can be seen in Figure 1 to Figure 5 below. This excerpt specifically comes from analysis performed on DJIA index from 2009(To see the other figures produced by the analysis refer to appendix A).

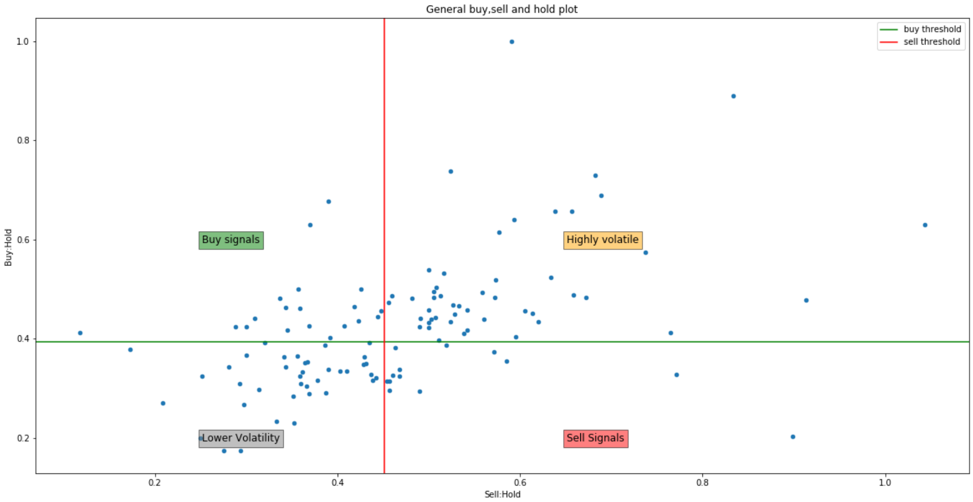


Figure 1 DJIA stock index volatility quadrant chart w.r.t to different news topics. please note each scatter point related to a specific news topic and its associated ratio of buy:hold and sell:hold signals

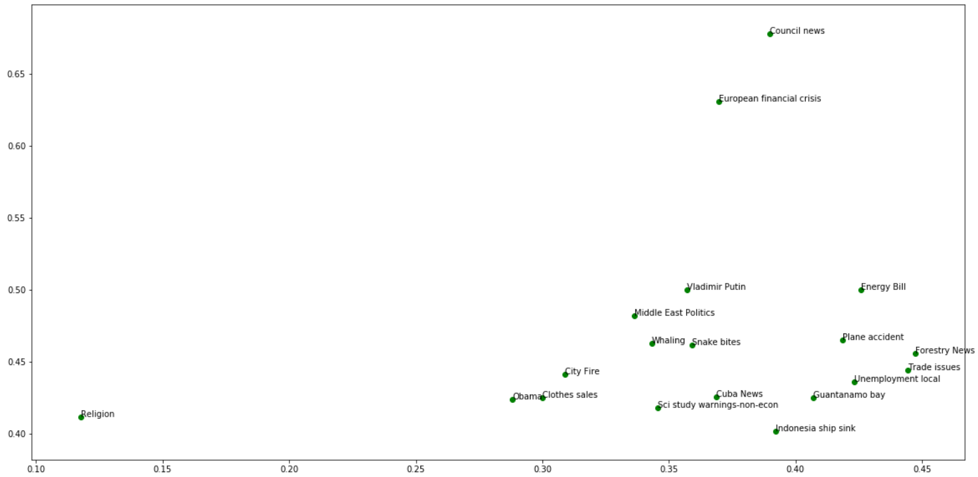


Figure 2 DJIA stock index 2009 Buy signal quadrant with new topics annotated to scatter points

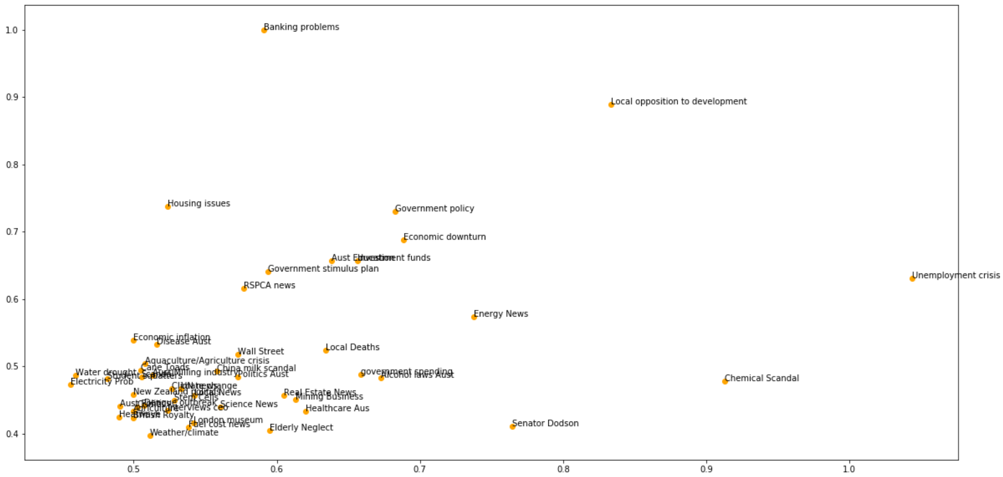


Figure 3 DJIA stock index 2009 high volatility signals quadrant with new topics annotated to scatter points

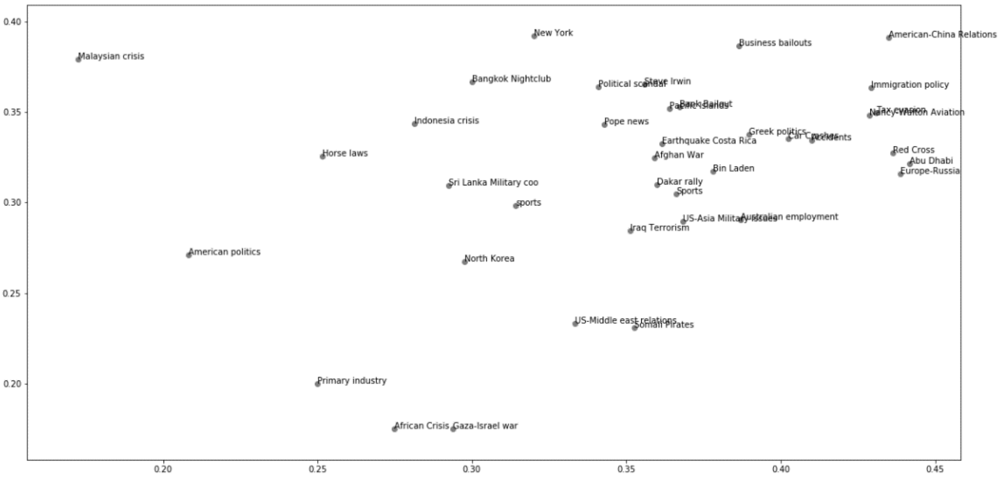


Figure 4 DJIA stock index 2009 Low volatility quadrant with new topics annotated to scatter points

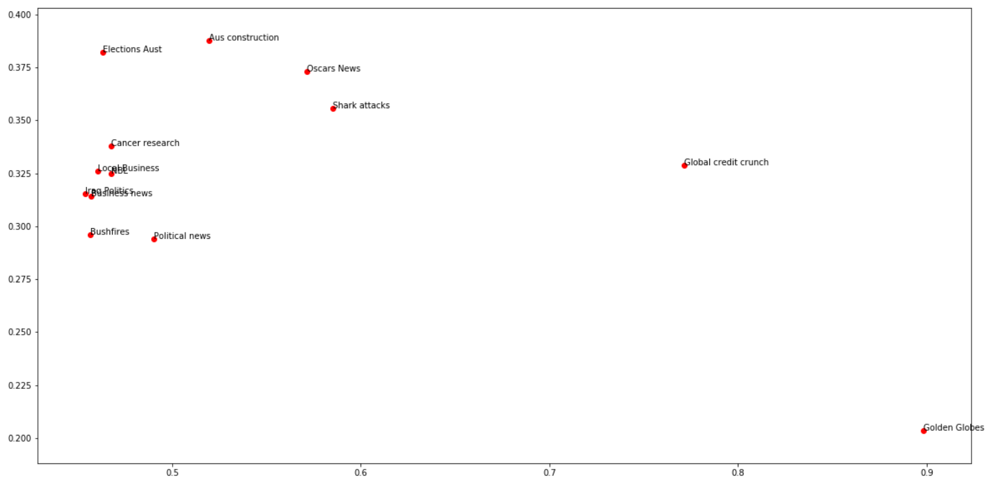


Figure 5 DJIA stock index 2009 Sell signal quadrant with new topics annotated to scatter points

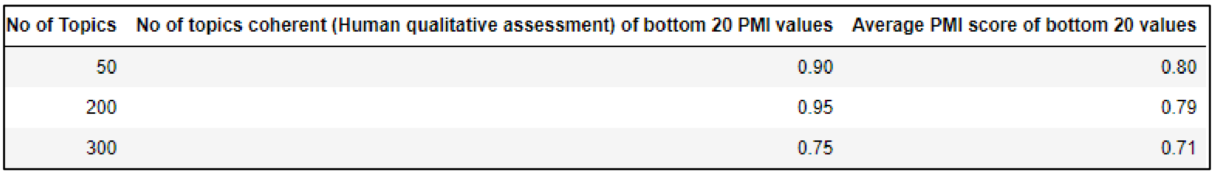
From reviewing the spread of non-financial or economic topics like "the Oscars" and "local news" items are distributed through all stock volatility quadrants instead of the low volatility quadrant exclusively. This is consistent throughout the different indices visualisations present (Reference appendix). With this level of noise of different news topics across the different stock volatility quadrants this explains why there is significant issues with accuracy with respect to utilising news corpus as a means to predicting stock volatility and denotes that significant feature selection from the news corpus is required in order to effectively ascertain effective pertinent topics within the news corpus to carry forward to a stock price prediction model.

Please note though there is one sources of bias that require further investigation specifically to confirm these results:

1. A sampled human qualitative review of the topic assignment approach that was performed using cumulative PMI Score in this project to confirm its accuracy for topic assignment.

What is a good topic model for short text:

From this analysis it was determined that an NMF model using a correlation matrix was superior in terms of generating coherent topics compared to other more established LDA topic models with overall maximum coherence when compared to human interpretation of 95% versus 52%. This result is validated by the fact that overall it produced superior semantic coherence compared to a human qualitative assessment (See Figure 6 and Table 1). Despite how consistently effective this approach is however there is issues with the NMF model.



No of Topics	No of topics coherent (Human qualitative assessment) of bottom 20 PMI values	Average PMI score of bottom 20 values
50	0.90	0.80
200	0.95	0.79
300	0.75	0.71

Table 1 NMF correlation matrix tuning results

Space complexity with the NMF model was a problem with this approach where the text corpus file size that could be processed by this model had to be <1300kB when computational power of 8GB of

RAM was used. The cause of this memory issue was that the size of the initial correlation matrix being the constant source of memory error used in the computation of this method and this appears to be the processing bottleneck.

What are effective evaluation metrics to determine the effectiveness of the topics model selected?

Overall it was found that the performance of the C_v metric which is recorded as having a correlation to human topic coherence of 73% [8] was found to correlate to a degree human coherence qualitative assessment with the LDA model when no data aggregation of the short texts was prevalent. This can be seen with the scatter plots in Figure 6 below where as topic size increase the semantic coherence metric C__v increases as the human quantitative assessment of semantic coherence decreased. However once data aggregation was utilised and the document text corpuses increased there was a significant divergence in terms of in coherency scores between human qualitative and this quantitative assessments (Reference Figure 6).

This is on first glance unusual as from the paper it states that the performance of this analysis typical improved as corpus sized increased [8]. However taking into account the Boolean sliding window element of how this C__v score being set at 10 words and each original topic string only being on average 40 characters long (Reference jupyter notebook for figure) means that it will infer quite disjointed word distributions across the entire topic as a result of the data aggregation despite more inherently coherent topics being generated by the LDA model and being interpreted by human as such as a result. This denotes that ultimately this metric can be used for either short or long text for coherence scores however for aggregated short text documents it is not an effective metric to use in conjunction with human qualitative assessments for semantic coherence.

