Text Analysis in Finance: A Survey

Issam Aattouchi[1], Mounir Ait Kerroum[1] and Saida Elmendili[2]
[1] *Computer science research Laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco*
[2] *Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco*

Keywords: Finance analysis, NLP, DL, Financial Data Sources, Classification, Dictionaries, Word embeddings, Word2vec, Bayesian Classifier.

Abstract: The new advancements in the domains of artificial intelligence and machine learning, have provided a vast opportunity to provides a means to interpret external data correctly, learn from such data, and exhibit flexible adaptation. The current combination of computing power and complex algorithms has made fields such as Artificial Intelligence, and particularly its Deep Learning branch, very popular and influential. The financial sector is undergoing a significant transformation in light of the new emerging artificial intelligence technologies and the increasingly competitive markets. This article introduces a survey of research on the analysis of the recent financial news, presenting and comparing different sources of information, methods and models of content analysis in the financial markets (return on assets, volatility, interest rate, etc.). The purpose of this surveys is providing a state-of-the-art, on the application of sentiment analysis methods, in the finance domain, to provide systematic approach in decision making.

1 INTRODUCTION

The last breakthroughs in Artificial Intelligence and Machine Learning have started getting a lot of attention recently. Finance is one particular area where sentiment analysis models started getting traction, combining both computational power and complex algorithms can now perform tasks that were historically assigned to humans, such as object recognition or speech. The inner force of these techniques results in their astonishing capacities of performing and dealing with unstructured data.

Natural Language Processing (NLP), as one of the most promising fields in Machine Learning and maybe its hottest area, has managed to resolve many tasks that seemed extremely difficult in the past: Information Retrieval, Information extraction, Machine translation, Spam filter, Sentiment analysis, etc. The main difficulty when applying NLP is the high dimensionality of text data. Hence, many methods are developed to quantify text into numerical vectors.

In finance, the flows of news update the investor's understanding and influence their sentiment and hence, shapes the financial markets (Asset returns, Volatility, Interest rates, etc.). Thus, one of the most critical challenges that an investor has to overcome is to extract useful information from text data to use it in decision making. More generally, obtaining an accurate forecast of stock market movement is the engine of financial prediction. Many algorithms are used for this purpose. News Analytics in Finance is the process of integrating quantified financial information into the analysis procedure, in order to improve the perception of the market by the investor.

2 DATA SOURCES

2.1 Data Sources Type

Investors receive data/news from different sources. However, some sources are far more reliable than others and cannot treated equally. In literature, we can distinguish four classes of news sources:

Everyday news: This includes mainstream Media which are broadcasted via newspapers, radio and television.

Pre-news: This includes sources that reporters process to give common news (e.g Securities and Exchange Commission reports).

Social media: These sources are less reliable than

Aattouchi, I., Ait Kerroum, M. and Elmendili, S.
Text Analysis in Finance: A Survey.
DOI: 10.5220/0010728200003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 50-59
ISBN: 978-989-758-559-3

the previous ones. They may contain dangerous inaccurate information due to the simplicity of spreading fake news. some financial statements are divided into two main categories (Mitra & Mitra, 2011):

Regular announcements: These are all types of news from reliable sources.

Event-driven announcements: News, social media streams and rumours.

2.2 Financial Data Sources

In this survey, we present the data sources most commonly used in Text Analysis in Finance:

Sedar - Public Company Documents Search (SEDAR, 2015): The majority of relevant financial information: quarterly earnings releases, financial statements, and recent stock prices, are all published in the "Investor Relations" section of company websites.

Edgar - Company Name Search (SEC.Gov | EDGAR - Search and Access, 2019): information relating to the annual reports on the security of public Customer companies is also available on the websites of the security regulators.

Financial Statement Data Sets (SEC.Gov | Financial Statement Data Sets, 2009): This data is extracted from corporate financial reports using eXtensible Business Reporting Language (XBRL). Compared to the numeric and textual data sets of the financial industry and their notes, the financial statement data sets are the most reliable.

Financial Performance Data (Industry Canada) (Financial Performance Data, 2015): This benchmarking tool for SMEs is based on Canadian tax return data. It includes industry averages (by NAICS) for income statement and sheet objects, profitability information and financial ratios.

