An exploration into the use of Natural Language Processing in Stock Market Prediction on the JSE

Robert Brink Marcus Gawronsky Christopher Kleyweg

17 May 2018

Introduction

According to the Efficient Market Hypothesis, markets take into account all relevant information in efficiently pricing securities (Fama, 1970). While many studies have explored this supposition under its strong, semi-strong and weak form, the growing volume, velocity and variety of market data has forced financiers to invest more-and-more in technology as a tool for investor decision making. Investors form a mosaic of information from financial reports, news articles and price data. With the growing trend towards automated trading, a requirement to explore new forms of unstructured and semi-structured data in order to remain competitive has emerged. This literature review aims to explore the use of text data in quantitative stock price prediction, surveying the growing number of techniques in Natural Language Processing to extract evolving market sentiments.

Literature Review

Natural Language Processing

The Oxford Dictionary contains 180 976 words which can be combined to form a multitude of sentences (Oxford University Press, 2018). While all of these sentences are unique, many share similar meanings. Algorithms need to find a way of representing these sentences in an efficient and useful manner which captures both the similarity and distinctiveness of these sentences for an array of applications from Human-Computer-Interaction to Document Retrieval.

The realm of Natural Language Processing (NLP) has seen increased attention in recent years with the growth of new techniques, datasets and computing capacity (Sohangir et al., 2014, Fisher et al. (2009), Cortis et al. (2017)). Early applications of NLP include the Bag-of-Words (BOW) approach, a simple technique described by Harris (1954) in which a document is represented by frequency table or matrix, counting the number of times a word appears in a document. While this approach has many drawbacks (Harris, 1954), it does produce document vectors which capture the common meaning of documents

sharing the same words. In English, different words and phrases can share or have different meanings. A key challenge of NLP comes in how to group these words and phrases in a way which is automated and efficient.

While predefined dictionaries do exist for grouping words, one simple and generalized technique has been through the use of stemming. Stemming is a non-parametric technique in which common suffixes or prefixes are removed from words based through a set of predefined rules. In an early algorithm developed by Lovins (1968), suffixes like *'ies'* are removed from words to reduce the dictionary of words contained in a corpus of documents and better capture their common meaning (Porter, 1980, Rani, Ramesh & Anusha (2015)).

In English, phrases can bear new meanings distinct from the words that comprise them. This phenomenon of *'collocation'* remains an extensive area of research in NLP (Lin, 1998). Many statistical techniques have been developed to extract these phrases, referred to as n-grams. While these techniques increase the dictionary of a given corpus, they can improve performance across applications by better capturing the meaning expressed by documents (Pecina, 2005).

One disadvantage in the non-parametric Bag-of-Words technique is that it does not weight words based on their distinctiveness in the corpus of documents. Articles like *'a'* or *'the'* occur throughout all English text, but contribute little meaning to a given sentence. While much work has been done analysing the information in a given piece of text (Shannon, 1951), term-frequency re-weighting has been a common staple with the Bag-of-Words approach to document vectorization (Spärck Jones, 1972, Robertson (2004)). Using Term Frequency-Inverse Document Frequency (TFIDF), the likelihood of choosing a word in a document is multiplied through by its log-likelihood across the corpus. While this technique can increase the sparsity of a given document vector, it can improve the performance of document similarity measures, valuable in topic analysis and document clustering (Huang, 2008).

Unsupervised techniques for dimensionality reduction have been popular extensions on the Bag-of-Words approach. Papers by Dumais et al. (1988) and Deerwester et al. (1990), through a technique called Latent Semantic Indexing (LSI), compare the use of factor analysis, principal component analysis and multidimensional scaling on count vectors in order to extract continuous document vectors which control for problems of matrix sparsity and polysemy in which many words have the same meaning.

A generative approach by Blei et al. (2003), referred to as Latent Dirichlet Allocation (LDA), uses a three-level hierarchical Bayesian model to extract document vectors from a given corpus. These document vectors describe each document as a mixture of latent predefined topics. Each word in the corpus has some probability of coming from a given topic and each topic has some probability of appearing in the corpus. The topics are then mixed according to a Dirichlet Distribution to form each document. This technique can be computationally challenging but has become a staple in many online applications.