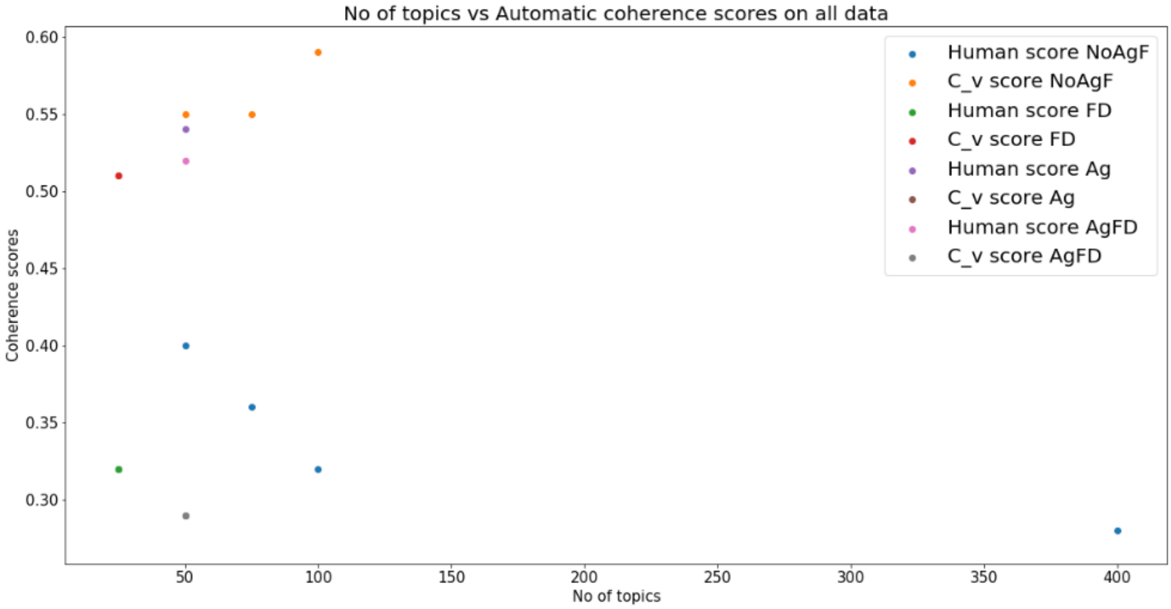


Figure 6 LDA comparison of human topic coherence scores to C_V scores NoAgF denotes data sets with no data aggregation or TDIDF dictionary, FD denotes TD-IDF data sets only, Ag denotes data aggregation only and AgFD denotes both processes in effect.

What topics within the news have the highest impact on stock price volatility?

As per the visualisations above due to the fact that the spread of topics across the entire range of the stock volatility no attempt was made to determine which topics had the highest impact on stock volatility. Despite this the visualisation created is a novel way to estimate whether or not text corpus is an effective source of information with regards to using it for stock volatility prediction.

Conclusion and future work

Ultimately when beginning this task the objective was to understand and characterise how news topics correlate to stock price volatility and from there then to predict stock prices using these different features. From the analysis performed it has been found with this news source it is not feasible to perform this. However in doing this an effective pipeline and set of visualisations to assess short text corpus and how it correlates to stock price volatility has been generated. As a result there a number that are proposed in order to further this approach:

1. Utilise this same NMF model on larger text corpuses or other short text corpuses to view the robustness of its performance on other datasets.
2. Perform a longer comprehensive analysis of the NMF model and topic assignment model against a topic labelled short text data set to determine if equivalent topics are generated ad to review the level of accuracy of this approach.
3. Generate a more sophisticated topic assignment model as opposed to the PMI Frequency Naïve Bayes MLE approach taken to better assign topics to each document.