It covers small and medium-sized businesses, with incomes between $30,000 to $5,000,000.

World Bank Open Data (Data Catalog | Data Catalog, 2015):On this site, you can find datasets such as:

- Unit-level data sets and indicators are a series of counts collected at regular periods over time.
- Unit-level data collected from many sources: sample studies, population censuses, and administrative systems.
- Data containing explicit geographic positioning information, in raster format or vector.

IMF Data (IMF Data, 2015): The International Monetary Fund publishes data in: the international finances, debt rates, currency exchange reserves, commodity payments and funds.To export some datasets, you might be required to register, user is able to export data as Excel, PDF, image, Powerpoint, .html and .mht files without registering in some cases.

European Union Open Data Portal (Data.Europa.Eu, 2015): The website funded by the European Union- is a comprehensive portal, featuring datasets from various sectors, which are free and easy to access. User don't need to register to download any datasets either.

Financial market statistics (Bank of Canada, 2015): Various Bank of Canada and financial market statistics such as bond yields, treasury bills, corporate paper rate, publishes data, from 1991 to present.

Bank of Canada - Interest Rates (Interest Rates, 2015): The Bank of Canada is a financial institution in Canada. Valet Web Services provides programmatic access to financial data from around the world. User can get financial data and information from the Bank of Canada using the Valet API, such as daily exchange rates. Historical time series data are available in Statistics Canada's CANSIM database.

FRED: Federal Reserve Economic Data (Federal Reserve Economic Data | FRED | St. Louis Fed, 2015): Federal Reserve Economic Data is a database maintained by the Research Division of the Federal Reserve Bank of St. Louis that contains over 765,000 economic time series from 96 sources. These open Data include money, finance and banking, national accounts, population, jobs and labour markets, production and business operations, prices, international data.

Quandl (Quandl, 2015): Analysts from the world's biggest hedge funds, asset managers, and investment banks use Nasdaq's Quandl platform. Investment professionals will benefit from this outstanding collection of financial, economic, and alternative datasets.

Table I summarizes these Data Sources with their own characteristics:

Table 1: Data Sources characteristics.

Data Sources	Type of Data/Size
SEDAR	Annual Information Form, Annual Report, Annual Statement of Payments, Financial Statements, Fund Facts, Fund Summary, Management's Discussion & Analysis (MD&A), Marketing Material, Material Change Report, Real Estate Offering Document, Report of Exempt Distribution.

EDGAR	The most common Documents include: 10-K (Annual report), 10-Q (Quarterly report), 8-K (Current report)..
Financial Statement Data Sets	Annual report, summary of the company's operations, events, an acquisition, bankruptcy, disposal of assets.
Financial Performance Data (Industry Canada)	Reports feature the number of businesses in the selected industry, detailed financial data on revenues and expenses industry average's for income statements and balance sheet items ..
World Bank Open Data	Time Series (14 681) Datasets and Indicators level data, Microdata (3 441) Unit-level data, Geospatial (777) Data.
IMF Data	World Economic Outlook (WEO) Databases, Statistical Data, Dataset Portals, Financial Data/Rates, Regional office document, financial reports.
EU Open Data Portal	open data sets across EU policy domains, including the economy, employment, science, environment and education.
Financial market statistics	Report of monetary authorities, Report of business performance and ownership, financial statements and performance, financial markets.
Bank of Canada	Measures the cost of overnight general collateral funding in Canadian dollars using Government of Canada treasury bills and bonds as collateral for repurchase transactions, Money Market Yields, 10-Year Lookup U.S. Prime Rate Charged by Banks, Federal Funds Rate, Commercial Paper.
Federal Reserve Economic Data	Interest Rates (1,000+), Exchange Rates (160+), Monetary Data (990+), Financial Indicators (2,000+), Banking (2,000+), Business Lending (2,300+), Foreign Exchange Intervention (3+), National Income & Product Accounts (13,000+), Federal Government Debt (3+), Flow of Funds (38,000+), U.S. Trade & International Transactions (410+), Banking and Monetary Statistics, 1914-1941 (1,200+), Daily Federal Funds Rate...
Quandl	Equity Prices, Equity Fundamentals, Equity Earnings, Estimates, Analyst Ratings, Futures, Economics, FX and Rates, Payment card transactions, Satellite imagery / GPS

The data sources cited in this paper are open, so you just need to choose the appropriate source to answer a specific financial problem.

