AUTOMATED FREQUENCY-BASED STATISTICAL AND LINGUISTIC FEATURE PROCESS MODELS FOR FINANCIAL NEWS SENTIMENT CLASSIFICATION

By

SEPIDEH FOROOZAN YAZDANI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

October 2017

FSKTM 2017 58

DEDICATION

"Dedicated to those who kept me on their shoulders"

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

AUTOMATED FREQUENCY-BASED STATISTICAL AND LINGUISTIC FEATURE PROCESS MODELS FOR FINANCIAL NEWS SENTIMENT CLASSIFICATION

By

SEPIDEH FOROOZAN YAZDANI

October 2017

Chairman : Masrah Azrifah Azmi Murad, PhD
Faculty : Computer Science and Information Technology

This thesis utilizes sentiment classification task within the field of artificial intelligence for financial news using the combination of machine learning, linguistics, and statistical methods. The motivation for this approach comes from human emotion and vital information that lies in the financial news like news reports and impacts on the market. In recent years, a huge amount of this information is accessible for investment and research analysis in a text format where investors and researchers can simply get access to the desired information through a variety of channels on the Internet.

Despite the studies conducted in automated sentiment classification of financial news, there are still challenges in some parts of text mining and financial news classification that concerns feature extraction, feature selection, and classification processes. Most existing literature on sentiment financial news typically relies on very simple linguistic features, such as Bag-of-Words (BOW) in which each piece of news is represented using distinct words with frequencies as a feature type, and only a few numbers of the studies have employed complicated approaches. Obviously, not all words are needed to reflect a given text. The primary downside of the BOW or unigrams is the huge number of linguistic features that it produces. The secondary downside is that linguistic features have too much information to become features while it is not clear which ones are important to the sentiment of financial news classification. Furthermore, since the extraction of words is based on their high frequency, typically low frequency-based linguistic features can be worth ignored. This research proposes two feature process models, Ngram-based and the NgramPOS-based models for the sentiment classification of financial news.

The Ngram-based model utilizes statistical approaches for feature processing in order to classify financial news. This high frequency-based model combines unigrams and bigrams along with Term Frequency-Inverse Document Frequency (TF-IDF)

(unsupervised feature weighting) while applying Document Frequency (DF) method with a certain threshold as dimensionality reduction method since it is suitable for high dimensional feature space.

NgramPOS-based model is able to enhance the performance of feature processing in Ngram-based model. NgramPOS-based model employs a combination of statistical and linguistic approaches to extract sentiment information as features in order to classify financial news. This low frequency-based model extracts the combination of sentiment-rich words and phrases as unigrams and bigrams using the defined POS-based fixed patterns along with the binary weighting method and applies Principle Component Analysis (PCA) as an unsupervised method to reduce the dimension of the extracted feature space.

Both models utilized RBF Support Vector Machine (SVM) with optimized parameters (C, γ) to classify the financial news as positive and negative news. Experiments showed that the combination of features as unigram and bigram along with TF-IDF and binary feature weighting methods in both models leads to the best result in financial news classification among, diverse feature spaces, with different accuracy for two models as 97.34% and 67.19% respectively.
.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MODEL PROSES FITUR AUTOMASI BERASASKAN FREKUENSI STATISTIK DAN LINGUISTIK UNTUK SENTIMEN PENGKLASIFIKASIAN BERITA KEWANGAN

Oleh

SEPIDEH FOROOZAN YAZDANI

Oktober 2017

Pengerusi : Masrah Azrifah Azmi Murad, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Tesis ini menggunakan tugas klasifikasi sentimen dalam bidang kecerdasan buatan untuk berita kewangan menggunakan gabungan pembelajaran mesin, linguistik dan kaedah statistik. Motivasi untuk pendekatan ini berasal dari emosi manusia dan maklumat penting yang terletak dalam berita kewangan seperti laporan berita dan kesan pasaran. Dalam tahun-tahun kebelakangan ini, sejumlah besar maklumat ini dapat diakses untuk analisis pelaburan dan penyelidikan dalam format teks di mana pelabur dan penyelidik hanya dapat mengakses maklumat yang dikehendaki melalui pelbagai saluran di Internet.

Walaupun kajian yang dijalankan dalam klasifikasi sentimen automatik berita kewangan, masih ada cabaran di beberapa bahagian penambangan teks dan klasifikasi berita kewangan yang menyangkut pengekstrakan ciri, pemilihan ciri, dan proses klasifikasi. Kebanyakan kesusasteraan sedia ada mengenai berita kewangan sentimen biasanya bergantung pada ciri linguistik yang sangat mudah, seperti *Bag-of-Words* (BOW) di mana setiap bahagian berita diwakili dengan menggunakan kata-kata yang berbeza dengan frekuensi sebagai jenis ciri, dan hanya beberapa kajian telah menggunakan pendekatan rumit. Jelas sekali, tidak semua perkataan diperlukan untuk mencerminkan teks yang diberikan. Kelemahan utama BOW atau unigram adalah jumlah ciri linguistik yang dihasilkannya. Kelemahan sekunder adalah, ciri linguistik mempunyai terlalu banyak informasi untuk menjadi cirri, sementara ia tidak jelas mana satu yang penting untuk sentimen klasifikasi berita kewangan. Tambahan pula, sejak pengekstrakan kata-kata didasarkan pada frekuensi tinggi mereka, ciri linguistik yang berasaskan kekerapan rendah boleh diabaikan. Penyelidikan ini mencadangkan dua model proses ciri, iaitu model berasaskan Ngram dan model berasaskan NgramPOS untuk pengkelasan sentimen berita kewangan.