Appendix A Stock Quadrant volatility charts

2009 Additional figures

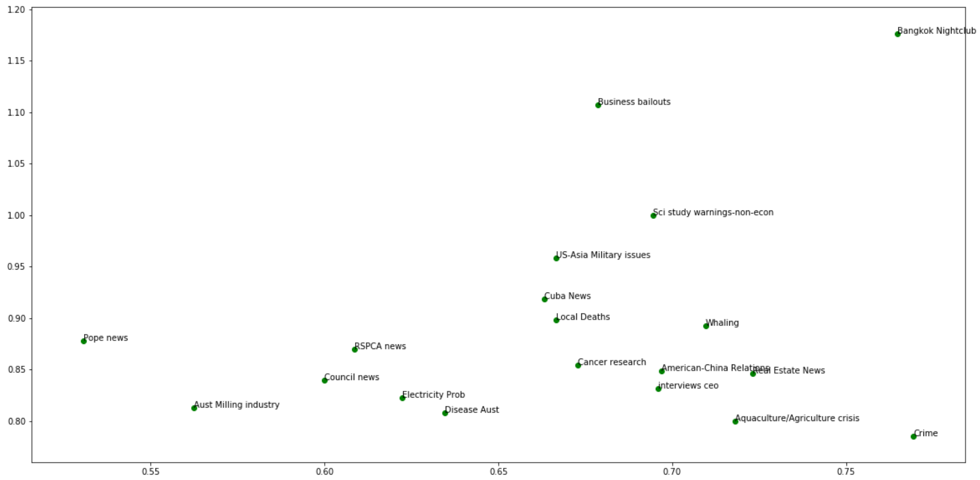


Figure 7 ASX200 buy signal stock volatility quadrant 2009

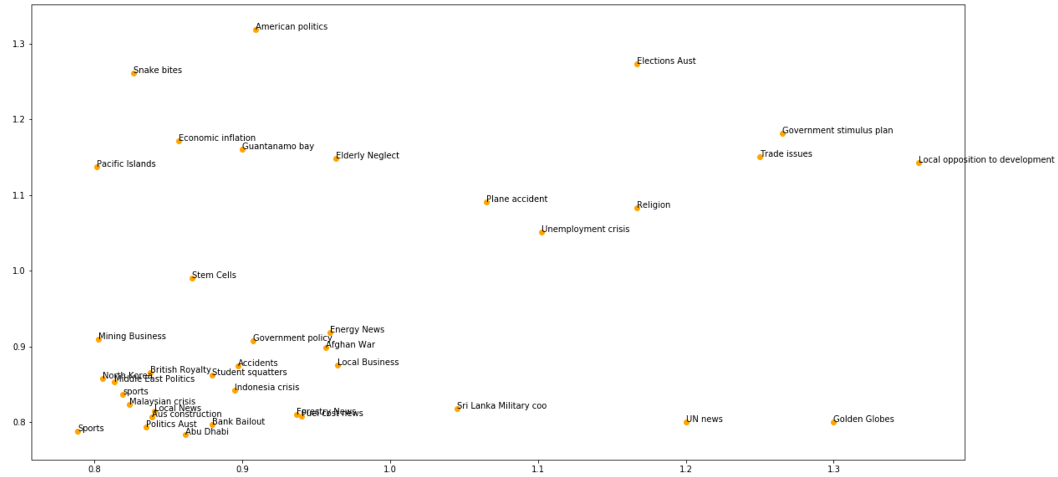


Figure 8 ASX200 high stock volatility quadrant 2009

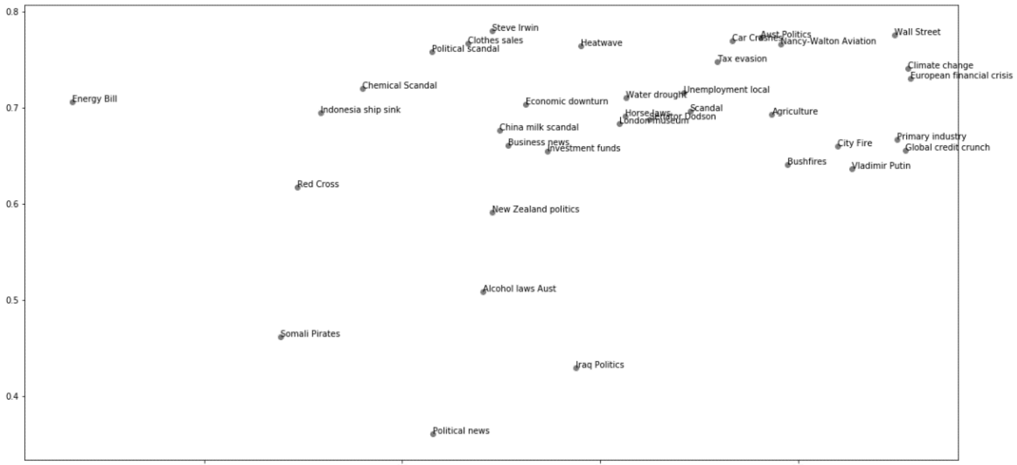


Figure 9 ASX200 low stock volatility quadrant 2009

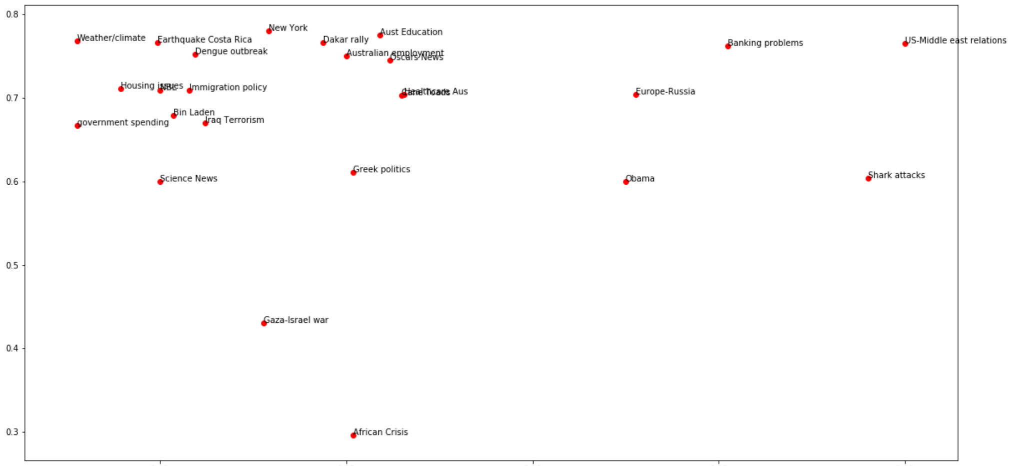


Figure 10 ASX200 sell signal stock volatility quadrant 2009

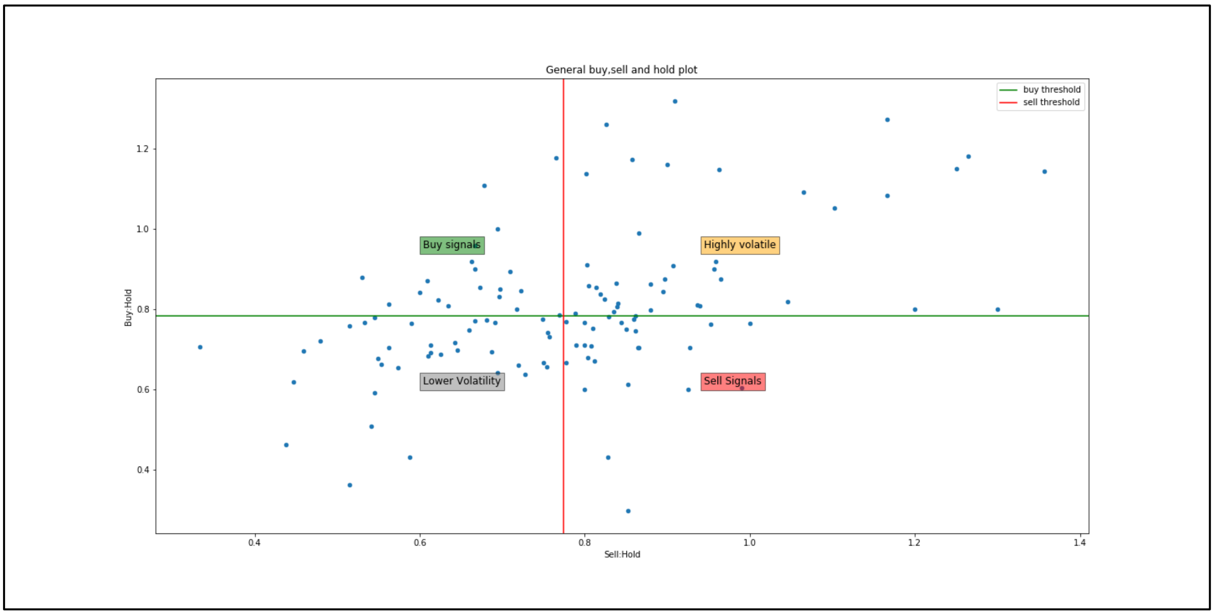


Figure 11 ASX200 stock volatility vs news topics quadrant 2009

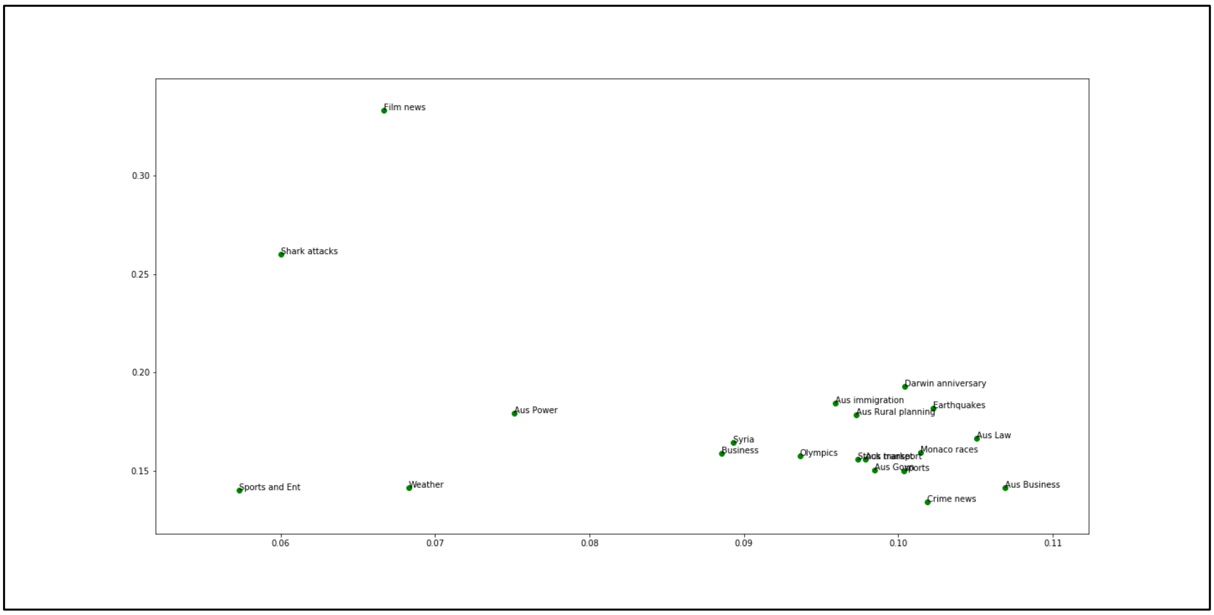


Figure 12 ASX200 Financials buy signal stock volatility quadrant 2009

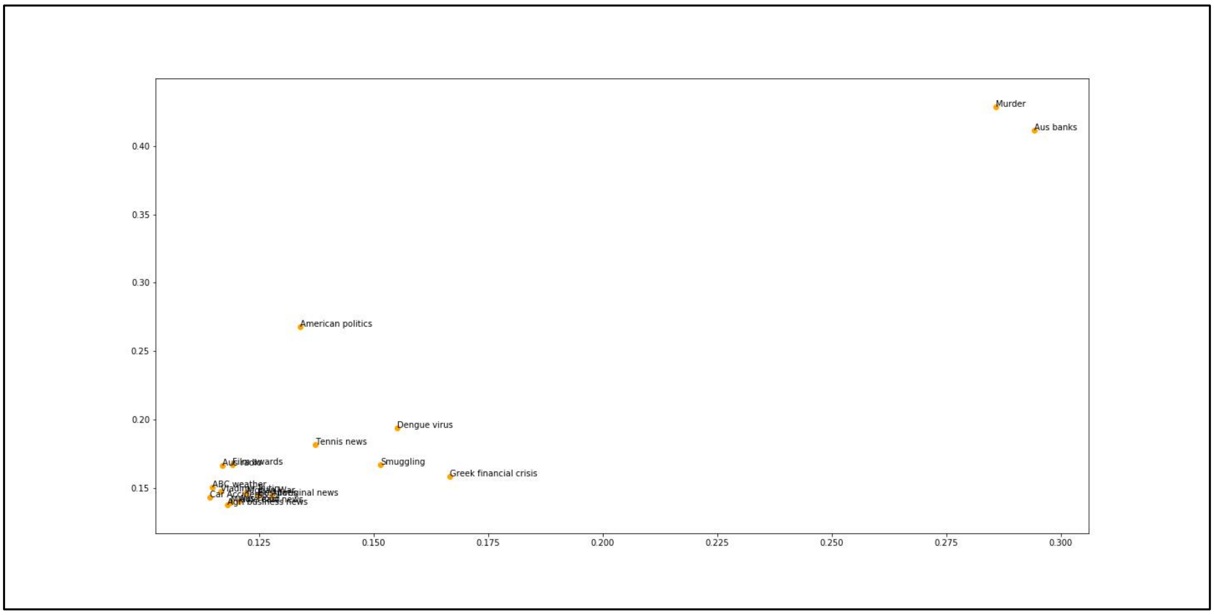


Figure 13 ASX200 Financials high stock volatility quadrant 2009

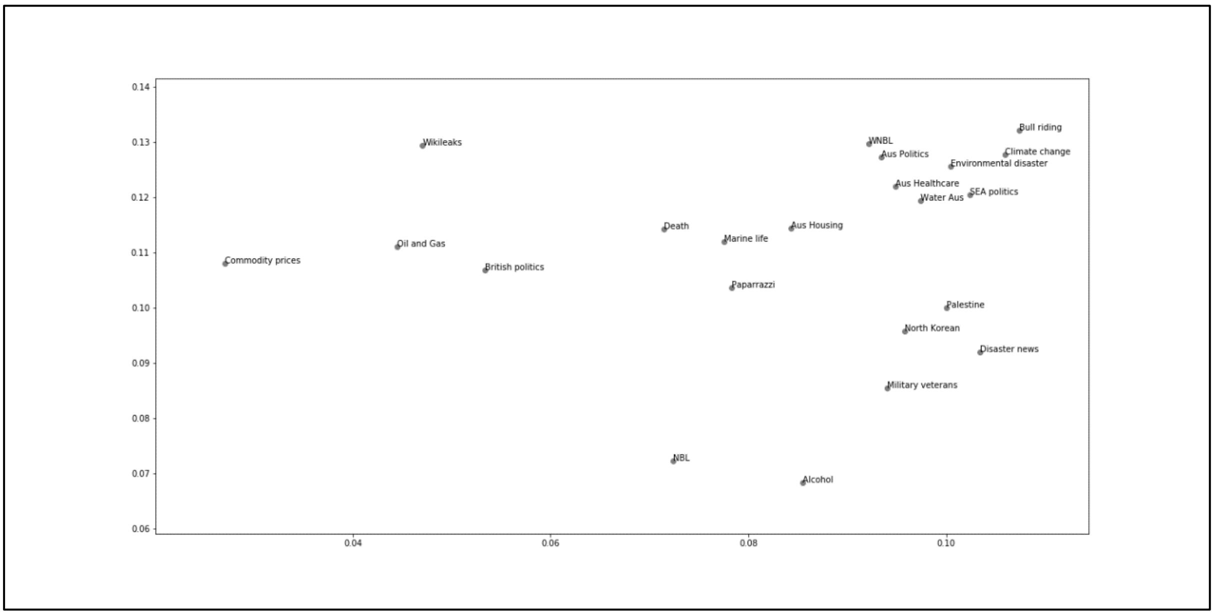


Figure 14 ASX200 Financials low stock volatility quadrant 2009

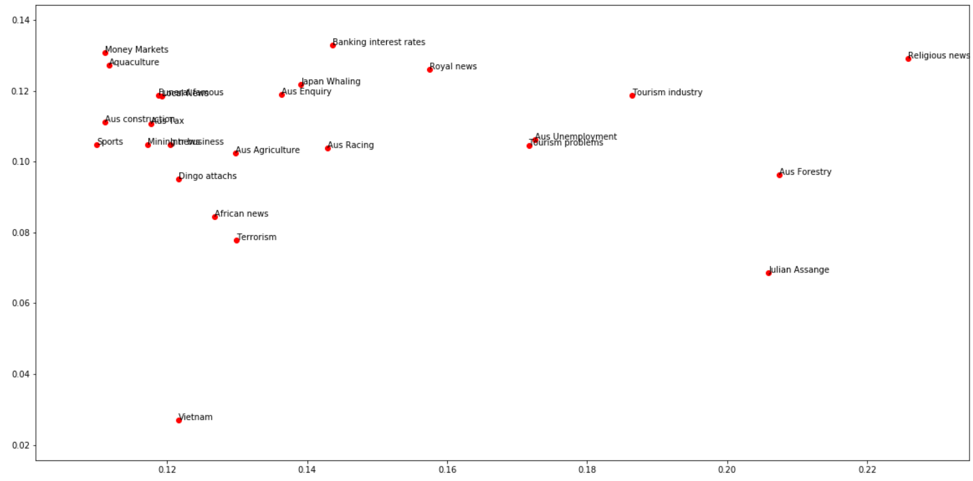


Figure 15 ASX200 Financials sell signal stock volatility quadrant 2009

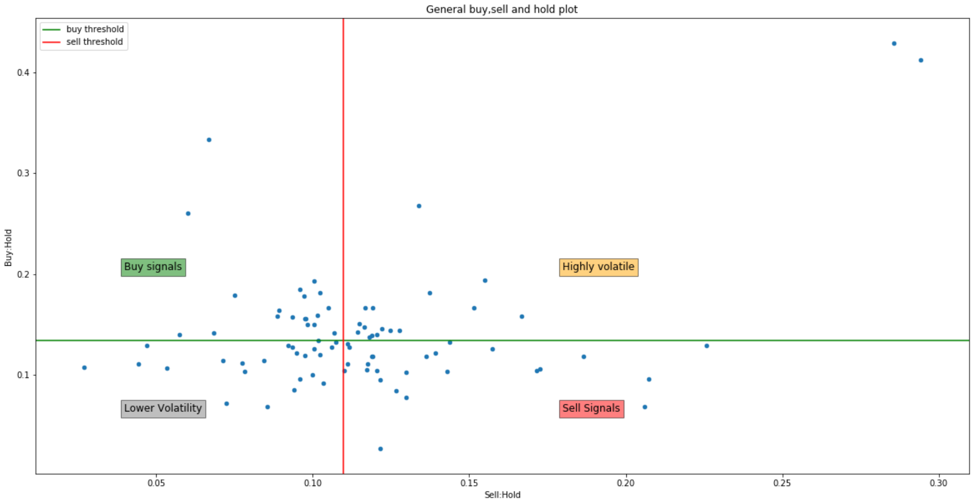


Figure 16 ASX200 financials stock volatility vs news topics quadrant 2009

2012 Figures

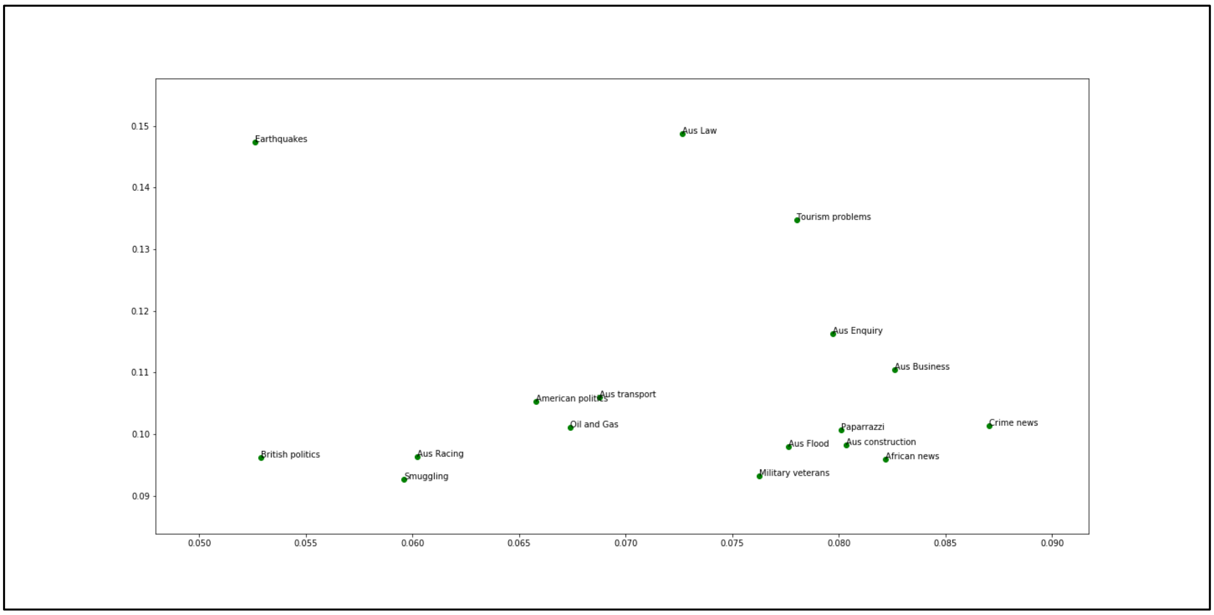


Figure 17 DJIA stock index 2012 Sell signal quadrant with new topics annotated to scatter points