3 TEXT PRESENTATION METHODS

The NLP faces a real challenge of the very high dimensionality of textual data. Thus, transform textual data into quantitative numerical data is a challenging task. In fact, the text was examined on three levels: text, content and the context.

Given that the leading role of the research is to transform the text into useful information, the complexity increases considerably with a considered level. Therefore, the primary applications are those that treat news simply as text.

Depending on the type of application used, textual data is represented in very different ways (el MENDILI, 2020). Note that for an investor, the objective is to assign a relevant score to the news.

It could simply use the set of words that define the sentences and take into account the context and the second degree of these sentences. It looks pretty natural that the complexity increases for the last choice.

3.1 Dictionaries

The simplest way to represent a text is the dictionary. It consists of taking the set of all possible words, labelling them as positive/negative and using the resulting group to assign a sentiment score to sentences simply. Note that finance can define a particular lexicon consisting only of financial words.

The General Inquirer (GI) integrated dictionary (Gilman, 1968) and the DICTION text analysis program (Wayback Machine, 2002) are the first two popular lexicons used in financial news analysis. Most researchers used the Harvard dictionary and The General Inquirer word lists, because it was the first lists readily available (LOUGHRAN & MCDONALD, 2011) (Doran et al., 2010; Engelberg, 2008; Ferris et al., 2012; Henry & Leone, 2009; Ozik & Sadka, 2012; TETLOCK, 2007; TETLOCK et al., 2008).

Research has shown that more than 73.8% of the number of words declared negative in the list proposed in / Harvard are words that are not negative, if used in the financial field (LOUGHRAN & MCDONALD, 2011).

To remedy this problem, Loughran and McDonald's developed a new dictionary of 3,532 unique lists of financially binding Words (Ding et al., 2017), using the US Security and Exchange Commission portal, from 1994 to 2008. The Loughran-McDonald Financial Sentiment Dictionary (LMFSD) is used by several subsequent research (Azmi Shabestari et al., 2019; Jangid et al., 2018; Kearney & Liu, 2013).

3.2 One-hot Encoding of Words and Characters

Text is the most common form of sequence data. It can be whether as a sequence of words or characters. The most common (and essential) numerical representation of text is the one-hot encoding.

The idea is elementary. We consider a space of dimension N =Number of words in a dictionary, and we represent each word as a binary vector of size N (The vector is all zeros except one index, which determines the order of the word in the dictionary).

A sentence is then naturally represented after tokenization (split words in a sentence by blank or punctuation) by a vector of size N. Each entry represents the number of occurrences of i-word of the dictionary in the sentence.

We can assign a weight to each word describing its polarity (This leads to a new vector with the same non-zeros elements but with a weighted value representing the cumulative effect) (Vargas et al., 2017).

3.3 Word Embeddings

The vectors obtained by one-hot encoding are sparse, binary and ultra-high dimensional. One can imagine representing text data with dense vector to gain both complexity and effectiveness.

The simplest way to associate a dense vector to a word is to choose the vector at random. However, the resulting space is also random and do not represent the text's inherent structure. To gain incoherence, one can say that the geometric distance between two vectors must reflect the semantic (or contextual) relationship between words.

For example, we may want to have similar distances between synonyms (or antonyms, actually) or a vector which enables us to pass from single to plural words. Precisely, the resulted space should somehow map reasonably the human language with numerical vectors. Similarly, the literature called semantic space models of meaning or vector space (Akbik et al., 2018; Harris, 1954; Landauer & Dumais, 1997; Peters et al., 2018; Sahlgren, M, 2006; Turney & Pantel, 2010).

3.4 Word2Vec (Skip-gram model)

Google released Word2Vec in 2013 as a useful tool. It includes two (nearly identical) models: Skip-gram and Continuous Bag of Words (Mikolov et al., 2013; Mikolov, Sutskever, et al., 2013).

Its goal is to convert each pair of words into a calculable numeric vector while maintaining the degree of resemblance and analogy between them. Word2vec has been the main spark of NLP since its publication, and it is now widely employed in the discipline.