While LSI and LDA have demonstrated significant improvements to document vectorization or embedding, works by Shannon (1951) and Huang et al. (1993) seem to indicate that a word's meaning can be derived from the contexts it and other words find themselves in rather than just their frequency (Baroni,

Dinu & Kruszewski, 2014). These contexts, known as skip-grams, refer to a sequence of words surrounding a target word and may be common for a number of different polysemes. In the sentence, *'The quick brown fox jumps'* the skip-gram for *'brown'* would be *'The quick fox jumps'*. This skip-gram may repeat again elsewhere in the corpus for the adjective *'red'*, in the sentence *'The quick red fox jumps'*, indicating that *'red'* and *'brown'* have similar contexts and may have similar meanings.

In a paper by Bengio et al. (2003), a feed-forward neural network is used with one hidden layer to predict a word's skip-gram (Alexandrescu, 2006). Using the output of this hidden layer, Bengio et al. (2003) demonstrate the value of this approach in extracting rich word-vectors which accurately capture the semantic meaning of words in some continuous vector space. While this technique remains tractable on small datasets and dictionaries, a breakthrough came with Mikolov et al. (2013a) and Mikolov et al. (2013b) who used negative sampling on words' skip-grams as a tool to re-parametrizing the model into something more computationally tractable (Goldberg & Levy, 2014). This technique has been extended on by Le & Mikolov (2014) in a method commonly referred to as Doc2Vec, which aims to find documents representations by using the same negative sample technique discussed in Mikolov et al. (2013a).

In recent years many natural language classification and regression problems have done away with document vectorization or embedding techniques to use deep learning as a tool for automatic feature extraction (Kim, 2014, Luong & Manning (2013), Zhang, Zhao & LeCun (2013)). Deep Learning architectures like the Long Short-Term Memory Unit (LSTM) or Convolutional Neural Network (CNN) have shown success in capturing either the locality or sequence-based dependencies contained in this unstructured data. These techniques may benefit from the distributed representation these models allow for (Bengio, 2009). While there exists limited theoretical studies detailing the factors underlying the success of these models recent works by Zhu (2013) and Guss & Salakhutdinov (2016) relate measures of data complexity from the field of topology to both natural language and the complexity of neural network models providing some justification for their implementation in the field.

Qualitative Information in Finance

With the ever-increasing availability of qualitative financial information in the form of news articles, blog posts, message boards and financially based social networks investors are no longer able to efficiently monitor and process massive amounts of unstructured data (Tirunillai & Tellis, 2012). Engelberg (2008) observes that whilst some financial information is still quickly incorporated into markets due to its ease of understanding, other information may be more ambiguous or costly to process and as a result is only reflected in the market over time. Due to this delay in market reaction, hedge funds are beginning to become interested in trading on processed textual data with NLP offering a competitive advantage in both time and cost efficiencies in processing information, allowing for significant profits to be made (Engelberg, 2008).

One of the most notable studies in sentiment analysis in the financial industry

is that of Tetlock (2007). Tetlock uses a General Inquirer approach, which counts the number of times certain words appear in a text based on predefined categories from Harvard's psychological dictionary and finds negative words to have a much stronger correlation with stock returns than other words, offering the greatest predicting power for both one-day market and firm returns. In addition, other existing research acknowledges negative word classifications to be the most effective in measuring tone and offers the greatest level of predicting power of financial variables (Antweiler & Frank, 2002, Das & Chen (2007), Chen, De & Hwang (2014), Tetlock, Saar-Tsechansky & Macskassy (2008), Engelberg (2008), Tirunillai & Tellis (2012)). Tirunillai & Tellis (2012) suggests a reason for this, believing investors to be more loss averse and hence negative information may elicit a stronger response from investors, whilst they choose to overlook positive information believing it to be more unreliable. Further developments of Tetlock's approach included in the innovation of the Loughran & McDonald (2011) Fin-Neg dictionary, a list of 1202 words which typically have a negative connotation in the financial domain, which proved to be more effective than the Harvard psychological dictionary which considers words to be negative in a general sense (Cortis et al., 2017).