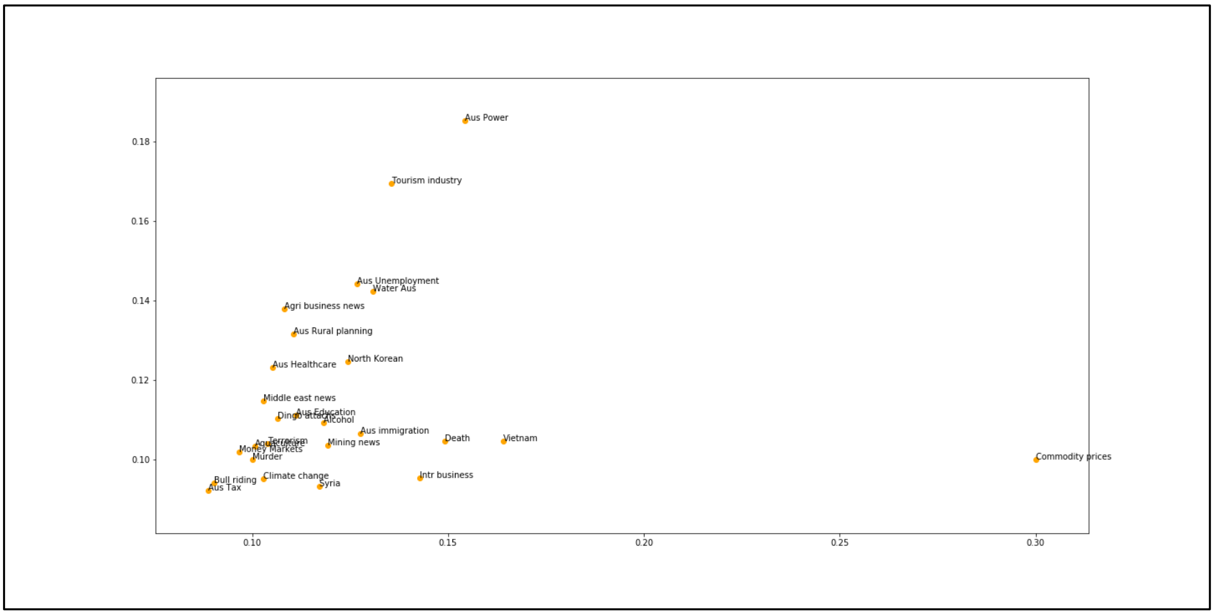


Figure 18 DJIA stock index 2012 high volatility quadrant with new topics annotated to scatter points

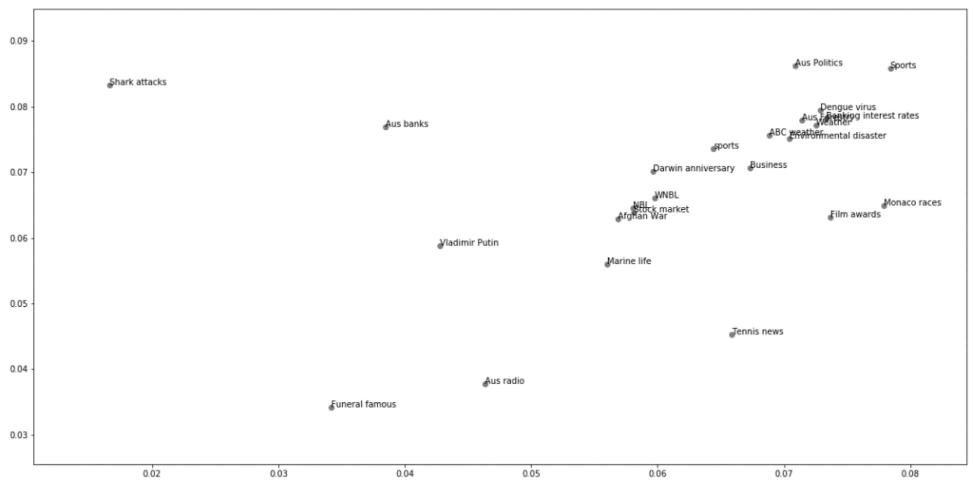


Figure 19 DJIA stock index 2012 low volatility quadrant with new topics annotated to scatter points

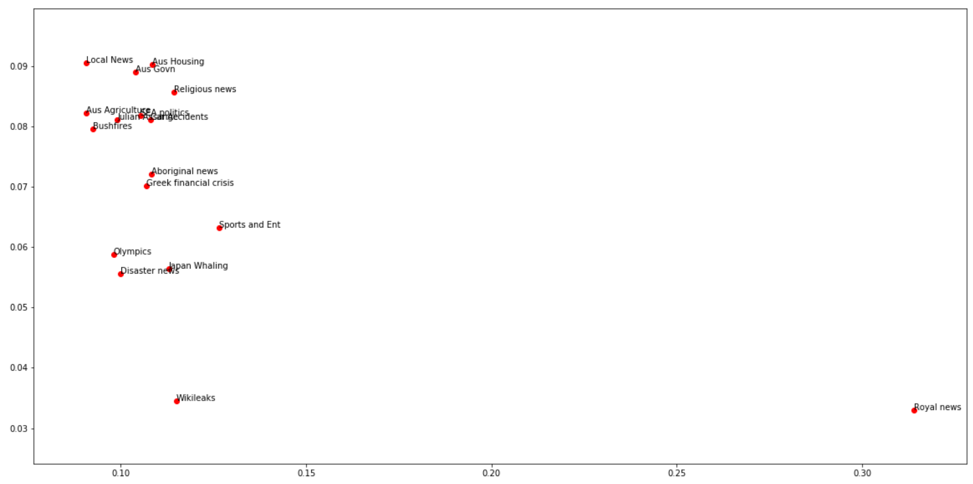


Figure 20 DJIA stock index 2012 sell signal quadrant with new topics annotated to scatter points

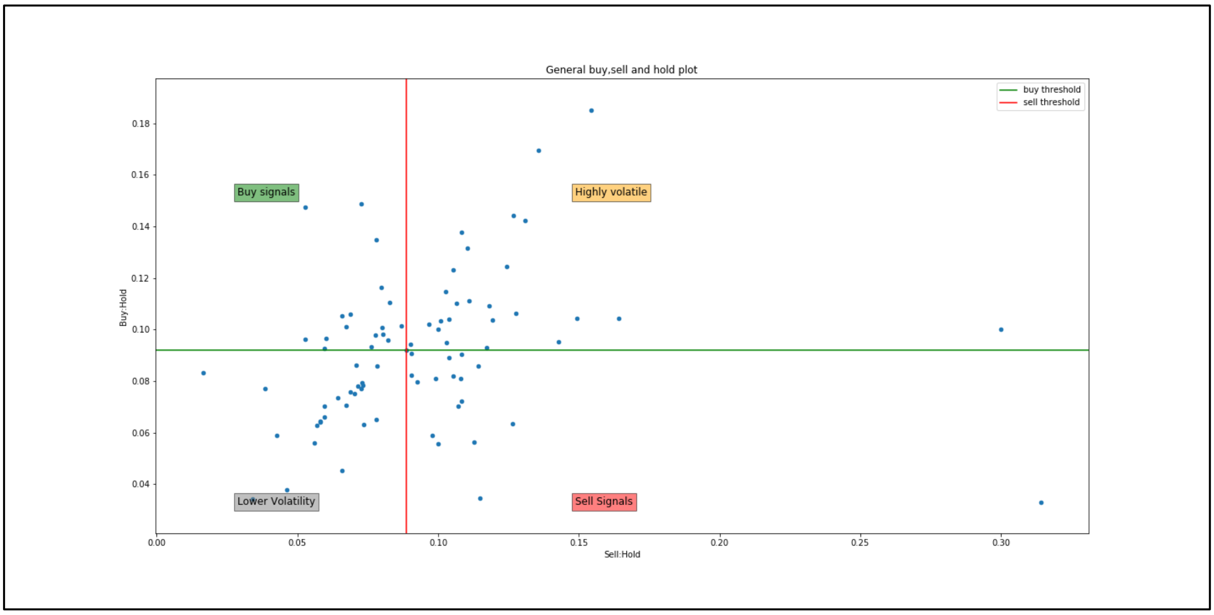


Figure 21 DJIA stock index 2012 scatters quadrant of stock volatility versus news topics

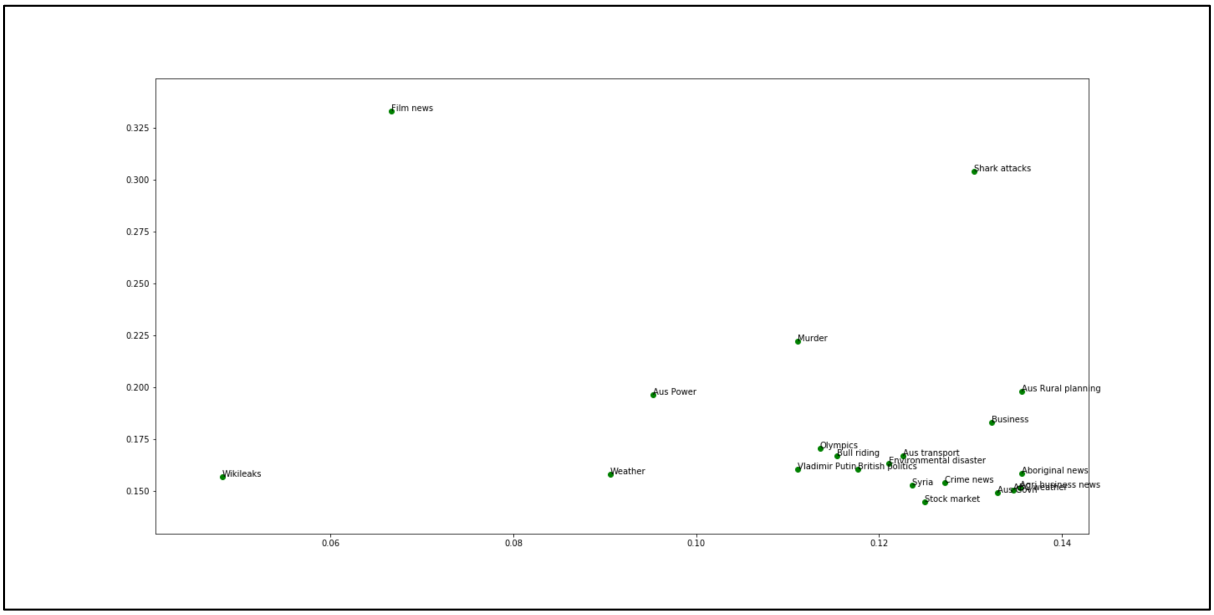


Figure 22 ASX200 stock index 2012 quadrant of buy signal stock volatility versus news topics

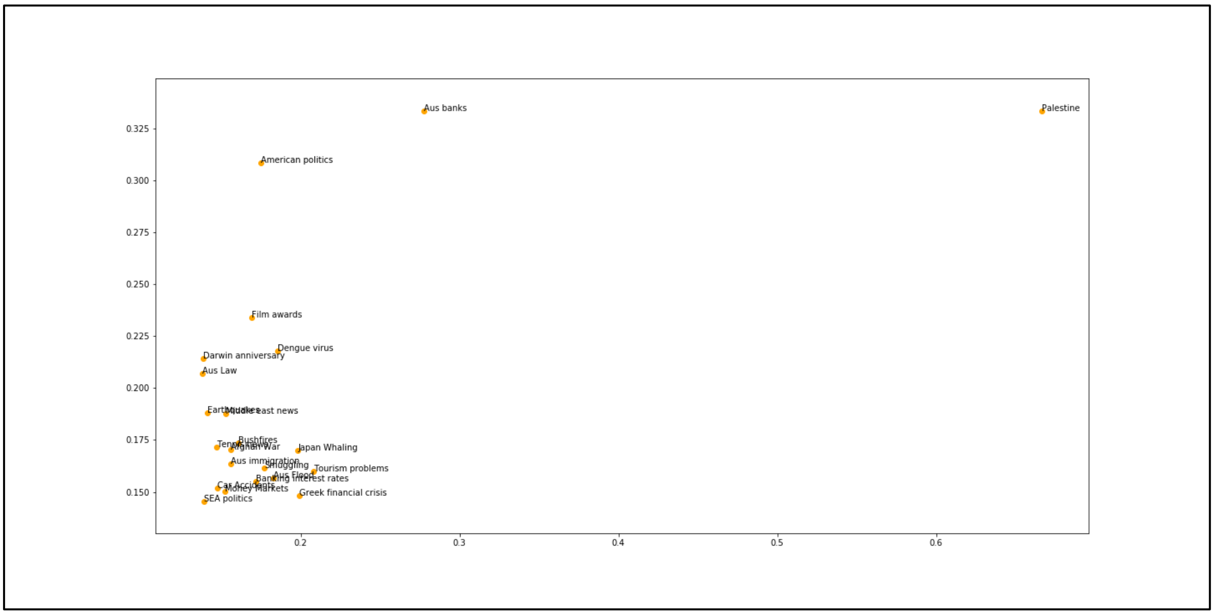


Figure 23 ASX200 stock index 2012 quadrant of high stock volatility versus news topics