Model berasaskan Ngram menggunakan pendekatan statistik untuk memproses ciri untuk mengklasifikasikan berita kewangan. Model berasaskan frekuensi tinggi ini menggabungkan *unigrams* dan *bigrams* bersama-sama dengan Frekuensi Istilah Frekuensi Dokumen Songsang (TF-IDF) (pengawalan ciri tanpa pengawasan) semasa menggunakan kaedah Frekuensi Dokumen (DF) dengan ambang tertentu sebagai kaedah pengurangan dimensi kerana ia sesuai untuk ruang ciri dimensi yang tinggi.

Model berasaskan NgramPOS dapat meningkatkan prestasi pemprosesan ciri dalam model berasaskan Ngram. Model berasaskan NgramPOS menggunakan kombinasi pendekatan statistik dan linguistik untuk mengekstrak maklumat sentimen sebagai ciri untuk mengklasifikasikan berita kewangan. Model berasaskan kekerapan rendah ini mengekstrak kombinasi kata-kata dan frasa yang kaya dengan sentimen sebagai *unigrams* dan *bigrams* menggunakan corak tetap berdasarkan POS yang ditetapkan bersama dengan kaedah penimbang binari dan menggunakan Analisis Komponen Prinsip (PCA) sebagai kaedah tanpa pengawasan untuk mengurangkan dimensi ruang ciri yang diekstrak.

Kedua-dua model menggunakan Mesin Vektor Sokongan RBF (SVM) dengan parameter yang dioptimumkan (C, γ) untuk mengklasifikasikan berita kewangan sebagai berita positif dan negatif. Eksperimen menunjukkan bahawa gabungan ciri-ciri seperti unigram dan bigram bersama dengan kaedah pemberat ciri TF-IDF dan binari dalam kedua-dua model membawa kepada hasil terbaik dalam klasifikasi berita kewangan di kalangan ruang ciri yang berbeza, dengan ruang ciri yang berbeza, dengan ketepatan yang berbeza untuk dua model masing-masing sebagai 97.34% dan 67.19%.

ACKNOWLEDGEMENTS

First of all, praise is to “Allah” the cherishers, and the sustainers of the world for giving me strengths, health and determination to complete this thesis. I wish to express my deep and sincere appreciation to the chair of my committee **Associate Professor Dr. Masrah Azrifah Azmi Murad** for her valuable ideas and support during the course of my thesis and also for the direction and guidance provided during the entire period of my studies without which, this thesis would not be possible. My deepest gratitude goes to my committee members, **Dr. Nurfadhlina Mohd Sharef, Ph.D, Prof. Yashwant Prasad Singh, Prof,** and **Dr. Ahmed Razman Abdul Latif** for their valuable guidance and advice throughout my study period at UPM.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Masrah Azrifah Azmi Murad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Nurfadhlina Mohd Sharef, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Ahmed Razman Abdul Latif, PhD
Senior Lecturer
Putra Business School
Universiti Putra Malaysia
(Member)

Yashwant Prasad Singh, PhD
Professor
Manav India Colege of Engineering India
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate Student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property of the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- There is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ________________________ Date: ______________________

Name and Matric No.: Sepideh Foroozan Yazdani,

Declaration by Members of Supervisory Committee

This is to confirm that:

- The research conducted and the writing of this thesis was under our supervision;
- Supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ________________________________
Name of
Chairman of
Supervisory
Committee: Masrah Azrifah Azmi Murad

Signature: _________________________________
Name of
Member of
Supervisory
Committee: Nurfadhlina Mohd Sharef

Signature: _________________________________
Name of
Member of
Supervisory
Committee: Yashwant Prasad Singh

Signature: _________________________________
Name of
Member of
Supervisory
Committee: Ahmed Razman Abdul latif

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAPL	Apple
AMZN	Amazon
ANNs	Artificial Neural Networks
BOW	Bag-of-Words
CHI	Chi-Square
CNG	Character N-gram
DF	Document Frequency
DT	Decision Tree
EMH	Efficient Market Hypothesis
GI	General Inquirer
GOOG	Google
IG	Information Gain
KCV	k-fold Cross-Validation
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
ME	Maximum Entropy
MI	Mutual Information
ML	Machine Learning
MMH	Maximum Margin Hyperplane
MSFT	Microsoft
NASDAQ	National Association of Securities Dealers Automated Quotations
NB	Naïve Bayes
NE	Named Entity
NPs	Noun Phrases
NYSE	New York Stock Exchange
PCA	Principal Component Analysis
POS	Part of Speech tag
RBF	Radial Basis Function
SGD	Stochastic Gradient Descent

SOM	Self Organizing Map
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
VSM	Vector Space Model
WSD	Word Sense Disambiguation

CHAPTER 1

INTRODUCTION

Financial news is considered as a significant factor to evaluate stock price by analysts and investors. Since the news conveys new information about the firm's fundamentals and qualitative information, therefore, financial documents can affect stock returns. Moreover, Tetlock (2007) showed the qualitative textual impact stock prices. According to the Efficient Market Hypothesis (Fama, 1965), all available information on stocks are reflected in market prices. Hence, news, particularly financial news including corporate, news article and, Internet message plays an essential role for investors when judging about stock price. This is because of the massive collection of the vital information contained in the news as the firm's fundamentals and prospect of other market participants. On the other hand, due to the rapid growth of financial news in the media for decades, it is difficult for investors to track and consider all available information, thus, automated classification of textual data seems vital.

Although the studies in automated classification of textual financial news are still in its infancy, but many attempts have been done in text mining to convert unstructured information to a usable format for classification task in machine learning. Nevertheless, there are still challenges in some parts of text mining and financial news classification such as feature extraction, feature selection, and classification processes.

Most existing literature (Koppel and Shtrimberg, 2006; Groth and Muntermann, 2011; Yu et al., 2013) on sentiment financial news typically rely on a very simple frequency-based textual presentation, such as *Bag-of-Words* (BOW) in which each piece of news is represented using distinct words with frequencies as a feature type. Moreover, some studies (Généreux et al., 2011; Zhai et al., 2011) have utilized unigrams, which have similar characteristics with distinct words as linguistic feature, since the extraction of both is based on their high frequency. On the other hand, a few other studies (Hagenau et al., 2013; Khadjeh Nassirtoussi et al., 2015) have employed complicated approaches for feature extraction such as noun phrases which is a type of low frequency-based feature.

