malaysia business news mining and impact on stock price

YONG KEH SOON

faculty of computer science and information technology

UNIVERSITY OF MALAYA

Kuala lumpur

june 2020

**malaysia BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE**

yong keh soon

RESEARCH PROJECT submitted in PARTIAL FULFILMENT OF the requirEmentS FOR the degree of master of data science

faculty of computer science and information technology

UNIVERSITY OF MALAYA

Kuala lumpur

**june 2020**

**UNIVERSITY OF MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: YONG KEH SOON (I.C/Passport No: 730726-08-5387)

Matric No: WQD 180065

Name of Degree: MASTER OF DATA SCIENCE

Title of Project Paper/Research Report/Dissertation/Thesis (“this Work”):

MALAYSIA BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE

Field of Study:

I do solemnly and sincerely declare that:

1. I am the sole author/writer of this Work;
2. This Work is original.
3. Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
4. I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
5. I hereby assign all and every rights in the copyright to this Work to the University of Malaya (“UM”), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
6. I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate’s Signature Date:

Subscribed and solemnly declared before,

Witness’s Signature Date:

Name:

Designation:

MALAYSIA BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE  
Abstract

There has been exponential growth in research on computer based natural language processing (NLP). In finance, text-based data such as news and corporate announcement had emerged into alternative sources and gained popularity among investors. It is an actively researched domain mainly within the western economy context. Research on news impact to Asia Pacific market is limited, especially in Malaysia. This research covered two main sectors within Kuala Lumpur Stock Exchange, financial services and telecommunication & media, with ten years of data. News were scrapped from two local media websites. Text preprocessing includes semantic extraction using pretrained SentiWordNet dictionary to determine positive and negative terms. Sentiment score is engineered from the term count as a quantitative measurement for each news article. The news sentiment is categorized into UP, STAY and DOWN, segregated by upper and lower boundary of tunable thresholds. Multiple models were constructed by mapping these categories into near term stock returns. Performance of the model was baselined with actual distribution of stock returns label, which were found to be quite evenly spread across the three categories. Both Results from both models and baselines were averaged for comparison. The final accuracy performed lower than baseline marginally. However, F1 means for all categories outperformed the baseline by over 15%. With this predictability, this research concluded that news has impact on stock price movement in the short term and the result is not due to randomness.

Keywords: News Classification, Sentiment Analysis, Natural Language Processing, Confusion Matrix, Kuala Lumpur Stock Exchange

MALAYSIA BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE  
Abstrak

Pertumbuhan eksponensial telah ada dalam bidang pemprosesan bahasa tabii oleh komputer. Dalam kewangan, data berdasarkan teks seperti berita dan pengumuman korporat telah muncul ke dalam sumber alternatif dan menjadi popular dalam kalangan pelabur. Ia adalah domain penyelidikan aktif, terutamanya dalam konteks ekonomi Barat. Penyelidikan tentang impak berita ke atas pasaran Asia Pasifik adalah terhad, terutamanya di Malaysia. Penyelidikan ini meliputi dua sektor utama dalam Bursa Saham Kuala Lumpur iaitu perkhidmatan kewangan dan telekomunikasi & media, dengan data 10 tahun. Berita telah diambil keluar daripada dua media laman web tempatan. Pemprosesan Text termasuk pengekstrakan semantik menggunakan kamus terlatih SentiWordNet untuk menentukan istilah-istilah positif dan negatif. Skor sentimen direkayasa daripada istilah yang dikira sebagai pengukuran kuantitatif daripada setiap artikel berita. Berita sentimen dikategorikan kepada UP, STAY dan DOWN, diasingkan oleh batasan atas dan bawah yang boleh ditala. Pelbagai model dibentuk dengan memetakan kategori-kategori ini kepada pulangan stok jangka masa dekat. Prestasi model tersebut adalah berdasarkan pengedaran sebenar label pemulangan stok, yang mana didapati disebarkan secara rata menyeberangi tiga kategori tersebut. Keputusan daripada model dan garis dasar telah dipuratakan untuk perbandingan. Prestasi ketepatan akhir adalah rendah sedikit daripada garis dasar. Walau bagaimanapun, F1-means untuk semua kategori yang melepasi tahap garis dasar sebanyak 15%. Dengan ramalan ini, kajian ini merumuskan bahawa berita membawa impak ke atas pergerakan harga bursa pada jangka pendek dan keputusan itu tidak berdasar kepada keberangkalian.