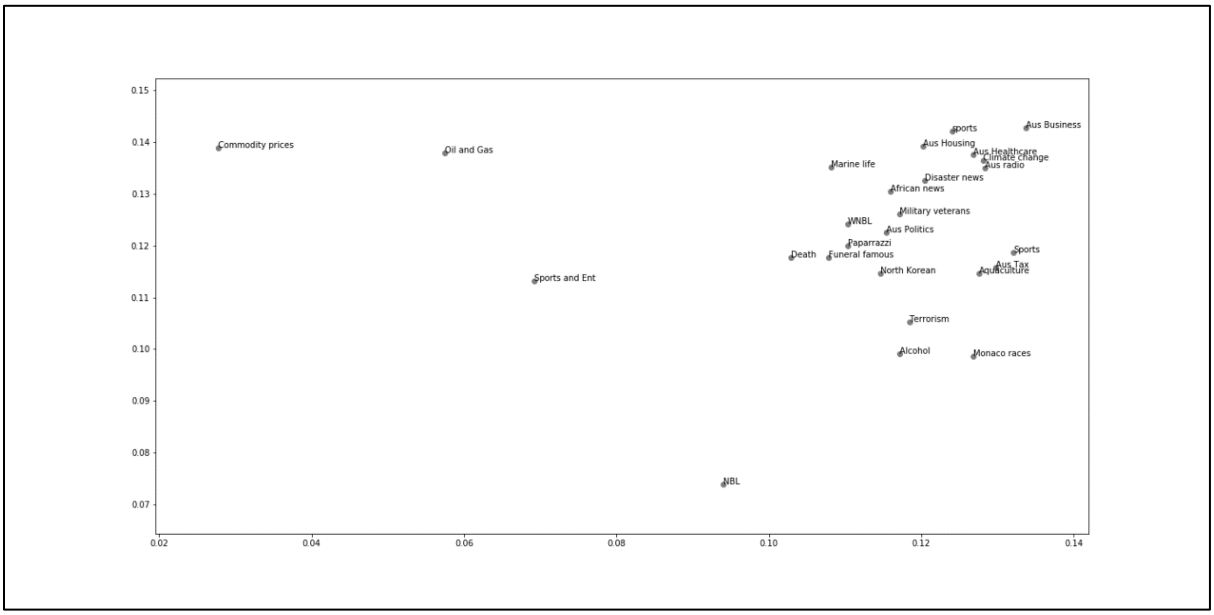


Figure 24 ASX200 stock index 2012 low stock volatility versus news topics

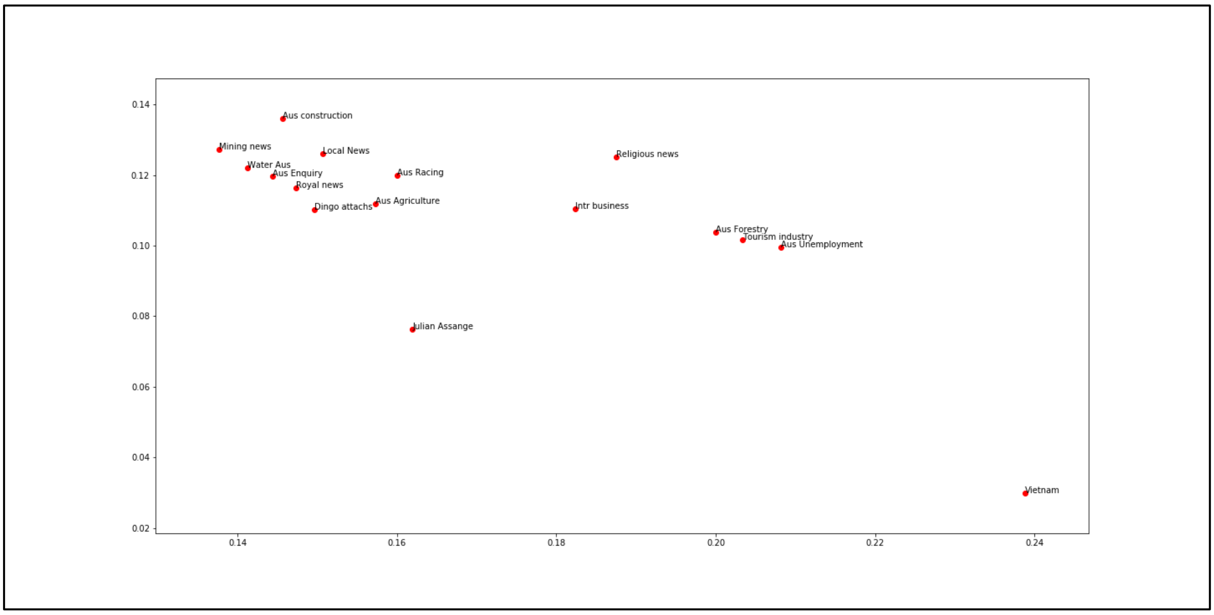


Figure 25 ASX200 stock index 2012 sell signal quadrant of stock volatility versus news topics

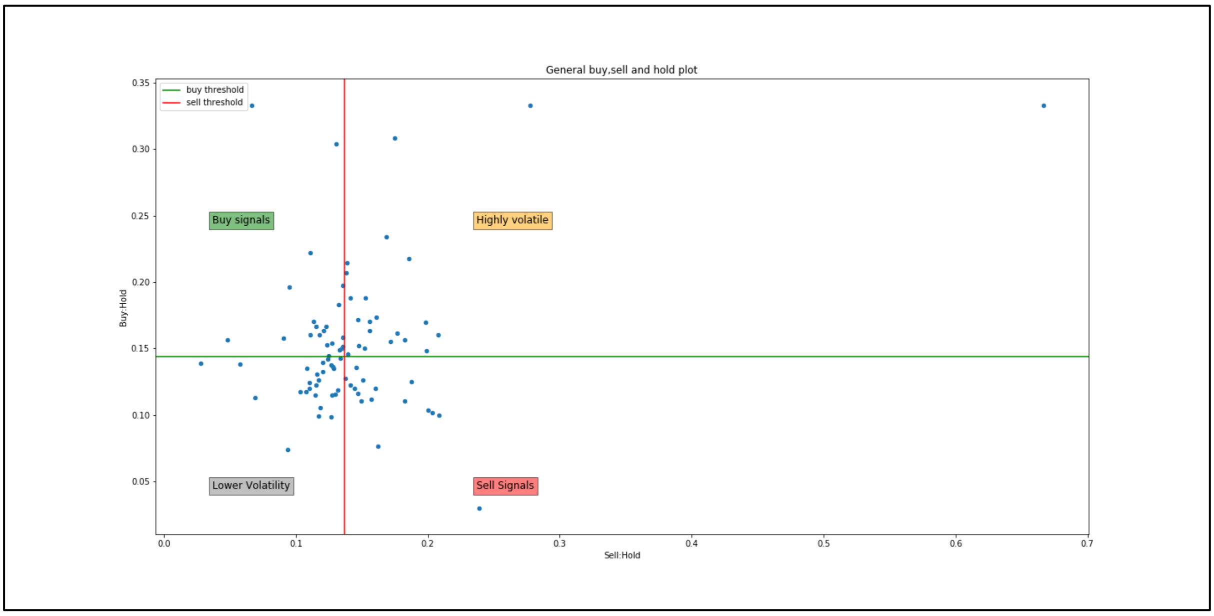


Figure 26 ASX200 stock index 2012 scatters quadrant of stock volatility versus news topics

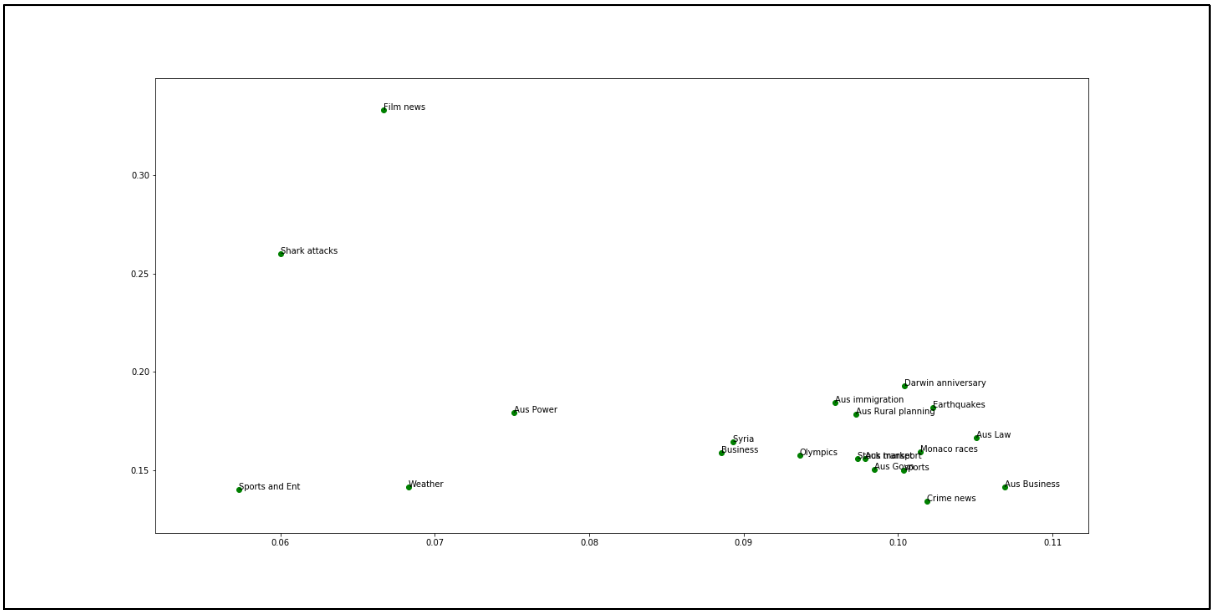


Figure 27 ASX200 financials stock index 2012 quadrant of buy signal stock volatility versus news topics

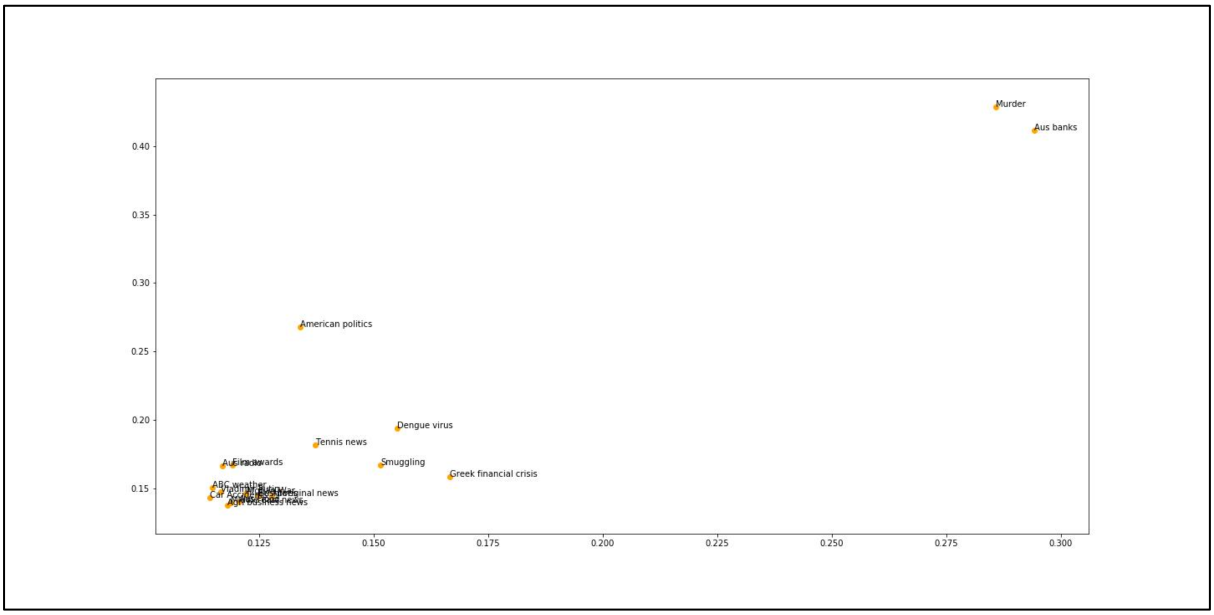


Figure 28 ASX200 financials stock index 2012 quadrant of high stock volatility versus news topics

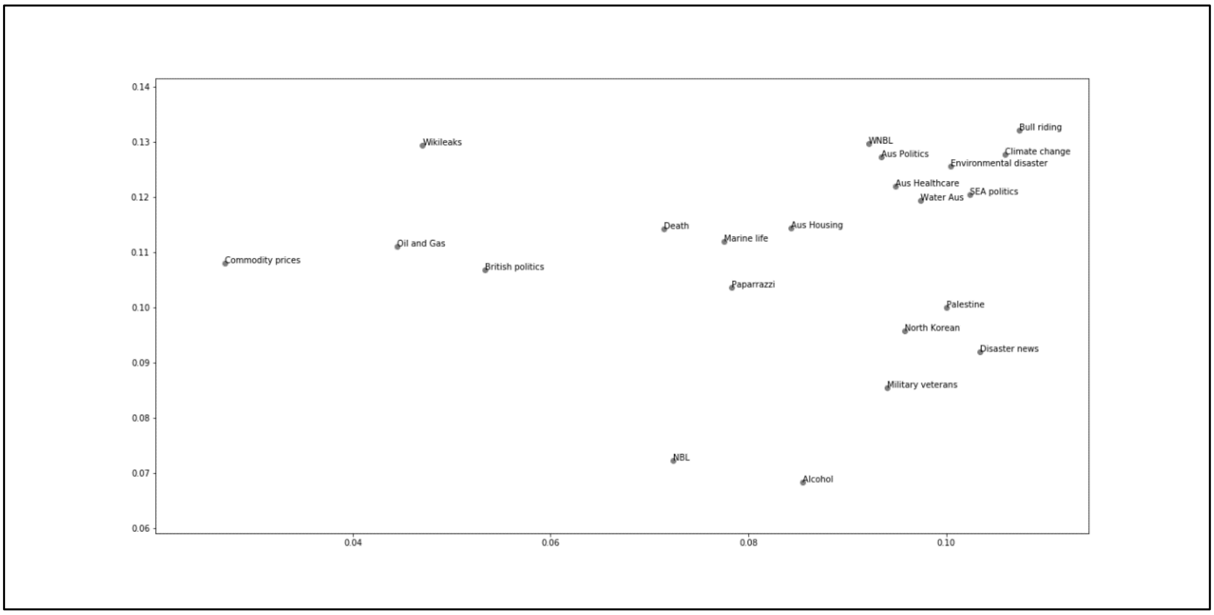


Figure 29 ASX200 financials stock index 2012 quadrant of low stock volatility versus news topics