For instance, a sample sentence is assumed as "Increase investment in universal economy" for how to capture the linguistic features. By dropping the word "in" as stop word the rest of the words ("Increase", "investment", "universal", "economy") will remain as features of the unigram type. While only the term "Increase" can be considered as a sentiment feature. Obviously, not all words are needed to reflect the polarity of the given text. The primary downside of the BOW or unigrams is the huge number of features that it produces using a big data set (Pestov, 2013). The secondary downside is that, linguistic features have too much information to become features while it is not clear which ones are important to sentiment of financial news classification.

By capturing sentiment expressions as bigram from the above sentence, a feature set including the terms “Increase investment”, “investment in”, “in universal”, “universal economy” is created. Surely, only term “Increase investment” expresses the polarity of the sentence. Certainly with a large volume of data, the process is encountered with a lot of the low frequency expressions which can be considered as an informative and sentiment feature. Since the extraction of words is based on their high frequency, typically low frequency-based linguist features that can be worth to sentiment classification are ignored.

As expressed, unfortunately most researches related to sentiment classification of financial news suffer from these significant weaknesses which were mentioned in above. All of these issues refer to the extraction and selection of specific and key terms as features to maintain the sense of dependency between terms which ultimately leads to an effective dimensionality reduction of features. For instance, extracting and selecting sentiment expressions as collocations and bigrams from the terms within the text can enhance the sentiment classification performance.

Moreover, some studies like Généreux et al. (2011) and Hagenau et al. (2013) applied some filter methods like Information Gain (IG) and Chi-Square (CHI) as feature selection. Although filter methods have lower risk of overfitting than wrapper and embedded methods, these techniques are scale-based methods that are not reliable for low-frequency terms (Dunning, 1993). Therefore, these methods cannot be reliable to extract the bigram or trigram features which have low frequency.

Khadjeh Nassirtoussi et al. (2014) presented a comprehensive systematic review on the text mining approaches used in the past for a market predictive purpose. The authors summarized all the machine learning algorithms employed in the financial news as follows: Support Vector Machine (SVM), Regression Algorithms, Naïve Bayes (NB), Decision Rules of Trees, Combinatory Algorithms, and Multi-algorithms. The studies by Sebastiani (2001) and Khadjeh Nassirtoussi et al. (2014) confirmed that SVM has been extensively and successfully applied as a textual classification and sentiment learning approach. However, none work of the existing works has applied the optimized classifiers while, Ageev and Dobrov (2003) have demonstrated that parameter optimization can essentially increase text classification performance.

As a summary of the discussion, the main focus of the existing works is to identify the polarity of news contents. Therefore, the existing approaches focused on the feature extraction and selection. The performance the related works can be improved if these weaknesses which have caused:

- Linguistic and statistical relevant features
- High Dimensionality of feature space
- Non optimized classifiers

can be addressed. This thesis focuses on these shortcomings and develops two feature process models for financial news sentiment classification based on statistical and linguistic approaches as low and high frequency feature process models to extract and select informative features to increase accuracy of polarity classification.

1.1 Motivation

With enlargement of Web 2.0 and the advent of social networks, blogs, and online news sources, analysts have to process enormous amounts of real-time unstructured data. For example, predicting the stock market trends and sentiment by the financial news is one of these instances. Financial news documents are produced in various types, such as recent earning statements, information about latest products, declaration of profits by a company, and similar issues. These sources usually contain the key but implicit factors which affect the stock market in different ways, for instance, effect on stock returns, volatility of price and also future firm earnings. Therefore, there is a vital need to discover the approaches to uncover the sentiment and polarity related expressions from these corpora of text. Obviously, this is a part in which sentiment analysis tool and its techniques can be employed to obtain the main concept of text by extracting important keywords from the financial news. Despite the large number of recent publications on sentiment mining in financial news, there are still many problems in this regard. Hence, it is necessary to have an improved technique for the extraction and determination of the sentiment and polarity of words, sentence, and phrase in order to extract the most representative expressions as features for sentiment analysis with high accuracy. This study focuses on the feature extraction and selection based on statistical and linguistic approaches to extract the prominent sentiment-rich features in order to enhance the sentiment classification performance of financial text.

1.2 Problem Statement

Analysis of the previous works on financial news classification indicates potential improvement for sentiment classification. According to the issues discussed in the introduction, most of the prior researches have solely relied on the BOW or unigrams (Koppel and Shtrimberg, 2006; Rachlin et al., 2007; Généreux et al., 2011; Yu et al., 2013) methods which unfortunately hardly capture any sentiment orientation of the content.

On the other hand, some early works have applied bigram model alone like Zhai et al., (2011), since it is perfect at capturing local dependencies. However, this model extremely suffers from data sparseness which is caused by the low frequency of bigrams. Unfortunately, since the extraction of words is based on their high frequency; typically low frequency-based linguist features that can be worth are ignored.

Moreover, due to the nature of textual data the news document classification will produce a high-dimensional feature space which most of features are irrelevant and leads to feature space sparsity. Most related works used two ways to handle this issue: the first approach employed common text preprocessing methods such as stemming and removing stopwords which were used by most researchers like (Pui Cheong Fung et al., 2003; Mittermayer, 2004; Joshi et al., 2016; Moore & Rayson 2017), and the second approach applied unsupervised selection method like IG or CHI that are strongly dependent on the training set and data set size (Généreux et al., 2011; Hagenau et al., 2013).