Kata Kunci : Klasifikasi Berita, Analisis Sentimen, Pemprosesan Bahasa Tabii, Confusion Matrix, Bursa Saham Kuala Lumpur

Acknowledgements

This work is dedicated to my family. My wife Swee Poh, who had never complained about countless weekends spent on projects and assignments in the past two years. And to my two girls Jolin and Janice, I know at times you wonder why your father invested so much time and effort in my pursue of this Master Degree. You gotten my response repeatedly, which is to find your interest, invest and enrich yourself with knowledge and stay curious. I hope you have a similar attitude towards life with me as the role model, and all you need is to find your passion.

And to mention my course mates, Shien Long, Vikas, Vee King, Sri and Leonard. Your never die attitude had made our assignment struggle seems all worth it and fun. I’m glad to have you in this journey.

Special thanks the lecturers in this program. Just to name a few, Dr. Teh Ying Wah who repeatedly set high expectation to his students in almost every lecture. Dr. Vimala our program coordinator who always keep us on track. Dr. Salimah who made the class fun and interesting, and Dr. Hoo Wai Lam who make many of us dig through tedious hadoop setup. I never thought my Unix skill can be advanced so much.

I must also thank my supervisor Dr. Ow Siew Hock for all her guidance and encouragement in this work. You set the standard of clear objective, simplified statements and stay on focus.

Lastly to my late father. I remember attending this master program orientation in the afternoon, on the day of your funeral. I always had your trust and support throughout my life that I’m deeply grateful for. Although you don’t have much words, this work is only possible because of your blessings.

Table of Contents

[MALAYSIA BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE Abstract iii](#_Toc42950586)

[MALAYSIA BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE Abstrak iv](#_Toc42950587)

[Acknowledgements v](#_Toc42950588)

[Table of Contents vi](#_Toc42950589)

[List of Figures ix](#_Toc42950590)

[List of Symbols and Abbreviations xi](#_Toc42950591)

[List of Appendices xii](#_Toc42950592)

[CHAPTER 1: introduction 1](#_Toc42950593)

[1.1 Background 1](#_Toc42950594)

[1.2 Problem Statement 2](#_Toc42950595)

[1.3 Objectives 2](#_Toc42950596)

[1.4 Research Questions 2](#_Toc42950597)

[1.5 Scope of Research 3](#_Toc42950598)

[CHAPTER 2: literature review 5](#_Toc42950599)

[2.1 Background 5](#_Toc42950600)

[2.2 Introduction 6](#_Toc42950601)

[2.3 Content Analysis 6](#_Toc42950602)

[2.3.1 News Semantic 9](#_Toc42950603)

[2.3.2 Feature Processing 11](#_Toc42950604)

[2.3.3 Dimension Reduction 12](#_Toc42950605)

[2.3.4 Feature Representation 13](#_Toc42950606)

[2.4 Prediction Model 14](#_Toc42950607)

[2.4.1 Dictionary Based Approach 14](#_Toc42950608)

[2.4.2 Machine Learning Based Approach 16](#_Toc42950609)

[2.5 Integrating News Categorization 17](#_Toc42950610)

[2.6 News Sources 18](#_Toc42950611)

[2.7 Conclusion 21](#_Toc42950612)

[CHAPTER 3: Research Methodology 22](#_Toc42950613)

[3.1 Terminology Used 22](#_Toc42950614)

[3.1.1 Financial Terminology 22](#_Toc42950615)

[3.1.2 Data Science Terminology 22](#_Toc42950616)

[3.2 Data Sources and Acquisition 23](#_Toc42950617)

[3.2.1 News Acquisition (Text) 23](#_Toc42950618)

[3.2.2 Stock Price Acquisition (Numerical) 25](#_Toc42950619)

[3.3 System Design 27](#_Toc42950620)

[3.4 Database Schema 28](#_Toc42950621)

[3.5 Data Preprocessing 29](#_Toc42950622)