Whilst qualitative news articles can offer investors a simplified explanation of hard to quantify aspect's of firm fundamentals (Tetlock, Saar-Tsechansky & Macskassy, 2008), ambiguity and implicitly expressed sentiment remain challenging for human interpretation, and even more challenging for computer algorithms acting with limited context (Das & Chen, 2007). However, naïve Bayes, a commonly used technique in deriving sentiment from news articles, ignores the challenge presented by context and performs rather well in practice (Antweiler & Frank, 2002, Pang et al. (2002)) only being marginally beaten by other probabilistic classifiers such as Maximum Entropy Classification and SVMs (Pang et al., 2002). Early significant research conducted by Gidófalvi (2001) uses Naïve Bayes to predict whether a news article is positive, negative or neutral in tone and attempts thereafter to predict the movement of the associated stock. A significant finding from this study reports predictive power in the interval starting 20 minutes prior to the article release and 20 minutes after news articles become publicly available. A later study by Antweiler & Frank (2002) attempts to classify stock message boards into buy, sell and hold signals. Whilst the study finds that message boards do not help predict long term stock returns, more bullish messages lead to greater trading volumes followed by negative returns the following day Hu & Tripathi (2015). Another study using Naïve Bayes to classify stock message boards by Leung & Ton (2015) demonstrates a size effect as higher message board activity impacts smaller stocks whilst large stocks experience no significant impact. One downfall of Naïve Bayes is that it requires labelled training data which is often difficult to obtain, is costly and impractical (Leung & Ton, 2015).

Tirunillai & Tellis (2012) suggests that User Generated Content (UGC) could produce new information about current performance in a more timeously manner as compared to news articles or analyst reports which are released less frequently. In addition, UGC reflects the actual opinions of users of the content as opposed to external influencers, and as a result can be used to determine the opinion of investors over an extended period of time Tirunillai & Tellis (2012). Chen, De & Hwang (2014) identifies the usefulness of peer-based advice in financial markets

and suggests that investors may begin to rely on peer based advice instead of more traditional financial analysis (Sohangir et al., 2014). Potential dangers of this include a wealth transfer from retail investors to institutional investors as less sophisticated traders are more vulnerable to language misprocessing (Loughran, 2018) and as a result may not be perfectly rational in their trading decisions.

O'Hare et al. (2009) develops a corpus of financial blogs in their analysis and use a simple Bag-Of-Words approach to determine sentiment. Because blogs tend to contain emotive opinions where authors offer their own market predictions, the challenge of implicitly expressed sentiment typically found in news articles is avoided. However, it was found that blogs tend to discuss multiple companies and as a result using document level sentiment classification would yield insignificant results. The study provides a solution to this problem leading to the early development of fine-grained analysis, in which company specific sub-documents are extracted and used for training and testing.

In recent developments, the rise of Big Data has resulted in deep learning based approaches to sentiment classification in the financial domain (Sohangir et al., 2014). Whilst previous studies have used either a rules based text classification approach or a computational statistical approach to classify words into either positive, negative or neutral polarities (Li, 2010), few studies thus far have taken into account grammatical structures to perform sentence level sentiment (Malo et al., 2013). Prior to the popularity of modern deep learning algorithms, Malo et al. (2013) introduced a concept called "financial entities", a predefined phrase bank of default neutral polarities that can either take a positive or negative meaning depending on the context, injecting specialized domain knowledge into a tree-kernel learning framework in order to derive sentence level sentiment. Advents of deep learning are further able to address sentence level sentiment through semantic indexing and data tagging, ensuring information pertaining to word order, proximity and relationships is not lost (Sohangir et al., 2014). Modern approaches in NLP in the financial domain typically include either a lexicon, machine learning, deep learning or hybrid approach (Cortis et al., 2017). In a recent study, Cortis et al. (2017) observes Convolutional Neural Networks (CNNs) to be more accurate than LSTM and Doc2Vec approaches in predicting stock market opinions posted in StockTwits. In addition, Ding et al. (2015) further illustrates the capabilities of CNNs to better illustrate the longer term influence of news events when compared to standard feedforward neural networks.