For instance, Généreux et al. (2011) employed different types of features such as unigrams, stems, financial terms, health metaphors and agent-metaphors along with some feature selection methods (IG, CHI, and DF) and two feature weighting methods (binary and TF). The authors achieved the highest accuracy with unigrams, IG feature selection, and TF feature weighting. As mentioned earlier, unigram alone cannot convey any sentiment and feature. On the other hand, IG method that was used as a feature selection method is sensitive to data size and TF feature scheme cannot be a proper feature weighting method since it only determines the importance of each feature in a document and not in the whole corpus.

Finally, majority of researches (Khadjeh Nassirtoussi et al., 2015; Joshi et al., 2016; Moore & Rayson 2017) related to financial news sentiment classification have used SVM as machine learning while none of these researches have used it as an optimized classifier learning method due to the focus on numerical financial data.

According to the above discussion the existing studies reveal some shortages on financial news classification. It can be categorized into these issues in different aspects: feature extraction and selection, feature weighting methods, and utilization of optimized classifiers. Therefore, this study focuses particularly on how to extract and select relevant features especially low frequency sentiment expression features in order to classify news documents as positive and negative by a high-performance sentiment classifier.

The weaknesses are extracted from the related works can be concluded as follows:

- Linguistic and statistical relevant features
- High Dimensionality of feature space
- Non optimized classifiers

1.3 Research Questions

This thesis is going to address the following questions:

- How does the combination of statistical features in a high frequency-based feature selection model (Ngram-model) affect the performance of sentiment classification of financial news?

- How does the integration of statistical and linguistic features in a low frequency-based feature selection model (NgramPOS-model) perform in improving the efficiency of sentiment classification accuracy?
- What is the impact of the unsupervised feature weighting methods on sentiment classification accuracy in high and low frequency-based feature selection model (Ngram-based model and NgramPOS-based model)?
- How does the dimensionality reduction affect financial news sentiment classification in high and low frequency-based feature spaces model (Ngram-based model and NgramPOS-based model)?
- How does the optimized SVM classifier improve classification performance?

1.4 Research Objectives

The main objective of this research is to propose two feature process models to sentiment classification of financial news based on the combination of statistical-based and linguistic-based approaches to improve the sentiment classification accuracy in terms of informative and sentiment feature extraction.

- To propose a high frequency-based feature selection model based on the combination of statistical features to sentiment classification of financial news.
- To propose a low frequency-based feature selection model based on the integration of statistical and linguistic features approaches to news sentiment classification.
- To propose various feature spaces using different unsupervised feature weighting methods in order to recognize the most appropriate method to each feature selection model to achieve the most accurate results in sentiment classification.
- To propose an optimized sentiment classification model for financial news based on supervised machine learning using dimensionality reduction approaches.

1.5 Research Scope

This study satisfies various challenges and research objectives of sentiment classification of financial news. The main focus of this thesis is on feature process models based on statistical, linguistic, unsupervised feature weighting, and dimensionality reduction methods to obtain an optimal feature space to classify financial news by supervised machine learning according to sentiment orientation of each of them which was aggregated from the *Google Finance News*. The collected data is considered as unbalanced to coordinate with the real case. Therefore, since the provided dataset contains more positive financial news documents than negative ones, hence experimental results will have higher sentiment classification accuracy for positive financial news text than negative ones.

1.6 Contributions of the Work

The major contribution of this thesis is to design and implement two feature process models to sentiment classification of financial news which is able to classify news document according to sentiment-rich features with a high rate of accuracy. The goal has been to find a way to process news documents for extracting and selecting richer sentiment features specially low frequency features using statistical and linguistic approaches associated with unsupervised feature weighting methods in order to represent financial textual news. To achieve this goal, the proposed sentiment classification schema integrates these statistical and linguistic methods as two high and low frequency-based models where these models extract and select sentiment words as unigrams and bigrams with based on the defined patterns in Part of Speech (POS), N-gram model, and apply the proper unsupervised feature weighting to news text presentation. These models overcome the limitations of the previous financial news sentiment classification systems. This study makes the following contributions:

- **Demonstrate the usefulness of combining high frequency-based statistical features to improve of news sentiment classification.**

Chapter 4 of this thesis will illustrate the new statistical-based framework for financial news sentiment classification, which is known as Ngram-based model. This model utilizes statistical methods as feature selection and extraction methods in order to classify financial news through the combination of unigrams and bigrams along with TF-IDF feature weighting while applying *Document Frequency* method with a certain threshold.

- **Demonstrate the strength of incorporation of low frequency-based prominent statistical and linguistic features to determine the polarity news as positive and negative**.

Chapter 5 of this thesis will propose an effective model for sentiment classification of financial news which is able to enhance the performance of feature extraction and selection in Ngram-based model. NgramPOS-based model employs a combination of statistical and linguistic methods to extract sentiment information as feature. This model utilizes the combination of sentiment-rich words and phrases as unigrams and bigrams (Uni-POS, Bi-POS) based on the defined fixed patterns along with binary weighting method while applying *Principle Component Analysis* (PCA) as a dimension reduction method in order to classify financial news documents.

- **Define various feature spaces using different unsupervised feature weighting methods in order to recognize the most appropriate method to each feature selection model to achieve the most accurate results in sentiment classification.**

The customized TF-IDF is generated by involving the number of unigrams, bigrams and composition of them for use in both models where the length of each news document is determined by the number of unigrams, bigrams, and their combination when each of them is used as feature to represent text news.

- **Demonstrate the efficiency of the optimized SVM classifiers as a supervised method to classify financial news documents as positive and negative.**

Chapter 4 of this thesis will experimentally demonstrate that optimized SVM classifiers as a supervised method can successfully classify financial news as different feature spaces with high accuracy rate.

1.7 Thesis Organization

This chapter has provided an overview of the whole thesis, which is structured as follow:

Chapter 2: Background and General Concepts. This chapter reviews the literature relevant to the thesis; and introduces the key concepts of research and revisits the available systems in financial news which have included textual sentiment as a major part. First, the basic concepts are defined in sentiment classification and are analyzed and then in second part, the related works are investigated in three subcategories: (i) financial textual sources, (ii) feature preprocessing, and (iii) machine learning methods for classification.