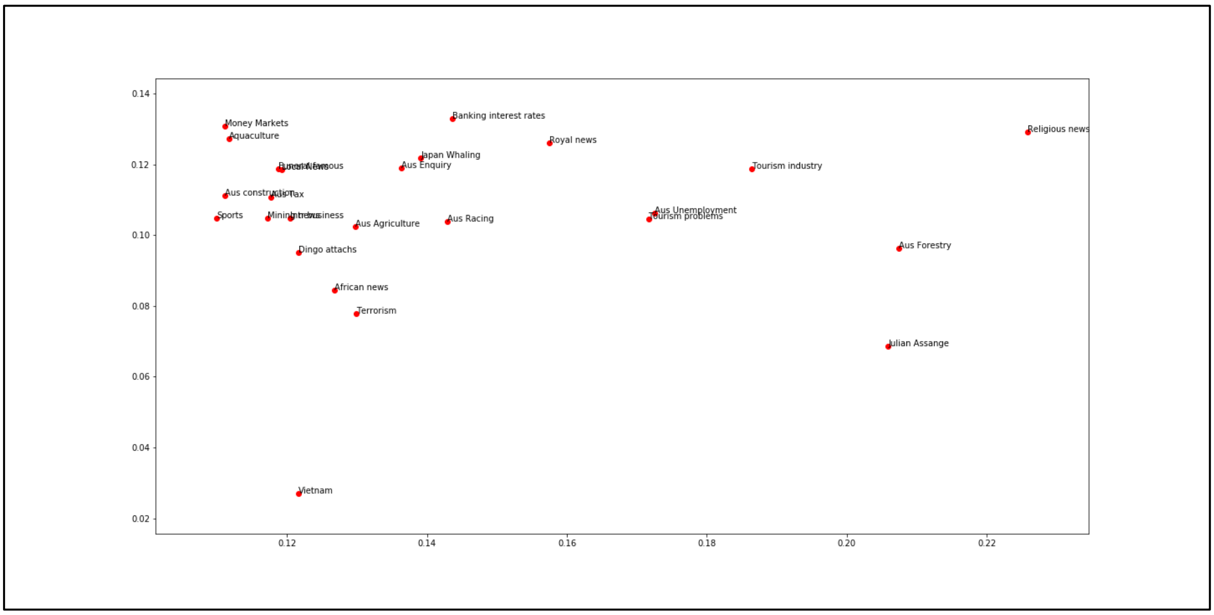


Figure 30 ASX200 financials stock index 2012 quadrant of sell signal stock volatility versus news topics

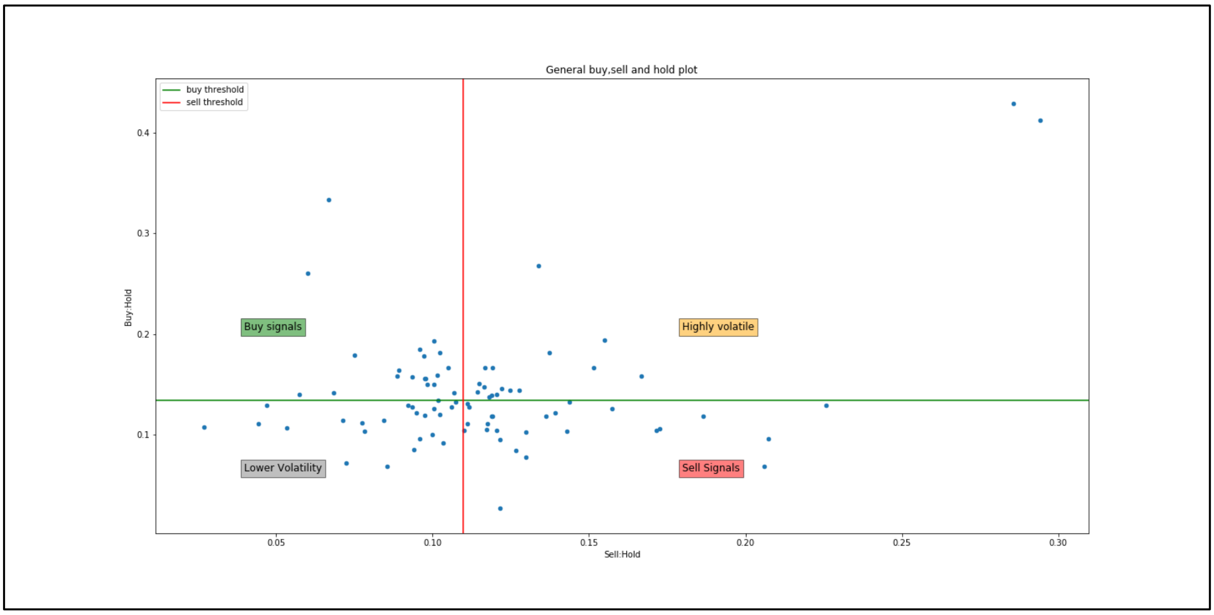


Figure 31 ASX200 financials stock index 2012 scatters quadrant of stock volatility versus news topics

2014

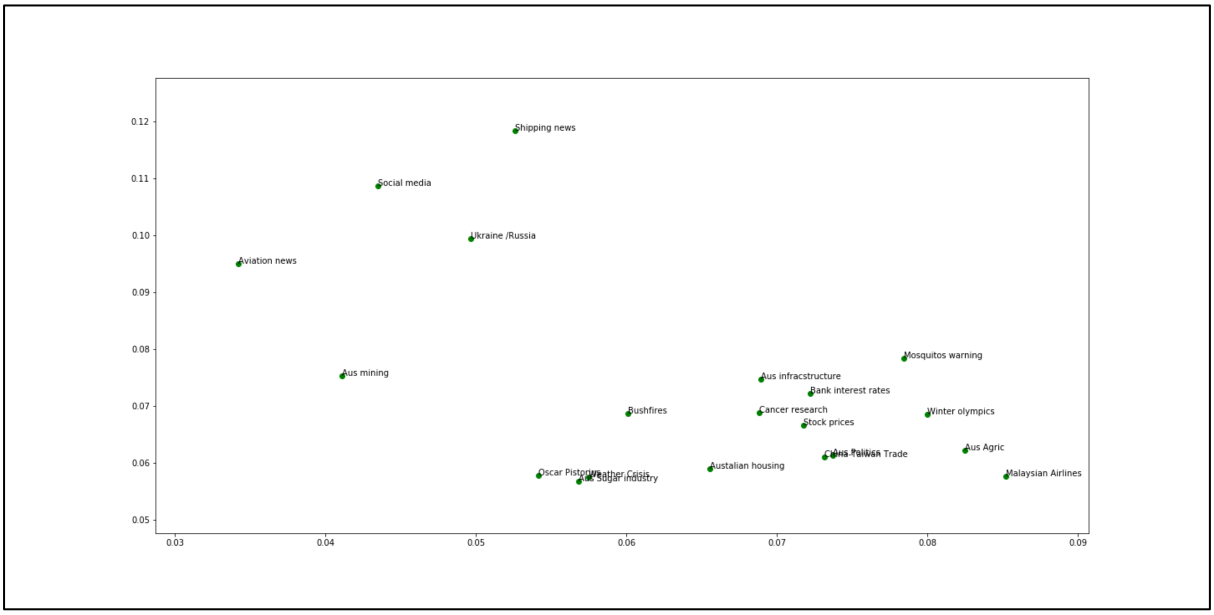


Figure 32 DJIA stock index 2014 buy signals of stock volatility versus news topics

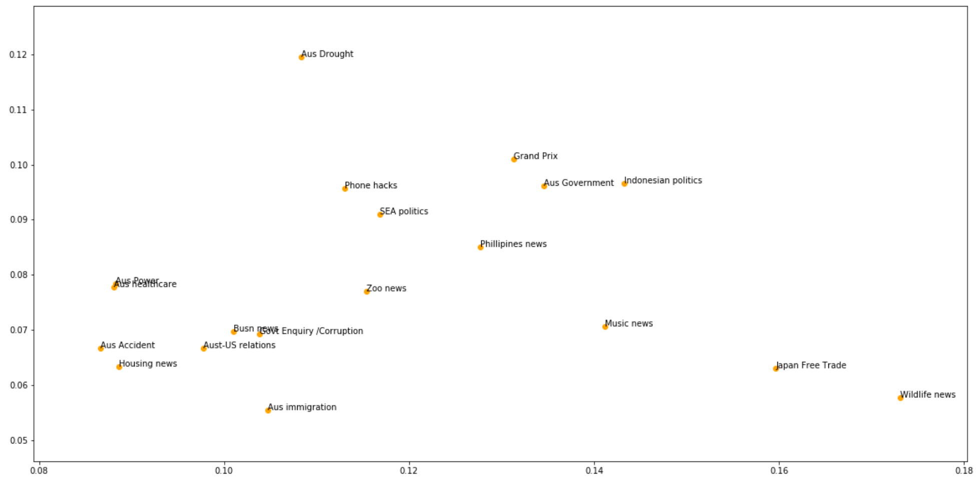


Figure 33 DJIA stock index 2014 high stock volatility versus news topics

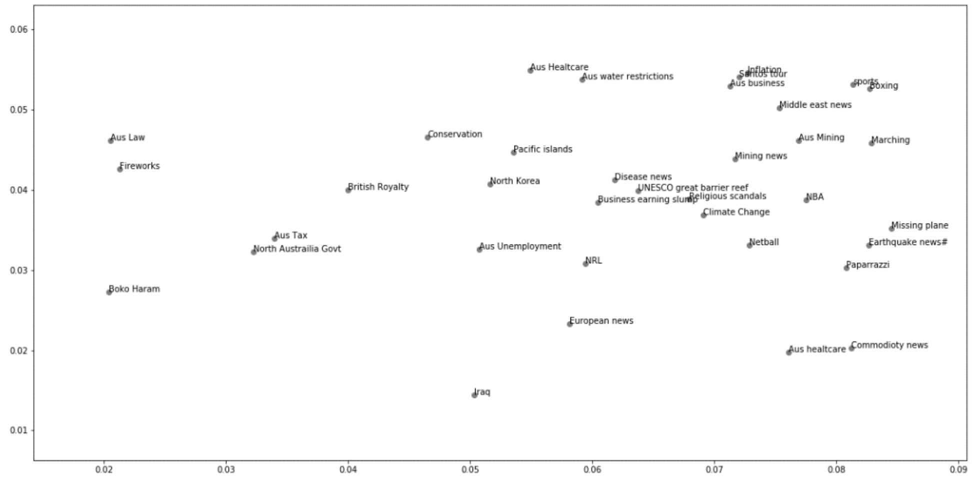


Figure 34 DJIA stock index 2014 low stock volatility versus news topics

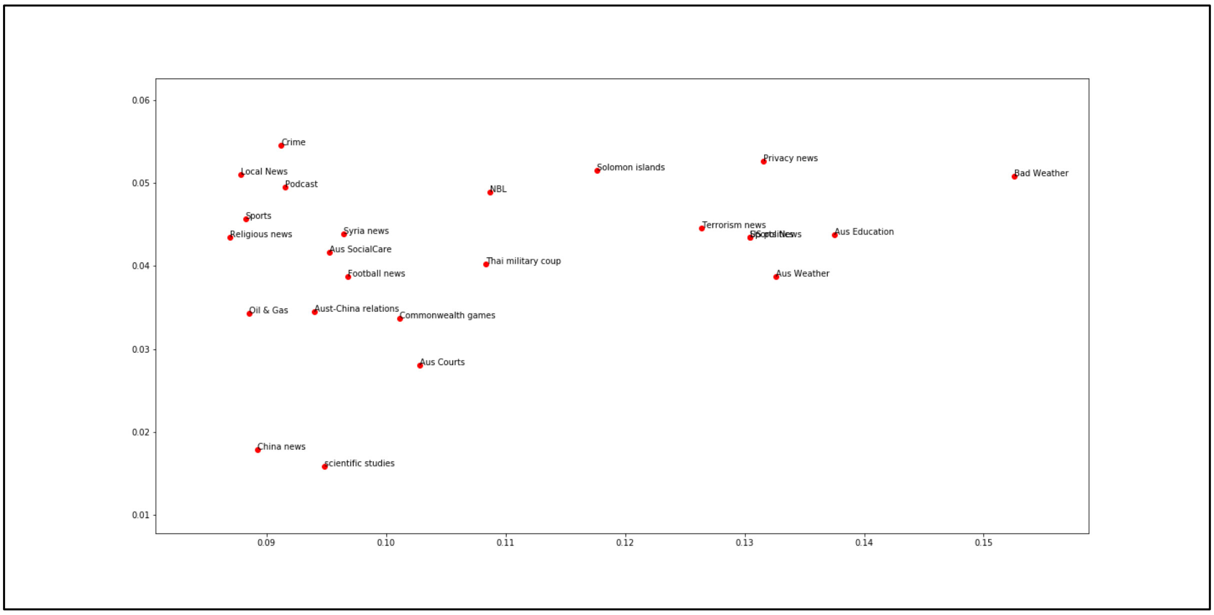


Figure 35 DJIA stock index 2014 buy signals of stock volatility versus news topics