We will introduce the Skip-gram model to understand its functioning. In the skip-gram, our primary focus is on one word (called Centre), and we try to predict words that will appear around it (called backgrounds). More precisely, we are mainly interested in the conditional probabilities of each set under the given centre word.

4 SENTIMENT CLASSIFICATION

We investigate below Das and Chen algorithm (Das & Chen, 2007) to classify news into positive/negative and scoring them. The goal is to detect the investor's sentiment from stock message boards.

To implement their algorithm, they needed to use specific databases (Mitra & Mitra, 2011):

Dictionary: Used to determine the nature of words (Adjective, adverb, etc.).

Lexicon: A collection of finance words.

Grammar: The training corpus of base messages used. Note that the lexicon and grammar define here the context of statements. We cite information about five classifiers:

4.1 Naive Classifier (NC)

Naïve Classifier is the primary language-dependent classifier. Given a sentence/news item, we can assign a simple sentiment score created on the lexical word count. Say, for instance, that if the number of positive/optimistic words exceeds the number, of negative/pessimistic words, the message is labelled positive and vice versa.

We can also naively assign a sentiment score (difference between the number, of positive and

negative words) (Shihavuddin et al., 2010).

4.2 Vector Distance Classifier

Using our dictionary of words D, we consider here a one-hot encoding representation of words. With a D-dimensional space.

Thus, we can represent each sentence by a vector (each entry represents simply the number of occurrences of the corresponding word in D).

Das and Chen (2007), has used this algorithm which is built on the same principle used by search engines.

4.3 Discriminant-based Classification

The previous classifiers did not distinguish between words. However, some words, are far more potent than others. As seen in (Natural Language Processing Tested in the Investment Process through New Partnership | J.P. Morgan, 2018), sentimental words that are frequently repeated/used (We do not consider linkage words) must be treated differently by the classifier.

Hence, this leads us to redefine the counting-based method by weighting words, the commonly used tool is Fisher's Discriminant.

4.4 Adjective-adverb Classifier

Another approach for Classification is to consider filter words, so that we take a subset of sentences that includes only segments with high emphasis (i.e., adverbs and adjectives).

Using dictionaries like CUVOALD (Computer Usable Version of the Oxford Advanced Learner's Dictionary), one can quickly build programs performing sentimental counting across these specific lexicons.

4.5 Bayesian Classifier

Bayesian Classification is the most famous technique used in practice. We can find its applications in almost all AI-powered fields as (Mbadi, S, 2018).

The main idea is to compute prior probabilities using Datasets and then calculate posteriors ones by the Bayes formula. The universality and the simplicity of this approach are maybe the reasons behind its success.

In the context of text classification, the classifier is trained on pre-trained corpus and try to learn some statistical features of this text. It uses word-based probabilities, and thus Bayesian classifier is an independent language, since he sees sentence as a simple bag of words.

4.6 Support Vector Machines

Support Vector Machine or simply SVMs are widely used classifiers similar to cluster analysis. However, they are mainly used to very-high-dimensional spaces, which makes sense in the context of a text represented by ultra-high dimensional spaces via one-hot encoding.

The main idea of SVMs is to find, given a training corpus, that best separate classes. Classification methods for sentiment extraction are applied to news or tweets datasets (Ghiassi et al., 2013; Li et al., 2009). Li et al. (2009) use several machine learning methods classifiers to obtain the sentiments from Tweets. Its implementation shows that the Support Vector Machine classifier is more efficient than the naive Bayes classifier and decision trees.

5 UNSUPERVISED FINANCIAL NETWORKS ANALYTICS

Rather than basing our analysis on text messages and attempting to give emotion scores or forecast market movements, we might consider the market as a vast network. That is, modelling interconnected stocks using flows of information.

It appears that strongly connected stocks reactions are highly correlated. Thus, this graph modelling will allow us to extract patterns/features that is used inferentially. However, how should we define our network? What do we mean by correlated stocks? There are many possible implementations. For example, Das and Sisk (2005), built a network based on the number of standard handles forwarded to pairs of stock.

5.1 Centrality

After modelling the problem with networks, the field of graph theory lends itself naturally. Centrality measures are among the most widely used indices based on network data. A node is called central, if it has strong connections (directly or indirectly), to other nodes.