Chapter 3: Research Methodology. This chapter presents the design principles and implementation of each phase of the research. All of the considerations involved in designing the proposed feature process models for financial news sentiment classification including data collection process, the relevant concepts for *Support Vector Machine* (SVM) classifier, and evaluation metrics are described in this chapter. The chapter begins with an overview of the research design and then continues giving details and purposes of the phases and experiments.

Chapter 4: Design and Development of Ngram-based Model using Statistical Approach. This chapter discusses the design, implementation and evaluation steps for first experiment (statistical-based model) in order to attain an Ngram-based feature process model. It begins with illustrating the design and development the model and experiments. Then continues with a brief review of the model and the methods related to the Ngram-based model and is followed by its detailed design. Finally, this chapter presents a comprehensive discussion on the implementation and the evaluation of the Ngram-based model in this chapter.

Chapter 5: Design and Development of NgramPOS-based Model using Linguistic Approach. This chapter discusses the design, implementation and evaluation steps for the second experiment (linguistic-based approach) in order to achieve an NgramPOS-based feature process model. It begins with a brief review of the model and the methods related to the NgramPOS-based model and followed by detailed design steps are explained.

Chapter 6: Conclusion and Future Work. The final chapter of this thesis, discusses the conclusions about the research describe throughout the dissertation, and lists the contributions. It will also present suggestions for future work.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *English*, *2*(4), 433–470. http://doi.org/10.1002/wics.101

Ageev, M. S., & Dobrov, B. V. (2003). Support Vector Machine Parameter Optimization for Text Categorization Problems. *Proceedings of International Conference ISTA*, 165–176. http://doi.org/10.1.1.107.9637

Aggarwal, C. C., & Zhai, C. (2012). A SURVEY OF TEXT CLASSIFICATION ALGORITHMS. In C. Aggarwal & C. Zhai (Eds.), *Mining Text Data (A survery of opinion mining and sentiment analysis)* (pp. 163–222). Springer New York Dorderecht Heidelberg London. http://doi.org/10.1007/978-1-4614-3223-4

Alpaydin, E. (2010). *Introduction to machine learning* (second).

Alvim, L., Vilela, P., Motta, E., & Milidiú, R. L. (2010). Sentiment of Financial News : A Natural Language Processing Approach. *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires*, 1–3.

Aphinyanaphongs, Y., Lulejian, A., Penfold Brown, D., Bonneau, R., & Krebs, P. (2016). Text Classification for Automatic Detection of E-Cigarette Use and Use for Smoking Cessation from Twitter: A Feasibility Pilot. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2016), Volume: 21*, 480–491. http://doi.org/10.14440/jbm.2015.54.A

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, *1*(1), 4–20. http://doi.org/10.4304/jait.1.1.4-20

Banchs, R. E. (2013). *Statistical Models. Text Mining with MATLAB*. Springer Science+Business Media New York. http://doi.org/10.1007/978-1-4614-4151-9

Benamara, F., Cesarano, C., & Reforgiato, D. (2007). Sentiment Analysis : Adjectives and Adverbs are better than Adjectives Alone. *In Proc of Int Conf on Weblogs and Social Media*, 1–4. http://doi.org/citeulike-article-id:9387439

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications*, *42*(6), 3105–3114. http://doi.org/10.1016/j.eswa.2014.11.038

Brücher, H., Knolmayer, G., & Mittermayer, M. (2002). Document Classification Methods for Organizing Explicit Knowledge. *Research Group Information Engi- Neering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland*, 1–26.

Butler, M., & Kešelj, V. (2009). Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports. *Advances in Artificial Intelligence*, *31*(1), 39–51. http://doi.org/10.1007/978-3-642-01818-3_7

Carvalho, J. P., Rosa, H., Brogueira, G., & Batista, F. (2017). MISNIS: An intelligent platform for twitter topic mining. *Expert Systems with Applications*, *89*, 374–388. http://doi.org/10.1016/j.eswa.2017.08.001

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164–175. http://doi.org/10.1016/j.dss.2010.07.012

Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2008). A Practical Guide to Support Vector Classification. *BJU International*, *101*(1), 1396–400. http://doi.org/10.1177/02632760022050997

Colas, F., & Brazdil, P. (2006). Comparison of SVM and some older classification algorithms in text classification tasks. *IFIP International Federation for Information Processing*, *217*, 169–178. http://doi.org/10.1007/978-0-387-34747-9_18

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *297,* 273–297. http://doi.org/10.1023/A:1022627411411

Daniel, M., Ferreira Neves, R., & Horta, N. (2016). Company event popularity for financial markets using Twitter and sentiment analysis. *Expert Systems With Applications*, *71*, 111–124. http://doi.org/10.1016/j.eswa.2016.11.022

Das, S., Martínez-Jerez, A., & Tufano, P. (2005). eInformation: A Clinical Study of Investor Discussion and Sentiment. *Financial Management*, *34*(3), 103–137. http://doi.org/10.1111/j.1755-053X.2005.tb00112.x

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

Dave, K., Way, I., Lawrence, S., & Pennock, D. M. (2003). Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews, 519–528.

Dhillon, I., & Modha, D. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1), 143–175.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Association for Computational Linguistics*, *19:1*, 61–74.

Esuli, A., Esuli, A., Sebastiani, F., & Sebastiani, F. (2005). Determining the Semantic Orientation of Terms through Gloss Classification. *In Proceedings of CIKM*, 617–624. http://doi.org/10.1.1.61.7611

Fama, E. (1965). Random Walks in Stock Market Prices. *Financial Analysts Journal*, *21*(5), 55–59.