Market Characteristics

In the prediction of stock market prices two main theories dominate the literature: the Efficient Market Hypothesis (EMH) and Random Walk Theory (Fama, 1970, Malkiel (1973)). These theories not only discuss the conditional probability and movement of prices but also the ability of the market to assimilated different sources of information. Critical literature by Osborne (1962) and Fama (1970) declare weak-form efficiency as the commonly accepted metric for efficient markets. Using weak-form as a measure of baseline efficiency, Seiler & Rom (1997) and Freud & Pagano (2000) confirm the New York Stock Exchange (NYSE) to be efficient. However, with certain stocks exhibiting a non-random walk behaviour,

the NYSE should be considered as less than weak-form efficient. While numerous studies by Heerden et al. (2013), Jefferis & Smith (2005) and Smith, Jefferis & Ryoo (2002) find the JSE to be weak-form efficient over the time periods 1990 to 2013, similar studies by Heerden et al. (2013) and Phiri (2014) dispute these findings demonstrating varying levels of market inefficiency. A later study by Noakes & Rajaratnam (2016) confirms this uncertainty, finding certain stocks to exhibit a non-random walk behaviour, similar to the NYSE. These contrasts in markets efficiency and structure may suggest contrary findings in the use of the Natural Language Models.

While the efficiency of South African Securities remains a continued and important area of research, many studies have analysed mean reversion as a key metric in understanding market over-reaction. A study by Page & Way (1992) on 204 "relatively well traded stocks" on the JSE between 1974 and 1989 found that losing portfolios yielded stronger mean reversion than winning portfolios, consistent with the hypothesis of over-reaction on the JSE (Muller, 1999, Hsieh & Hodnett (2011)). This key phenomenon may suggest event-based studies on the JSE demonstrate results which are more significant or larger in their effect, and may suggest that South African market participants assimilate market information in a way which differs to participants in overseas markets.

The term Investment Style refers to the core investment philosophy adopted by a given market participant (Robertson, Firer & Bradfield, 2000). Common styles include value investing where investors purchase securities which they believe have higher intrinsic values than market values and benefiting from market corrections after-the-fact, momentum investing where investors aims to capitalise on current market trends continuing and size investing where investors form a trading strategy around the size of the stocks traded, either small or large capitalization stocks. While studies in South Africa and the United States by Rensburg & Robertson (2003), Basu (1977) and Fama & French (1992) demonstrate the dominance of value and size investment styles across these two markets, works by Kruger & Toerien (2014) find the consistency of these size effects on the JSE to depend on market stability. These findings may suggest the dominance of certain financial information in these markets and the consistency of similar studies analysing the effects of qualitative information on market prices.

Referrences

Alexandrescu, A. 2006. Factored Neural Language Models. (June):1–4.

Antweiler, W. & Frank, M.Z. 2002. DOI: 10.2139/ssrn.282320.

Baroni, M., Dinu, G. & Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 238–247. DOI: 10.3115/v1/P14-1023.

Basu, S. 1977. Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios : A Test of the Efficient Market Hypothesis. *the journal*

of finance. 32(3):663–682.

Bengio, Y. 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*. 2(1):1–127. DOI: 10.1561/2200000006.

Bengio, Y., Ducharme, R., Vincent, P. & Janvin, C. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*. 3:1137–1155. DOI: 10.1162/153244303322533223.

Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I. & Edu, J.B. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3:993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Chen, H., De, P. & Hwang, B.-h. 2014. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *The Review of Financial Studies*. 27(5):1367–1403. DOI: 10.1093/rfs/hhu001.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S. & Davis, B. 2017. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 519–535. DOI: 10.18653/v1/S17-2144.

Das, S.R. & Chen, M.Y. 2007. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*. 53(9):1375–1388. DOI: 10.1287/mnsc.1070.0704.

Deerwester, S., Furnas, G.W., Landauer, T.K. & Harshman, R. 1990. Indexing by Latent Semantic Analysis Scott. *Journal of the American Society of Information Science*. 41(6):391–407.

Ding, X., Zhang, Y., Liu, T. & Duan, J. 2015. Deep Learning for Event-Driven Stock Prediction. (Ijcai):2327–2333.

Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. & Harshman, R. 1988. Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*. 281–285. DOI: 10.1145/57167.57214.

Engelberg, J. 2008. Costly Information Processing: Evidence from Earnings Announcements. *SSRN Electronic Journal*. 61(7):819–834. DOI: 10.2139/ssrn.1107998.