Ambrus et al. (2018), Ambrus and Elliott (2020), lighten how this measure, can be computed, they explain that risk-sharing in shape studies where transfers between pairs of agents can only depend on

the income generated by the agents to which they are both linked in a pre-existing network.

5.2 Communities

Another essential feature of graphs is communities. More precisely, a community is a cluster of nodes, which can be detected using classical algorithms: Lloyd's heuristic or the walk trap algorithm (Note that the problem of clustering is NP-hard, these algorithms do not always give the optimal solutions but are very practical).

Finally, communities tend to react uniformly due to the strong presence of contagion phenomena inside them. As proposed by (Schweitzer et al., 2009), currently, the risk system literature has started to use the topology of financial networks to measure systemic risk. The complexity of the economic system is perhaps extended with new paradigms using the different economic networks (Billio et al., 2012; Creel et al., 2015; Diebold & Yilmaz, 2015; Hautsch et al., 2014).

5.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet allocation (David M Blei et al., 2003), describe the Topics in a word space and describe the documents in a Topics space, based on calculating the DIRICHLET distribution of words for each Topic and the calculating of the DIRICHLET distribution of Topic for each document. The space of the projection of a document is a latent space of low dimension.

More specifically, LDA is a three-level hierarchical Bayesian model, in which each item is considered a finite mixture of latent topics. This algorithm is used for different purposes as: Topics extraction, Reduction of dimension, novelty detection, summarization, similarity and relevance judgments, etc.

The algorithm's goal is to depict short descriptions of texts (or any other collection of data) that enable processing of the text corpora while preserving the essential statistical relationships that are useful for basic tasks. LDA generally works best due to its generative nature. LDA is different in how it considers documents as a mixture of topics and topics as a distribution over words as shown in (Asadi Kakhki et al., 2018; Feuerriegel et al., 2016).

6 DEEP LEARNING IN FINANCE

Deep learning is a subfield of machine learning, which involves computers processing large amounts of data using artificial neural networks that mimic the structure of the human brain.

Whenever new information is incorporated, the existing connections between neurons are subject to change and expand, this operation allows the system to learn things without human intervention, autonomously, while improving the quality of its decision-making and forecasts. Among the different techniques, Deep Multilayer Perceptron (DMLP), CNN, RNN, LSTM, Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), and Autoencoders (AEs).

About 15 studies on the application of deep learning to a specific finance domain are examined in this part. Figure 1 depicts the link between commonly used DL models:

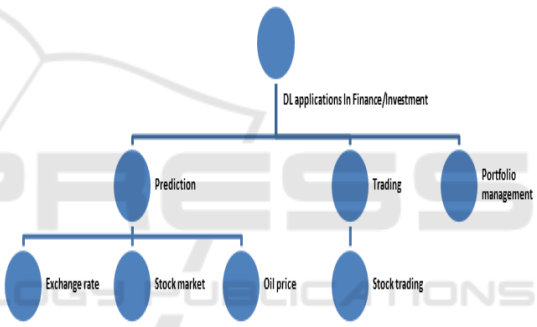

Figure 1: link between commonly used DL models

We give a review of some DL model's applicability in two financial fields: Prediction and trading.

6.1 Prediction

We look at this subject from three aspects of prediction: **exchange rates**, **stock markets**, and **oil prices**:

Galeshchuk and Mukherjee (2017), assert that a single hidden layer NN or SVM performs worse than a basic model such as moving average (MA). However, they discovered that due of the consecutive layers of DNN, CNN could obtain higher classification accuracy in predicting the direction of the change in exchange rate.

Ravi et al. (2017), use MLP (FNN), chaos theory, and multiobjective evolutionary algorithms to create

a hybrid model. Their Chaos+MLP + NSGA-II model has a remarkably low mean squared error (MSE) of 2.16E-08.

Kim et al. (2015), suggest a deep convolutional neural network architecture for predicting whether or not a customer is suitable for bank telemarketing. The number of layers, learning rate, initial value of nodes, and other factors that should be established when building a deep convolutional neural network are discussed and presented.

Lee et al. (2017), employed FFNN, Support Vector Regressor (SVR), and RBM-based DBN to analyze the revenues and profitability of organizations. They present a corporate performance prediction model based on deep neural networks that employs financial and patent metrics as predictors. An unsupervised learning phase and a fine-tuning phase are included in the proposed paradigm.