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

Feldman, R., & Sanger, J. (2006). *THE TEXT MINING HAND BOOK*. New York: Cambridge University Press. Retrieved from www.cambridge.org

Foreman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Gao, Y., & Sun, S. (2010). An empirical evaluation of linear and nonlinear kernels for text classification using Support Vector Machines. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, (Fskd), 1502–1505. http://doi.org/10.1109/FSKD.2010.5569327

Généreux, M., Poibeau, T., & Koppel, M. (2011a). Sentiment Analysis Using Automatically Labelled Financial News Items. In *Affective Computing and Sentiment Analysis* (Vol. 45, pp. 101–114). the series Text, Speech and Language Technology, Springer. http://doi.org/10.1109/MIS.2016.31

Généreux, M., Poibeau, T., & Koppel, M. (2011b). Sentiment Analysis Using Automatically Labelled Financial News Items Michel. In K. Ahmad (Ed.), *Computing and Sentiment Analysis, Text, Speech and Language Technology* (Vol. 45). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-007-1757-2

Glavaš, G., & Šnajder, J. (2014). Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications*, *41*(15), 6904–6916. http://doi.org/10.1016/j.eswa.2014.04.004

Graovac, J. (2014). Text Categorization Using n-Gram Based Language Independent Technique. *35 Godina Racunarske Lingvistike U Srbiji, Book of Abstracts, to Appear in the Proceedings of the Conference*. http://doi.org/10.3233/IDA-140663

Graovac, J., Kovačević, J., & Pavlović-lažetić, G. (2015). Language Independent n-Gram-Based Text Categorization with Weighting Factors : A Case Study. *JIDM - Journal of Information and Data Management*, *6*(1), 4–17.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*(4), 680–691. http://doi.org/10.1016/j.dss.2010.08.019

Gunal, S., Ergin, S., Gulmezoglu, M. B., & Gerek, O. N. (2006). On feature extraction for spam e-mail detection. *Multimedia Content Representation, Classification and Security*, *4105*, 635–642.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. http://doi.org/10.1016/j.aca.2011.07.027

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features.

Decision Support Systems, *55*, 685–697. http://doi.org/10.1016/j.dss.2013.02.006

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *0*, 1–14. http://doi.org/10.1016/j.knosys.2017.05.001

Han, E. S., Karypis, G., & Kumar, V. (2001). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. *5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 1–7. http://doi.org/10.1109/34.310689

Han, J., & Kamber, M. (2006). *Data Mining (Concepts and Techniques)*. (J. Widom & S. Ceri, Eds.). Elsevier (Morgan Kaufmann).

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining*. http://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Hatzivassiloglou, V., McKeown, K. R., Pang, B., Lee, L., Vaithyanathan, S., Ku, L.-W., … Chen, H.-H. (2009). Predicting the Semantic Orientation of Adjectives. *ACM Transactions on Information Systems (TOIS)*, *21*(4), 315–346. http://doi.org/10.3115/979617.979640

Hsu, C., Chang, C., & Lin, C. (2010). A Practical Guide to Support Vector Classification. *Bioinformatics*, *1*(1), 1–16. http://doi.org/10.1177/02632760022050997

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 04*, *4*, 168. http://doi.org/10.1145/1014052.1014073

Hull, D. (1996). Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, *47*(1), 70–84. http://doi.org/10.1002/(SICI)1097-4571(199601)47:1<70::AID-ASI7>3.3.CO;2-Q

Islam, S. M. A., Heil, B. J., Kearney, C. M., & Baker, E. J. (2017). Protein classification using modified n-gram and skip-gram models, *0*(0), 1–7.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning ECML '98*, 137–142. http://doi.org/10.1007/BFb0026683

Josan, G. S., & Lehal, G. S. (2008). Size of N for Word Sense Disambiguation using N gram model for Punjabi Language. *International Journal of Translation*, *20*.

Joshi, K., H. N., B., & Jyothi Rao. (2016). STock Trend Prediction Using News Sentiment Analysis. *CoRR*, *abs/1607.0*. Retrieved from http://arxiv.org/abs/1607.01958

Joshi, M. V. (2002). On evaluating performance of classifiers for rare classes. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 641–644. http://doi.org/10.1109/ICDM.2002.1184018

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, *15*(9), 28. Retrieved from http://www.jstatsoft.org/v15/i09/paper

Kearney, C., & Liu, S. (2014). International Review of Financial Analysis Textual sentiment in fi nance : A survey of methods and models ☆. *International Review of Financial Analysis*, *33*(Cc), 171–185. http://doi.org/10.1016/j.irfa.2014.02.006

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670. http://doi.org/10.1016/j.eswa.2014.06.009

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. http://doi.org/10.1016/j.eswa.2014.08.004

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, *1*(1), 4–20. http://doi.org/10.4304/jait.1.1.4-20

Kim, D., & Kim, S. (2012). Investor Sentiment from Internet Message Postings and Predictability of Stock Returns Investor Sentiment from Internet Message Postings and Predictability of Stock Returns, *107*(February), 708–729.

Kit, C., Xu, Z., & Webster, J. J. (2003). Integrating Ngram Model and Case-based Learning For Chinese Word Segmentation, (July), 160–163.