[3.5.1 Handling Missing Data and Errors 29](#_Toc42950623)

[3.5.1.1 Missing Numerical Data and Errors 29](#_Toc42950624)

[3.5.1.2 Missing News Data and Errors 30](#_Toc42950625)

[3.5.2 Feature Engineering 30](#_Toc42950626)

[3.5.2.1 Stock Return 30](#_Toc42950627)

[3.5.2.2 Sector Index Return 32](#_Toc42950628)

[3.5.2.3 News Sentiment 34](#_Toc42950629)

[3.5.3 News To Stock Entity Mapping 42](#_Toc42950630)

[3.5.4 Data Aggregation 43](#_Toc42950631)

[3.5.5 Returns Labelling 44](#_Toc42950632)

[CHAPTER 4: Data Interpetation 46](#_Toc42950633)

[4.1 News 46](#_Toc42950634)

[4.1.1 Overall News 46](#_Toc42950635)

[4.1.2 Sector News 47](#_Toc42950636)

[4.2 Share Price 48](#_Toc42950637)

[4.2.1 Share Returns 48](#_Toc42950638)

[4.2.2 Returns Label 49](#_Toc42950639)

[CHAPTER 5: Data Modeling 51](#_Toc42950640)

[5.1 The Model 51](#_Toc42950641)

[5.2 Performance Metrics 52](#_Toc42950642)

[5.3 Performance Baseline 53](#_Toc42950643)

[CHAPTER 6: RESULTS 56](#_Toc42950644)

[6.1 Results By Group of Stocks and Index 56](#_Toc42950645)

[6.2 Individual Stocks Results 57](#_Toc42950646)

[6.3 Comparison To Other Research 58](#_Toc42950647)

[CHAPTER 7: CONCLUSION 59](#_Toc42950648)

[CHAPTER 8: Future Work 60](#_Toc42950649)

[References 61](#_Toc42950650)

[Appendix A: CODES AND DATA (GITHUB) 64](#_Toc42950651)

List of Figures

[Figure 2.1: Evolution of NLP Research 5](#_Toc42382098)

[Figure 2.2: Sample Business News, (Ding, Zhang, Liu, & Duan, 2014) 9](#_Toc42382099)

[Figure 2.3: From News To Decision 9](#_Toc42382100)

[Figure 2.4: News Feature Processing 12](#_Toc42382101)

[Figure 2.5: Dictionary Based Approach 15](#_Toc42382102)

[Figure 2.6: Machine Learning Based Approach 16](#_Toc42382103)

[Figure 2.7: Proposed Predictive System, adopted from (Shynkevich et al., 2016) 18](#_Toc42382104)

[Figure 3.1: News Provider Search Page 25](#_Toc42382105)

[Figure 3.2: News Article Web Page 25](#_Toc42382106)

[Figure 3.3: Eikon API Call For Pricing Data In Python 26](#_Toc42382107)

[Figure 3.4: CIMB Price Chart 27](#_Toc42382108)

[Figure 3.5: Overall System Design 28](#_Toc42382109)

[Figure 3.6: Database Schema 28](#_Toc42382110)

[Figure 3.7: MBBM.KL Stock Price Return Illustration 31](#_Toc42382111)

[Figure 3.8: Graphical Representation of Senti-WordNet Opinion Properties 36](#_Toc42382112)

[Figure 3.9: Sentiment Scoring For News Article and Term Labeling 38](#_Toc42382113)

[Figure 3.10: Sentiment Score Processing Flowchart 40](#_Toc42382114)

[Figure 3.11: Example Sentiment Scores DataFrame 42](#_Toc42382115)

[Figure 3.12: Example News To Stock Entity Mapping DataFrame 43](#_Toc42382116)

[Figure 3.13: Example News Sentiment To Entity Mapping 43](#_Toc42382117)

[Figure 3.14: Example News Sentiment Scores 44](#_Toc42382118)

[Figure 3.15: Example Share Returns Label 45](#_Toc42382119)

[Figure 4.1: News Count By Sources 46](#_Toc42382120)

[Figure 4.2: News and Market Capital Scatter Plot 47](#_Toc42382121)