A constrained Boltzmann machine is used in the learning phase. The fine-tuning step employs a backpropagation algorithm and a recent training data set that reflects the most recent trends in the association between predictors and business performance. When compared to the SVR-based model, the suggested approach has a prediction error reduction of 1.3–1.5 times. The suggested model, in particular, outperforms SVR-based models in predicting the performance of companies that experience earnings surprises or shocks. This demonstrates that the proposed model has long-term predictability, whereas general prediction models exhibit a decline in prediction accuracy over time.

Table II summarizes these models with their own characteristics:

Table 2: Prediction model's characteristics.

Source	Method Performance	Environment	Feature Set	accuracy
Galeshchuk and Mukherjee (2017)	CNN	-	Exchange rate	92.62% (for USD/JPY 83.72% for GBP/USD)
Ravi et al. (2017)	Hybrid (Chaos, MLP, MOPSO)	MATLAB/Gretl	Exchange rates	99.91%
Kim et al. (2015)	DCNN	-	Bank marketing	76.70%
Lee et al. (2017)	FFNN+SVR +RBM	Boltzmann machine	Revenues organizations	90%

According to several papers, only a hybrid paradigm might perform better. These approaches bring together effectiveness and performance, to get promising results.

6.2 Trading

Algorithmic trading (or Algo-trading) is defined as the use of algorithmic models to make buy-sell decisions. These decisions might be based on simple rules, efficient processes, mathematical models or complicated techniques, as in machine/deep learning.

Karaoglu and Arpaci (2017) present a technique for detecting trading signals based on a dynamic threshold selection, by combining many different rules-based systems and augmenting them using the Recurrent Neural Network algorithm.

Recurrent Neural Networks (RNNs) learn the connection weights of subsystems with arbitrary input sequences, making them ideal for time series data. With the purpose of detecting probable excessive movements in a noisy stream of time series data, the suggested model is based on Piecewise Linear Representation and Recurrent Neural Network. To discover irregularities, they employ an exponential smoothing approach. The tests revealed that this model delivers successful trading data results, but its scalability needs to be improved.

S. Wang et al. (2017), present a unique State Frequency Memory (SFM) recurrent network, to record multi-frequency trading patterns from historical market data and create long and short-term predictions over time. Stock prices are determined by, short and/or long-term commercial and trading activity that represent various trading patterns and frequency. These patterns, are frequently elusive, because they are influenced by a variety of uncertain political economic elements in the real world, such as corporate performance, government regulations, and even breaking news that spreads across markets.

Deng et al. (2017), employed Fuzzy Deep Direct Reinforcement Learning (FDDR) to predict stock prices and provide trading signals. They present a recurrent deep neural network (NN) for the modeling and trading of real-time financial signals. The proposed model, interacts with deep representations and makes trading decisions in an unknown environment to acquire the ultimate rewards.

Table III summarizes these models with their own characteristics:

Table 3: Algo-trading model's characteristics.

Source	Method Performance	Environment	Feature Set	accuracy
Karaoglu and Arpaci (2017)	LSTM, PLR, RNN	Big Data/SPARK	Stock Exchange	98.4%
Zhang et al. (2017)	SFM recurrent network	-	Stock Price	88.9%
Deng et al. (2017)	RL + Fuzzy Deep Direct Reinforcement Learning (FDDR),DMLP	Keras	Price Data	97%

The majority of Algo-trading research focused on predicting stock or index prices. Meanwhile, LSTM was the most used DL model.

7 CONCLUSION

Finance has long been one of the most researched domains for deep learning.

Algorithmic trading, Stock market forecasting, portfolio allocation, credit risk assessment, asset pricing, and the derivatives market are among the areas where deep learning researchers have been working on developing models, that can provide real-time working solutions for the financial industry.

Indeed, unsupervised methods and in-depth learning are essential tools for analyzing financial news. In our study, we described different techniques for extracting useful information. We have presented tools that perform feature extraction from a corpus. And as a synthesis of this study, we found that using a hybrid approach that brings together the effectiveness of several methods will get promising results. In our future paper, it is relevant to propose a new framework applied to real problems in text analysis in finance.

This approach will therefore be implemented and tested in practice on a case study, to present the experimental strengths.