Knowles, F. (1996). Lexicographical Aspects of Health Metaphors in Financial Text.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(12), 1137–1143. http://doi.org/10.1067/mod.2000.109031

Koppel, M., & Shtrimberg, I. (2006). Good News or Bad News ? Let the Market Decide. *Computing Attitude and Affect in Text: Theory and Applications the Information Retrieval*, *20*, 297–301. http://doi.org/10.1007/1-4020-4102-0_22

Korde, V., & Mahender, C. N. (2012). Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. http://doi.org/10.5121/ijaia.2012.3208

Kumar, B. S., & Ravi, V. (2016). Knowle dge-Base d Systems A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128–147. http://doi.org/10.1016/j.knosys.2016.10.003

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded Methods. *Studies in Fuzziness and Soft Computing*, *207*, 137–165. http://doi.org/10.1007/978-3-540-35488-8

Lan, M. L. M., Tan, C. L. T. C. L., Su, J. S. J., & Lu, Y. L. Y. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 721–735. http://doi.org/10.1109/TPAMI.2008.110

Lantz, B. (2013). *Mahine Learning with R*. (P. Arondekar, P. More, A. Ramchandran, & H. Vairat, Eds.). Birmingham B3 2PB, UK.: Packt Publishing Ltd. Retrieved from www.packtpub.com

Li, F. (2010). The information content of forward- looking statements in corporate filings-A na??ve bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102. http://doi.org/10.1111/j.1475-679X.2010.00382.x

Lin, H., & Lin, C. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *Submitted to Neural Computation*, 1–32. http://doi.org/10.1.1.14.6709

Lin, Y. (1999). Support Vector Machines and the Bayes Rule in classification. *Technical Report No.1014, Department of Statistics, University of Wiscousin, Madison.*

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies,; # 16; Variation: Synthesis Lectures on Human Language Technologies ;; # 16.*, (May), 1–108. http://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, L. L. L., Kang, J. K. J., Yu, J. Y. J., & Wang, Z. W. Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. *2005 International Conference on Natural Language Processing and Knowledge Engineering*, *9*, 597–601. http://doi.org/10.1109/NLPKE.2005.1598807

Loughran, T., & Mcdonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Distionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. http://doi.org/10.1111/j.1540-6261.2010.01625.x

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, *179*(13), 2208–2217. http://doi.org/10.1016/j.ins.2009.02.014

Maragoudakis, M., & Serpanos, D. (2016). Exploiting Financial News and Social Media Opinions for Stock Market Analysis using MCMC Bayesian Inference. *Computational Economics*, *47*(4), 589–622. http://doi.org/10.1007/s10614-015-9492-9

Mayne, A. (2010). *Sentiment Analysis for Financial News*. University of Sydney.

Mejova, Y., & Srinivasan, P. (2011). Exploring Feature Definition and Selection for Sentiment Classifiers. In *Fifth International AAAI Conference on Weblogs and Social Media* (pp. 546–549).

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 1st ed. Elsevier*. http://doi.org/10.1016/B978-0-12-386979-1.00009-8

Miranda, J., Montoya, R., & Weber, R. (2005). Linear Penalization Support Vector Machines for Feature Selection. In *Pattern Recognition and Machine Intelligence* (Vol. 3776, pp. 188–192).

Mittermayer, M.-A. (2004). Forecasting Intraday stock price trends with text mining techniques. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, *0*(C), 1–10. http://doi.org/10.1109/HICSS.2004.1265201

Mohammad, H., & M.N., S. (2015). A R EVIEW ON E VALUATION M ETRICS F OR D ATA C LASSIFICATION E VALUATIONS. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(2), 1–11. http://doi.org/DOI: 10.5121/ijdkp.2015.5201

Moore, A., & Rayson, P. (2017). Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 581–585. Retrieved from http://www.aclweb.org/anthology/S17-2095

Moreo, a., Romero, M., Castro, J. L., & Zurita, J. M. (2012). Lexicon-based Comments-oriented News Sentiment Analyzer system. *Expert Systems with Applications*, *39*(10), 9166–9180. http://doi.org/10.1016/j.eswa.2012.02.057

Mullen, T., Mullen, T., Collier, N., & Collier, N. (2004). Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 412–418). Retrieved from research.nii.ac.jp/~collier/group/mullen/Papers/emnlp_corrected.pdf

Myllymgki, P., & Tirri, H. (1993). Bayesian Case-Based Reasoning with Neural Networks Bayesian Case-Based Reasoning with Neural Networks. In *IEEE International Conference on Neural Network'93* (Vol. 1, pp. 422–427). http://doi.org/10.1109/ICNN.1993.298594

Na, J., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews. In *Advances in Knowledge Organization* (Vol. 9, pp. 49–54).

Nasukawa, T., & Yi, J. (2003). Sentiment Analysis : Capturing Favorability Using Natural Language Processing. *2nd International Conference on Knowledge Capture*, 70–77. http://doi.org/10.1145/945645.945658

Ng, V., Dasgupta, S., & Arifin, S. M. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *Proc. COLING/ACL Main Conf. Poster Sess.*, *pp*(July), 611–618. Retrieved from http://portal.acm.org/citation.cfm?id=1273152

Ooi, H. S., Schneider, G., Lim, T., Chan, Y., Eisenhaber, B., & Eisenhaber, F. (2010). Data Mining Techniques for the Life Sciences, *609*, 129–144. http://doi.org/10.1007/978-1-60327-241-4

Ozik, G., & Sadka, R. (2012). Media and Investment Management. *SSRN Electronic Journal*, *33*(0). http://doi.org/10.2139/ssrn.1633705

Paltoglou, G., & Thelwall, M. (2010). A study of Information Retrieval weighting schemes for sentiment analysis. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, (July), 1386–1395. Retrieved from http://www.aclweb.org/anthology/P10-1141

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis using Subjectivity Summation based on Minimum Cuts. *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271. http://doi.org/10.3115/1218955.1218990

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. *Foundations and Trends® in Information Retrieval* (Vol. 2). http://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, (July), 79–86. http://doi.org/10.3115/1118693.1118704

Pederson, T. (2001). A Desision Tree of Bigrams is an Accurate Predictor of Word Sence. In *proceeding of the second NAACL* (pp. 79–86).