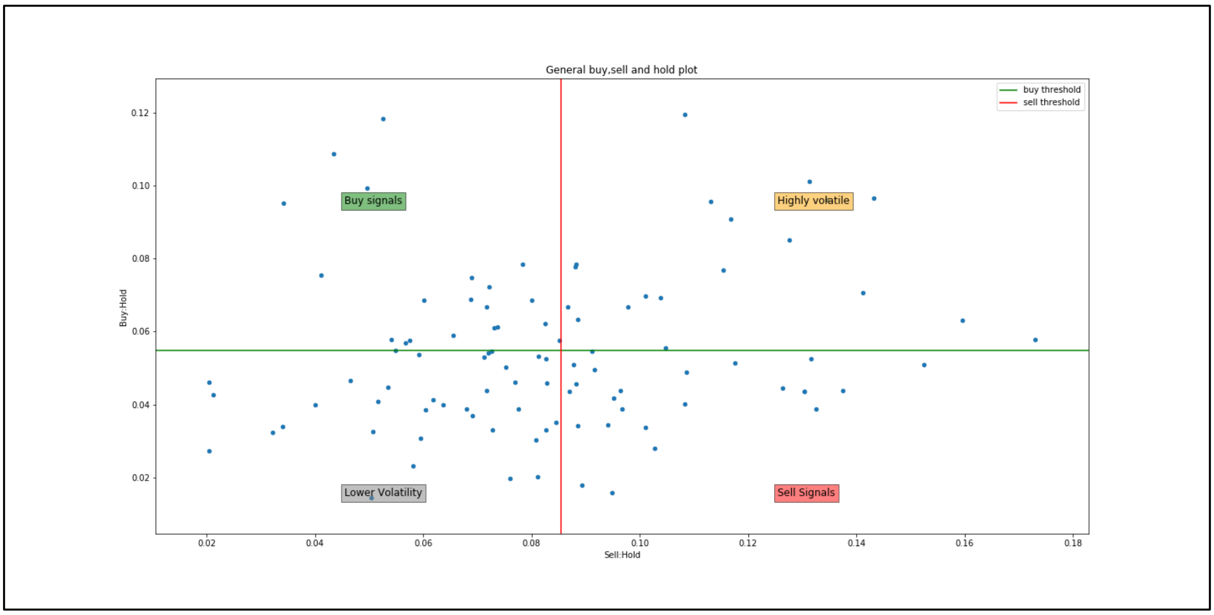


Figure 36 DJIA stock index 2014 scatters quadrant of stock volatility versus news topics

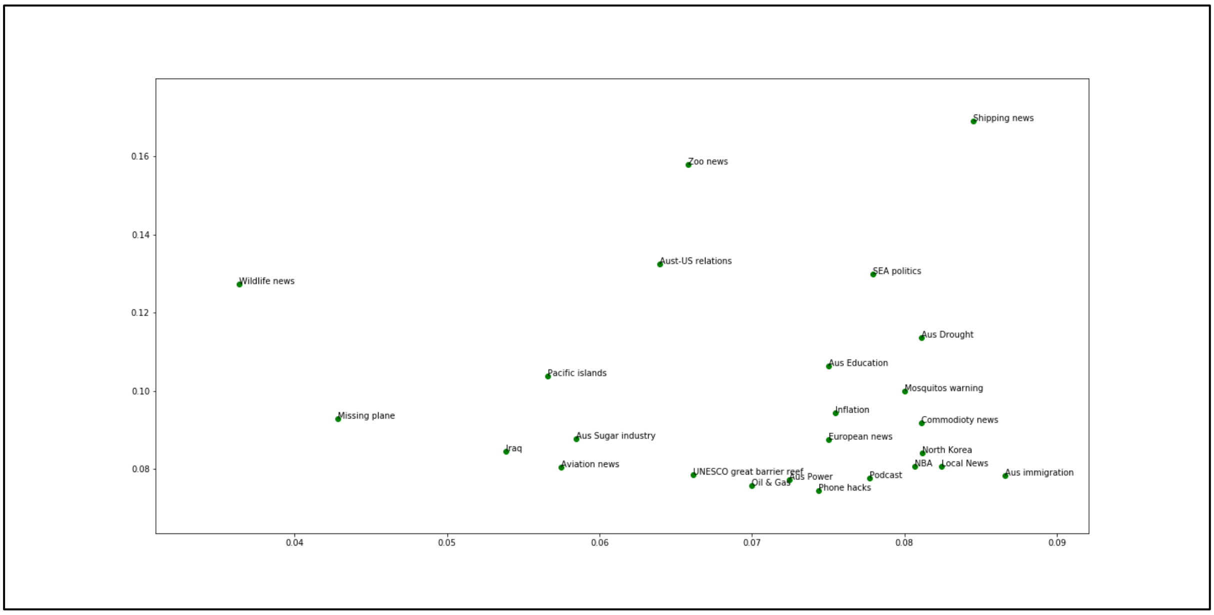


Figure 37 ASX200 stock index 2014 scatters quadrant of sell signals stock volatility versus news topics

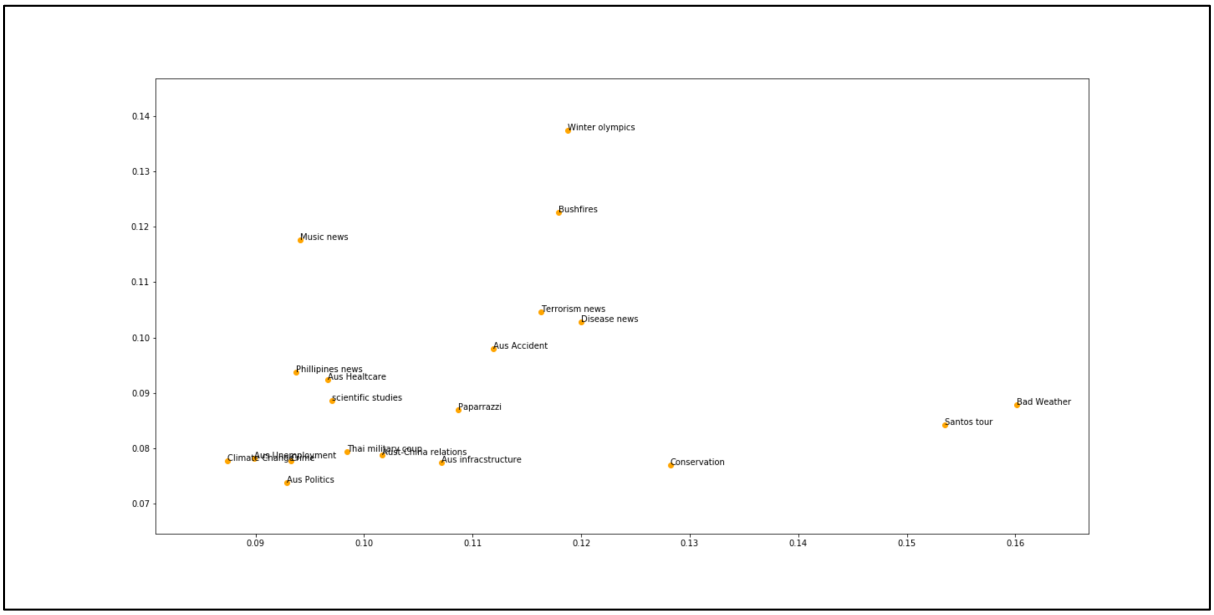


Figure 38 ASX200 stock index 2014 scatters quadrant of high stock volatility versus news topics

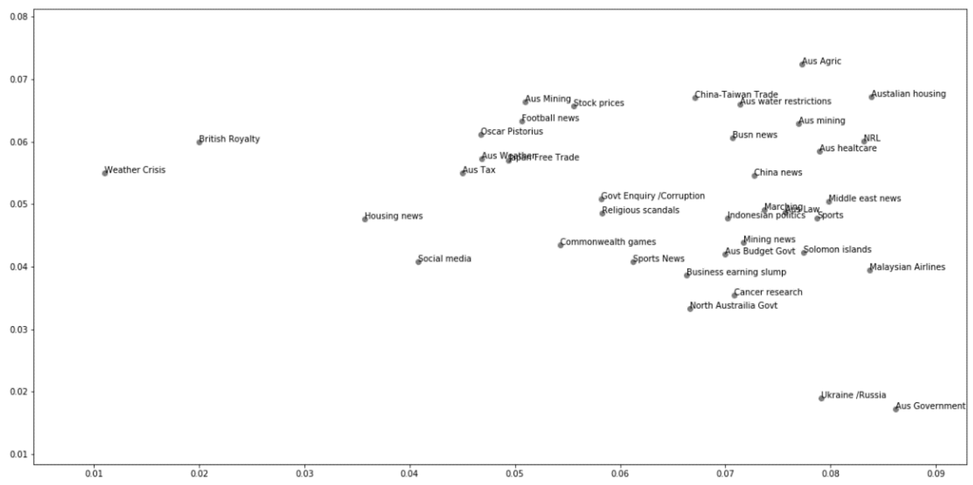


Figure 39 ASX200 stock index 2014 scatters quadrant of low stock volatility versus news topics

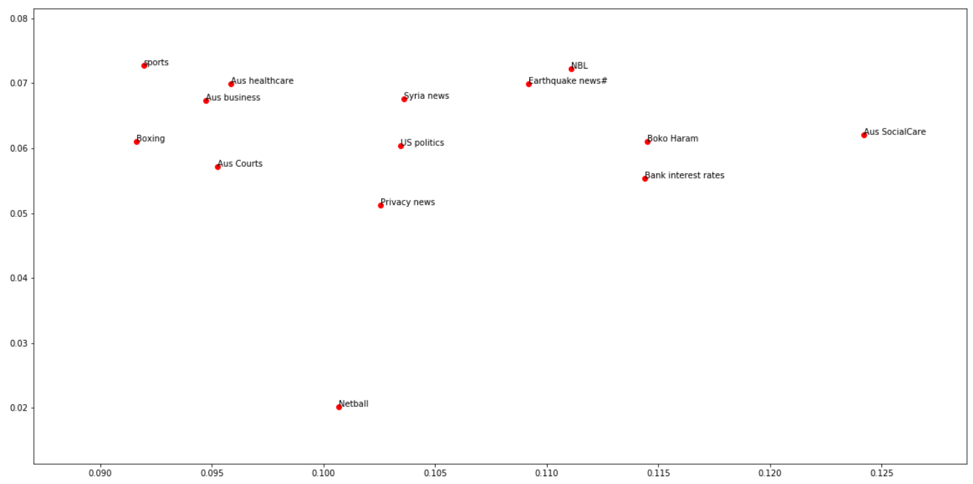


Figure 40 ASX200 stock index 2014 scatters quadrant of sell signal stock volatility versus news topics

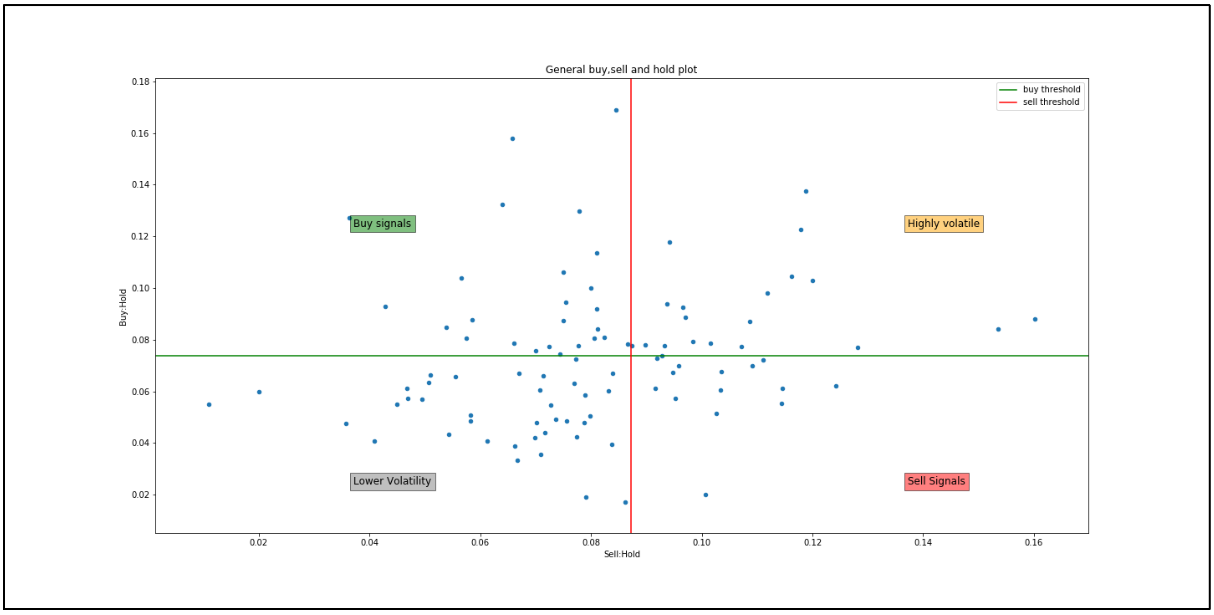


Figure 41 ASX200 stock index 2014 scatters quadrant of stock volatility versus news topics