REFERENCES

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. 1638-1649. https://aclanthology.org/C18-1139

Ambrus, A., & Elliott, M. (2020). Investments in social ties, risk sharing, and inequality. The Review of Economic Studies, 88(4), 1624–1664. https://doi.org/10.1093/restud/rdaa073

Ambrus, A., Gao, W., & Milan, P. (2018). Informal Risk Sharing with Local Information. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.3220524

Asadi Kakhki, S. S., Kavaklioglu, C., & Bener, A. (2018). Topic Detection and Document Similarity on Financial News. Advances in Artificial Intelligence, 322–328. https://doi.org/10.1007/978-3-319-89656-4_34

Azmi Shabestari, M., Moffitt, K., & Sarath, B. (2019). Did the banking sector foresee the financial crisis? Evidence from risk factor disclosures. Review of Quantitative Finance and Accounting, 55(2), 647–669. https://doi.org/10.1007/s11156-019-00855-y

Bank of Canada,. (2015). Bank of Canada,. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010013901

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104(3), 535–559. https://doi.org/10.1016/j.jfineco.2011.12.010

Creel, J., Hubert, P., & Labondance, F. (2015). Financial stability and economic performance. Economic Modelling, 48, 25–40. https://doi.org/10.1016/j.econmod.2014.10.025

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

Das, S. R., & Sisk, J. (2005). Financial Communities. The Journal of Portfolio Management, 31(4), 112–123. https://doi.org/10.3905/jpm.2005.592103

Data Catalog | Data Catalog. (2015). World Bank. https://datacatalog.worldbank.org

data.europa.eu. (2015). Europa Data. https://data.europa.eu/data/datasets?locale=en

David M Blei, Andrew Y Ng, & Michael I Jordan. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. IEEE Transactions on Neural Networks and Learning Systems, 28(3), 653–664. https://doi.org/10.1109/tnnls.2016.2522401

Diebold, F. X., & Yilmaz, K. (2015). Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring (Illustrated ed.). Oxford University Press.

Ding, Y., Yu, C., & Jiang, J. (2017). A Neural Network Model for Semi-supervised Review Aspect Identification. Advances in Knowledge Discovery and Data Mining, 668–680. https://doi.org/10.1007/978-3-319-57529-2_52

Doran, J. S., Peterson, D. R., & Price, S. M. (2010). Earnings Conference Call Content and Stock Price: The Case of REITs. The Journal of Real Estate Finance and Economics, 45(2), 402–434. https://doi.org/10.1007/s11146-010-9266-z

el MENDILI, S. (2020). Towards a Reference Big Data architecture for sustainable smart cities. International Journal of Advanced Trends in Computer Science and Engineering, 9(1), 820–827. https://doi.org/10.30534/ijatcse/2020/118912020

Engelberg, J. (2008). Costly Information Processing: Evidence from Earnings Announcements. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1107998

Federal Reserve Economic Data | FRED | St. Louis Fed. (2015). Federal Reserve. https://fred.stlouisfed.org

Ferris, S. P., Hao, G. Q., & Liao, S. M. Y. (2012). The Effect of Issuer Conservatism on IPO Pricing and Performance*. Review of Finance, 17(3), 993–1027. https://doi.org/10.1093/rof/rfs018

Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation. 2016 49th Hawaii International Conference on System Sciences (HICSS). Published. https://doi.org/10.1109/hicss.2016.137

Financial performance data. (2015). FP DATA. http://www.ic.gc.ca/eic/site/pp-pp.nsf/eng/home

Galeshchuk, S., & Mukherjee, S. (2017). Deep networks for predicting direction of change in foreign exchange rates. Intelligent Systems in Accounting, Finance and Management, 24(4), 100–110. https://doi.org/10.1002/isaf.1404

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16), 6266–6282. https://doi.org/10.1016/j.eswa.2013.05.057

Gilman, R. C. (1968). The General Inquirer: A Computer Approach to Content Analysis.Philip J. Stone , Dexter C. Dunphy , Marshall S. Smith , Daniel M. Ogilvie. American Journal of Sociology, 73(5), 634–635. https://doi.org/10.1086/224539

Harris, Z. S. (1954). Distributional Structure. WORD, 10(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hautsch, N., Schaumburg, J., & Schienle, M. (2014). Financial Network Systemic Risk Contributions. Review of Finance, 19(2), 685–738. https://doi.org/10.1093/rof/rfu010