Pedregosa, F., Grisel, O., Weiss, R., Passos, A., & Brucher, M. (2011). *Scikit-learn : Machine Learning in Python*. *Journal of Machine Learning Research* (Vol. 12). http://doi.org/10.1007/s13398-014-0173-7.2

Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, *24*, 131–139. http://doi.org/10.1145/563932.563921

Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers and Mathematics with Applications*, *65*(10), 1427–1437. http://doi.org/10.1016/j.camwa.2012.09.011

Pham Xuan, N., & Le Quang, H. (2014). A New Improved Term Weighting Scheme for Text Categorization. *Advances in Intelligent Systems and Computing*, *271*, 261–270. http://doi.org/10.1007/978-3-319-11680-8

Porter, M. F. (2006). An algorithm for suffix stripping. *Emerald Inisght*, *14*(3), 130–137. http://doi.org/10.1108/eb046814

Pui Cheong Fung, G., Xu Yu, J., & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. *IEEE/IAFE Conference on Computational Intelligence for Financial Engineering, Proceedings (CIFEr), 2003–Janua*, 395–402. http://doi.org/10.1109/CIFER.2003.1196287

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADM RAL: A data mining based financial trading system. In *IEEE Symposium on Computational Intelligence and Data Mining*, (pp. 720–725). http://doi.org/10.1109/CIDM.2007.368947

Radovanović, M., & Ivanovi, M. (2008). APPROACHES AND APPLICATIONS. *October*, *38*(3), 227–234.

Ringsquandl, M., & Petković, D. (2013). Analyzing Political Sentiment on Twitter. *AAAI Spring Symposium Series Analyzing Microtext*, 40–47. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/viewFile/5702/5909

Salton, G., & Buckley, C. (1997). Term weighting approaches in automatic text retrieval. *Readings in Information Retrieval*.

Salton, G., Wong, A., & Yang, C. S. (1975). Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620. http://doi.org/10.1145/361219.361220

Schölkopf, B. (2002). Learning with kernels. *Journal of the Electrochemical Society*, *129*(November), 2865. http://doi.org/10.1198/jasa.2003.s269

Schölkopf, B., & Smola, A. (2005). Support Vector Machines and Kernel Algorithms. (pp. 1–22). Retrieved from http://eprints.pascal-network.org/archive/00001021/

Schölkopf, B., & Smola, A. J. (2002). Support Vector Machines and Kernel Algorithms. *The Handbook of Brain Theory and Neural Networks*, 1119–1125.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464. http://doi.org/10.1016/j.dss.2012.03.001

Schumaker, R. P., Zhang, Y., Huang, C., & Rochelle, N. (2009). Sentiment Analysis of Financial News Articles, 1–21.

Sebastiani, F. (2001). Machine Learning in Automated Text Categorization. http://doi.org/10.1145/505282.505283

Sheen, S., Anitha, R., & Sirisha, P. (2013). Malware detection by pruning of parallel ensembles using harmony search. *Pattern Recognition Letters*, *34*(14), 1679–1686. http://doi.org/10.1016/j.patrec.2013.05.006

Singh, U., Goyal, V., & Rani, A. (2014). Disambiguating hindi words using n-gram smoothing models. *An International Journal of Engineering Sciences*, *10*(June 2014), 26–29.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. http://doi.org/10.1016/j.ipm.2009.03.002

Srivastava, A., & Sahami, M. (2009). *Text Mining, classification, clustering, and applications*. *Computer (Long. Beach. Calif).* http://doi.org/10.1111/j.1751-5823.2010.00109_1.x

Stone, F. J., & Hunt, E. B. (1963). A Computer Approach to Content Analysis, Studies Using the General Inquirer System. *Managing Requirements Knowledge, International Workshop on*, *0*, 241. http://doi.org/http://doi.ieeecomputersociety.org/10.1109/AFIPS.1963.1

Tam, V., Santoso, A., & Setiono, R. (2002). A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization. *Object Recognition Supported by User Interaction for Service Robots*, *4*, 235–238. http://doi.org/10.1109/ICPR.2002.1047440

Taylor, A., Marcus, M., & Santorini, B. (2003). the Penn Treebank: an Overview. *Treebanks*, 5–22. http://doi.org/10.1007/978-94-010-0201-1_1

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*(3), 1139–1168. http://doi.org/10.1111/j.1540-6261.2007.01232.x

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems With Applications*, *57*, 117–126. http://doi.org/DOI: 10.1016/j.eswa.2016.03.028

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (July), 417–424. http://doi.org/10.3115/1073083.1073153

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classificatio. *Information Processing and Management*, *50*, 104–112.

Vapnik, V. P. (1995). *The nature of statistical learning Theory*. Springer, NewYork.

Vu, T., Chang, S., Ha, Q., & Collier, N. (2012). An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter. In *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, Mumbai.* (pp. 23–38).

Wiebe, J., Bruce, R., & O'Hara, T. (1998). D e v e l o p m e n t and U s e of a G o l d - S t a n d a r d D a t a Set for S u b j e c t i v i t y Classifications. in Proceedings of the Association for Computational Linguistics (ACL-1999).

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just How Mad Are You ? Finding Strong and Weak Opinion Clauses. *Proceedings of AAAI*, 761–767.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *School of Computer Science, Carne- Gie Mellon University*.

Yang, Y., & Pederson, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14 th International Conference on Machine Learning* (pp. 412–420).

Yu, B., Xu, Z., & Li, C. (2008). Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems*, *21*(8), 900–904. http://doi.org/10.1016/j.knosys.2008.03.045

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on fi rm equity value : A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919–926. http://doi.org/10.1016/j.dss.2012.12.028

Zhai, J. J., Cohen, N., & Atreya, A. (2011). CS224N Final Project : Sentiment analysis of news articles for financial signal prediction, 1–8.

Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. *Advances in Neural Networks*, *4493*, 1087–1096.

Zhang, W., & Skiena, S. (2010). Trading Strategies to Exploit Blog and News Sentiment. *Icwsm*, *34*(Chan 2003), 375–378. http://doi.org/10.1016/j.jbankfin.2009.11.025

Zhen, Z., Wang, H., Han, L., & Shi, Z. (2011). Categorical Document Frequency Based Feature Selection for Text Categorization. *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, *2*, 65–68. http://doi.org/10.1109/ICM.2011.365