Fama, E.F. 1970. American Finance Association Efficient Capital Markets : A Review of Theory and Empirical Work. *The Journal of Finance*. 25(2).

Fama, E.F. & French, K.R. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance*. 47(2):427–465. DOI: 10.1111/j.1540-6261.1992.tb04398.x.

Fisher, I., Garnsey, M., Hughes, M., Fischer, I., Garnsey, M. & Hughes, M. 2009. Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*. 16(1-2):21–31. DOI: 10.1002/isaf.

Gidófalvi, G. 2001. Using news articles to predict stock price movements. *Department of Computer Science and Engineering University of California San*

Diego. (December 2004):9. DOI: 10.1111/j.1540-6261.1985.tb05004.x.

Goldberg, Y. & Levy, O. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. (2):1–5. DOI: 10.1162/jmlr.2003.3.4-5.951.

Guss, W.H. & Salakhutdinov, R. 2016. On Characterizing the Capacity of Neural Networks using Algebraic Topology. (2012).

Harris, Z.S. 1954. Distributional Structure. *WORD*. 10(2-3):146–162. DOI: 10.1080/00437956.1954.11659520.

Heerden, D. van, Rodrigues, J., Hockly, D., Lambert, B., Taljard, T. & Phiri, A. 2013. Efficient market hypothesis in South Africa: Evidence from a threshold autoregressive (TAR) model. *MPRA Paper*. 10(50544):1–16. Available: http://ideas.repec.org/p/pra/mprapa/50544.html.

Hsieh, H.-H. & Hodnett, K. 2011. Tests of the overreaction hypothesis and the timing of mean reversals on the JSE Securities Exchange (JSE): The case of South Africa. *Journal of Applied Finance & Banking*. 1(1):107–130.

Hu, T. & Tripathi, A. 2015. The performance evaluation of textual analysis tools in financial markets. *25th Annual Workshop on Information Technologies and Systems, WITS 2015*. 1–17.

Huang, A. 2008. Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*. (April):49–56.

Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Lee, K.F. & Rosenfeld, R. 1993. The sphinx-ii speech recognition system: An overview. *Computer Speech and Language*. 7(2):137–148. DOI: 10.1006/csla.1993.1007.

Jefferis, K. & Smith, G. 2005. The changing efficiency of African stock markets. *South African Journal of Economics*. 73(1):54–67. DOI: 10.1111/j.1813-6982.2005.00004.x.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. DOI: 10.3115/v1/D14-1181.

Kruger, R. & Toerien, F. 2014. The Consistency of Equity Style Anomalies on the JSE during a Period of Market Crisis. *The African Finance Journal*. 16(1):1–18.

Le, Q. & Mikolov, T. 2014. Distributed Representations of Sentences and Documents. In *31st international conference on machine learning*. ed.

Leung, H. & Ton, T. 2015. The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks. *Journal of Banking and Finance*. 55(December 2008):37–55. DOI: 10.1016/j.jbankfin.2015.01.009.

Li, F. 2010. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach The Information Content of Forward-Looking Statements in Corporate Filings -A Naïve Bayesian Machine Learning Approach. *Source Journal of Accounting Research Journal of*

Accounting Research Journal of Accounting Research. 48(5):1049–1102. DOI: 10.HH/j.1475-679X.2010.00382.x.

Lin, D. 1998. Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology*. 57–63. DOI: 10.1.1.56.1687.

Loughran, T. 2018. Linguistic tone and the small trader: Measurement issues, regulatory implications, and directions for future research. *Accounting, Organizations and Society*. (March):0–1. DOI: 10.1016/j.aos.2018.03.001.

Loughran, T. & McDonald, B. 2011. When is a Liability not a Liability? *Journal of Finance*. 66(1):35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

Lovins, J.B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*. 11(June):22–31. Available: http://journal.mercubuana.ac.id/data/MT-1968-Lovins.pdf.

Luong, M.-t. & Manning, C.D. 2013. Better Word Representations with Recursive Neural Networks for Morphology. 104–113.

Malkiel, B.G. 1973. *A random Walk Down Wall Street*. ed.