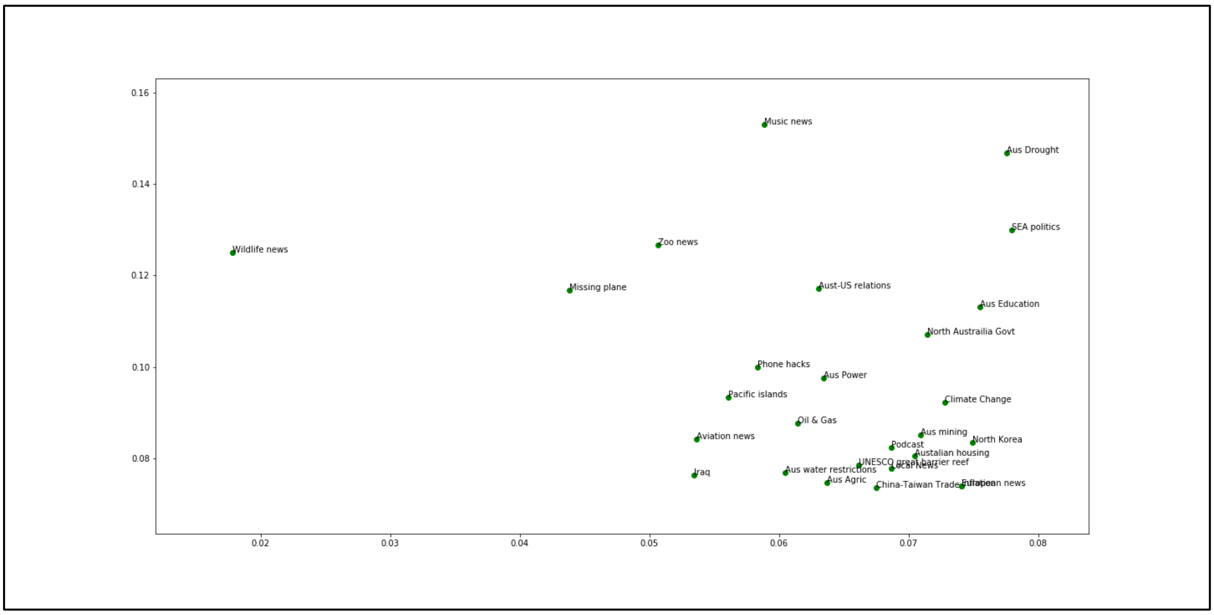


Figure 42 ASX200 financials stock index 2014 scatters quadrant of buy signal stock volatility versus news topics

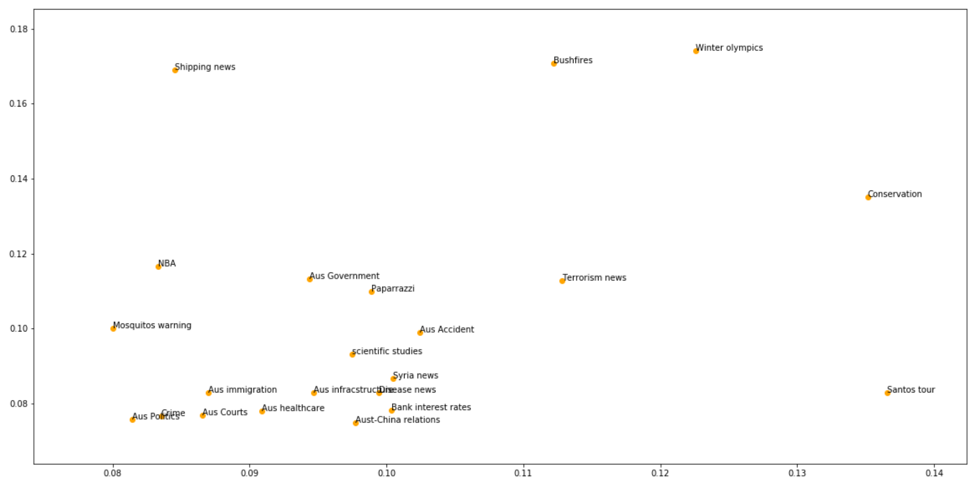


Figure 43 ASX200 financials stock index 2014 scatters quadrant of high stock volatility versus news topics

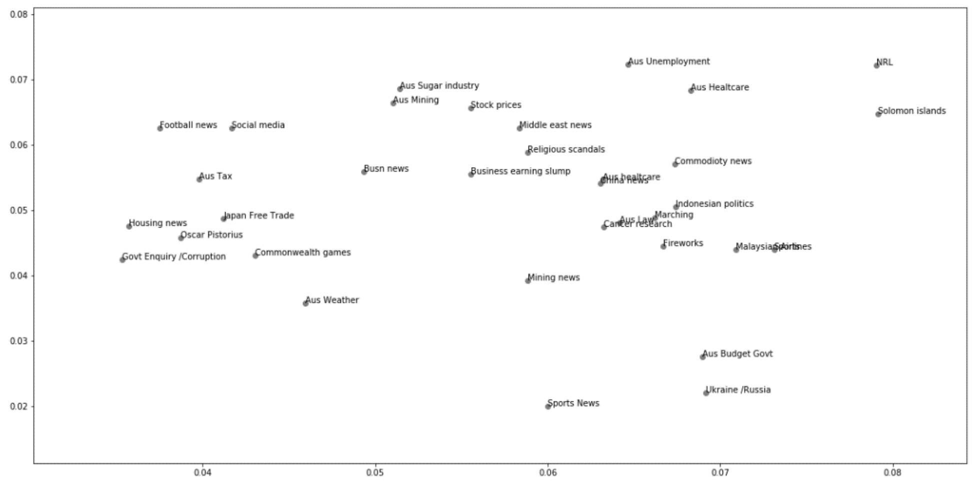


Figure 44 ASX200 financials stock index 2014 scatters quadrant of low stock volatility versus news topics

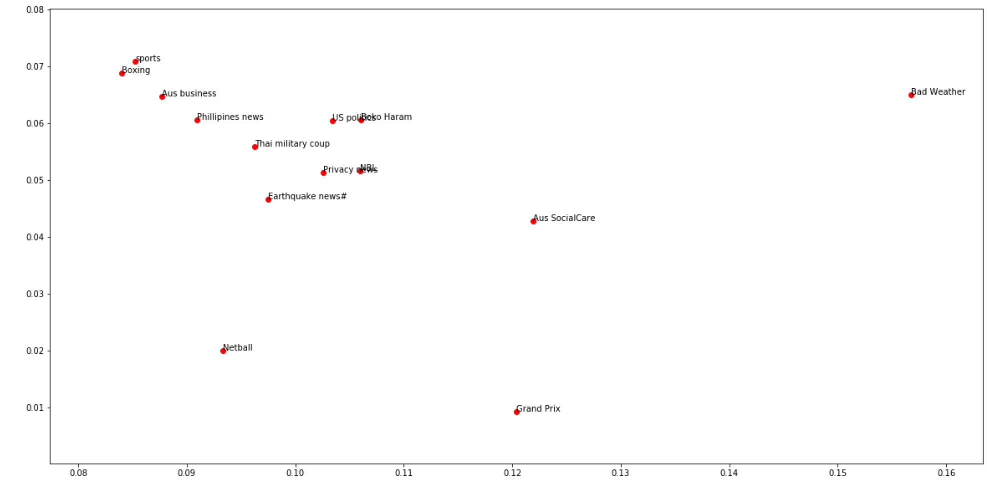


Figure 45 ASX200 financials stock index 2014 scatters quadrant of sell signal stock volatility versus news topics

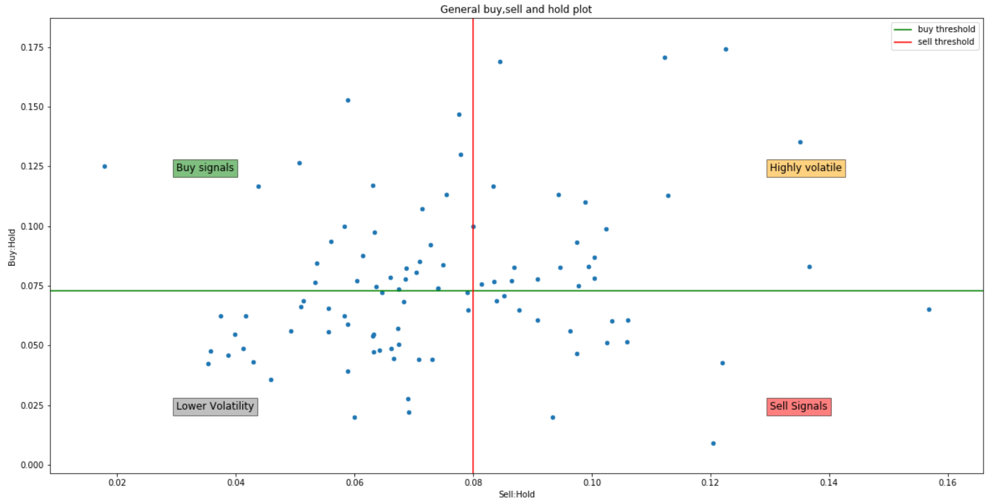


Figure 46 ASX200 financials stock index 2014 scatters quadrant of buy signal stock volatility versus news topics

References

[1] S. A. Y. W. C. L. N. Arman Khadjeh Nassirtoussi, “Text mining for market prediction: A systematic review,” *Expert Systems with Applications,* vol. 41, pp. 7653-7670, 2014.

[2] S. indices, “www.spindices.com,” Standard & Poor's company, [Online]. Available: https://au.spindices.com/indices/equity/dow-jones-industrial-average. [Accessed 30 Oct 2018].

[3] S. i. ASX200, “www.spindices.com,” Standard and Poor's company, [Online]. Available: https://au.spindices.com/indices/equity/sp-asx-200 . [Accessed 30 Oct 2018].

[4] S. A. 2. financials, “www.spindices.com,” Standard & Poor's company , [Online]. Available: https://us.spindices.com/indices/equity/sp-asx-200-financials-sector. [Accessed 31 Oct 2018].

[5] R. Kulkarni, “Million headlines,” KonivaC, [Online]. Available: https://www.kaggle.com/therohk/million-headlines/home . [Accessed 1 Oct 2018].

[6] S. A. Y. W. C. L. N. Arman Khadjeh Nassirtoussi, “Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension redcuction algorithm with semantics and sentiment,” *Expert Systems with Applications,* vol. 42, no. 1, pp. 306-324, 2015.

[7] B. D. D. Hong Liangjie, “"Empirical study of topic modeling in twitter.",” in *ACM Proceedings of the first workshop on social media analytics*, Washington, 2010.

[8] A. B. H. Michael Röder, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the eighth ACM international conference on Web search and data mining.* , New York, 2015.

[9] C. W. v. A. W. Jacobi, “Quantitative analysis of large amounts of journalistic texts using topic modelling,” *Digital Journalism,* vol. 4, no. 1, pp. 89-106, 2016.

[10] B. D. D. Liangjie Hong, “Empirical Study of Topic Modeling in Twitter,” in *Proceedings of the first workshop on social media analytics ACM.*, 2010.

[11] Andersen Torben G, “The distribution of realized stock return volatility.,” *Journal of financial economics,* vol. 61, no. 1, pp. 43-76, 2001.

[12] C. W. Wei X, “LDA-based document models for ad-hoc retrieval.,” in *29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, 2006.

[13] J. G. S. L. X. C. Y. W. Yan Xiaohui, “Learning Topics in Short Texts by Non-negative Matrix Factorization on term correlation matrix,” in *Proceedings of the 2013 SIAM International Conference on Data Mining*, Austin Texas, 2013.

[14] J. J. W. J. H. J. L. E. Y. H. L. X. Zhao WX, “"Comparing twitter and traditional media using topic models.",” in *Springer European conference on information retrieval.*, Berlin, Heidelberg, , 2011.