Henry, E., & Leone, A. J. (2009). Measuring Qualitative Information in Capital Markets Research. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1470807

IMF Data. (2015). IMF. https://www.imf.org/en/Data

Interest Rates. (2015). Bank of Canada. https://www.bankofcanada.ca/rates/interest-rates/

Jangid, H., Singhal, S., Shah, R. R., & Zimmermann, R. (2018). Aspect-Based Financial Sentiment Analysis using Deep Learning. Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. Published. https://doi.org/10.1145/3184558.3191827

Karaoglu, S., & Arpaci, U. (2017). A Deep Learning Approach for Optimization of Systematic Signal Detection in Financial Trading Systems with Big Data. International Journal of Intelligent Systems and Applications in Engineering, Special Issue(Special Issue), 31–36. https://doi.org/10.18201/ijisae.2017specialissue31421

Kearney, C., & Liu, S. (2013). Textual Sentiment Analysis in Finance: A Survey of Methods and Models. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.2213801

Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. (2015). Predicting the success of bank telemarketing using deep convolutional neural network. 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR). Published. https://doi.org/10.1109/socpar.2015.7492828

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240. https://doi.org/10.1037/0033-295x.104.2.211

Lee, J., Jang, D., & Park, S. (2017). Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability. Sustainability, 9(6), 899. https://doi.org/10.3390/su9060899

Li, N., Liang, X., Li, X., Wang, C., & Wu, D. D. (2009). Network Environment and Financial Risk Using Machine Learning and Sentiment Analysis. Human and Ecological Risk Assessment: An International Journal, 15(2), 227–252. https://doi.org/10.1080/10807030902761056

LOUGHRAN, T., & MCDONALD, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Mbadi, S. (2018). Predicting Stock Market Movement Using an Enhanced Naïve Bayes Model for Sentiment Analysis Classification

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs]. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv:1310.4546 [cs, stat]. http://arxiv.org/abs/1310.4546

Mitra, G., & Mitra, L. (2011). The Handbook of News Analytics in Finance (1st ed.). Wiley.

Natural language processing tested in the investment process through new partnership | J.P. Morgan. (2018). J.P. Morgan. https://www.jpmorgan.com/news/natural-language-processing-tested-in-the-investment-process-through-new-partnership?source=cib_di_jp_mal0418

Ozik, G., & Sadka, R. (2012). Media and Investment Management. SSRN Electronic Journal. Published. https://doi.org/10.2139/ssrn.1633705

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Published. https://doi.org/10.18653/v1/n18-1202

Quandl. (2015). Quandl. https://www.quandl.com/

Ravi, V., Pradeepkumar, D., & Deb, K. (2017). Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. Swarm and Evolutionary Computation, 36, 136–149. https://doi.org/10.1016/j.swevo.2017.05.003

Sahlgren, M. (2006). The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., & White, D. R. (2009). Economic Networks: The New Challenges. Science, 325(5939), 422–425. https://doi.org/10.1126/science.1173644

SEC.gov | EDGAR - Search and Access. (2019, December 16). EDGAR. https://www.sec.gov/edgar/search-and-access

SEC.gov | Financial Statement Data Sets. (2009, January 1). FS Data Set. https://www.sec.gov/dera/data/financial-statement-data-sets.html

SEDAR. (2015). SEDAR. https://www.sedar.com/search/search_form_pc_en.htm

Shihavuddin, A., Mir Nahidul Ambia, Mir Mohammad Nazmul Arefin, Mokarrom Hossain, & Adnan Anwar. (2010). Prediction of stock price analyzing the online financial news using Naive Bayes classifier and local economic trends. 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE). Published. https://doi.org/10.1109/icacte.2010.5579624

TETLOCK, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

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. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, 141–188. https://doi.org/10.1613/jair.2934

Vargas, M. R., de Lima, B. S. L. P., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). Published. https://doi.org/10.1109/civemsa.2017.7995302

Wayback Machine. (2002). Rhetorica.Net. https://web.archive.org/web/*/rhetorica.net

Zhang, L., Aggarwal, C., & Qi, G. J. (2017). Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Published. https://doi.org/10.1145/3097983.3098117