Malo, P., Sinha, A., Takala, P., Ahlgren, O. & Lappalainen, I. 2013. Learning the roles of directional expressions and domain concepts in financial news analysis. *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*. 945–954. DOI: 10.1109/ICDMW.2013.36.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. 1–12. DOI: 10.1162/153244303322533223.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013b. DOI: 10.1162/153244303322533223.

Muller, C. 1999. Investor overreaction on the Johannesburg Stock Exchange. *Investment Analysts Journal*. 49(1):5–17. DOI: 10.1017/CBO9781107415324.004.

Noakes, M.A. & Rajaratnam, K. 2016. Testing market efficiency on the Johannesburg Stock Exchange using the overlapping serial test. *Annals of Operations Research*. 243(1-2):273–300. DOI: 10.1007/s10479-014-1751-y.

Osborne, M...F...M... 1962. Periodic Structure in the Brownian Motion of Stock Prices. *Operations reseach*. 10(3):345–379.

Oxford University Press. 2018. Available: https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P.P., Gurrin, C., Smeaton, A.F. & OHare, N. 2009. Topic-Dependent Sentiment Analysis of Financial Blogs. *International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*. 9–16. DOI: 10.1145/1651461.1651464.

Page, M.J. & Way, C.V. 1992. Stock Market Over-reaction: The South African Evidence. *Investment Analysts Journa*. 36:35–49.

Pang, B., Lee, L., Rd, H. & Jose, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceeding EMNLP '02 Proceedings of*

the ACL-02 conference on Empirical methods in natural language processing. (July):79–86.

Pecina, P. 2005. An extensive empirical study of collocation extraction methods. *Proceedings of the ACL Student Research Workshop on - ACL '05*. (June):13. DOI: 10.3115/1628960.1628964.

Phiri, A. 2014. Evidence from Linear and Nonlinear Unit Root Tests. 13(4):369–387.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program*. 14(3):130–7. DOI: 10.1108/00330330610681286.

Rani, S., Ramesh, B. & Anusha, M. 2015. Evaluation of stemming techniques for text classification. *Journal of Computer ...* . 43(3):165–171.

Rensburg, P. van & Robertson, M. 2003. Style characteristics and the cross-section of JSE returns. *Investment Analysts Journal*. 31(57):7–15. DOI: 10.1080/10293523.2003.11082444.

Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*. 60(5):503–520. DOI: 10.1108/00220410410560582.

Robertson, M., Firer, C. & Bradfield, D. 2000. Identifying and correcting misclassified South African equity unit trusts using style analysis.

Shannon, C.E. 1951. The redundancy of English. *Cybernetics; Transactions of the 7th Conference, New York: Josiah Macy, Jr. Foundation*. 248–272.

Smith, G., Jefferis, K. & Ryoo, H.J. 2002. African stock markets: Multiple variance ratio tests of random walks. *Applied Financial Economics*. 12(7):475–484. DOI: 10.1080/09603100010009957.

Sohangir, S., Wang, D., Pomeranets, A. & Khoshgoftaar, T.M. 2014. Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization. *Journal of Big Data*. DOI: 10.1186/s40537-017-0111-6.

Spärck Jones, K. 1972. A Statistical Interpretation of Term Specificity and its Retrieval. *Journal of Documentation*. 28(1):11–21. DOI: 10.1108/eb026526.

Tetlock, P.C. 2007. Giving Content to Investor Sentiment : The Role of Media in the Stock Market. 62(3):1139–1168.

Tetlock, P.C., Saar-Tsechansky, M. & Macskassy, S. 2008. More than words: Quantifying language to measure firms fundamentals. *The Journal of Finance*. 63(3):1437–1467. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2008.01362.x/full.

Tirunillai, S. & Tellis, G.J. 2012. Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance. *Marketing Science*. 31(2):198–215. DOI: 10.1287/mksc.1110.0682.

Zhang, X., Zhao, J. & LeCun, Y. 2013. Character-level Convolutional Networks for Text Classification. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 3057–3061. DOI:

10.1063/1.4906785.

Zhu, X. 2013. Persistent homology: An introduction and a new text representation for natural language processing. *IJCAI International Joint Conference on Artificial Intelligence*. 1953–1959.