[Figure 4.3: Sector News Count 47](#_Toc42382122)

[Figure 4.4: Sector Price Returns Overall Chart 49](#_Toc42382123)

[Figure 4.5: Shares Price Returns Chart By Stock 49](#_Toc42382124)

[Figure 4.6: Returns Label Distribution For Stocks 50](#_Toc42382125)

[Figure 4.7: Returns Label Distribution For Sector Index 50](#_Toc42382126)

[Figure 5.1: Prediction Model 51](#_Toc42382127)

[Figure 5.2: Three Categories Confusion Matrix 53](#_Toc42382128)

[Figure 5.3: Baselining Method 54](#_Toc42382129)

List of Tables

[Table 1.1: Top 10 Largest Capital Stocks (Financial Services) 3](#_Toc42382130)

[Table 1.2: Top 10 Largest Capital Stocks (Telecommunication and Media) 4](#_Toc42382131)

[Table 2.1: Paper Summary Of Content Analysis Method 7](#_Toc42382132)

[Table 2.2: News Event Classification In 8-K Reports (Schumaker & Chen, 2009) 19](#_Toc42382133)

[Table 2.3: List Of Financial Event Types In 8-K Reports (Lee et al., 2014) 19](#_Toc42382134)

[Table 2.4: Results Of Proposed Model (Lee et al., 2014)) 20](#_Toc42382135)

[Table 2.5: Result of Temporal Aspect Model (Lee et al., 2014) 20](#_Toc42382136)

[Table 3.1: Financial Terminology 22](#_Toc42382137)

[Table 3.2: Data Science Terminology 22](#_Toc42382138)

[Table 3.3: News Search Keywords 24](#_Toc42382139)

[Table 3.4: Stock Price Data Count 29](#_Toc42382140)

[Table 3.5: MBBM.KL Stock Price Return Example 31](#_Toc42382141)

[Table 3.6: Sector Index Price Calculation 32](#_Toc42382142)

[Table 3.7: Stock Market Capital Weighing 33](#_Toc42382143)

[Table 3.8: Market Capitalization Weighted Financial Services Index 33](#_Toc42382144)

[Table 3.9: Structure of WordNet 34](#_Toc42382145)

[Table 3.10: Various Meaning of The Same Word in WordNet 36](#_Toc42382146)

[Table 3.11: Senti-WordNet SynSet Scores Example 37](#_Toc42382147)

[Table 3.12: News Sentiment Score Example 1 39](#_Toc42382148)

[Table 3.13: News Sentiment Score Example 2 39](#_Toc42382149)

[Table 3.14: Penn Treebank POS Tag Mapping to WordNet 41](#_Toc42382150)

[Table 5.1: Performance Baseline 55](#_Toc42382151)

[Table 6.1: Model Performance Per Group 57](#_Toc42382152)

[Table 6.2: Model Performance Per Stock 57](#_Toc42382153)

List of Symbols and Abbreviations

|  |  |  |
| --- | --- | --- |
| API | : | Application Programming Interface |
| FS | : | Financial Services |
| IPO | : | Initial Public Offering |
| KLCI | : | Kuala Lumpur Composite Index |
| KLFI | : | Kuala Lumpur Finance Index |
| KLSE | : | Kuala Lumpur Stock Exchange |
| POS | : | Part Of Speech |
| RIC | : | Reuters Instrument Code |
| SWN | : | Senti-WordNet |
| TELCO | : | Telecommunication and Media |
| WN | : | WordNet |
| SWN | : | Senti WordNet |
| SVM | : | Support Vector Machine |
| ASTR.KL | : | Astro Malaysia Holdings Berhad |
| CIMB.KL | : | CIMB Group Holdings Berhad |
| MBBM.KL | : | Maybank Berhad |
| MXSC.KL | : | Maxis Berhad |
| NLTK | : | Natural Language Toolkit |
| OCKG.KL | : | OCK Group Berhad |

List of Appendices

|  |  |
| --- | --- |
| Appendix A: Codes and Data (Github) | 64 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# introduction

## Background

News are widely known to have its influence on stock price movements. In today’s Internet age, news sources and volume are growing rapidly. Emerging technologies had enabled automated news acquisition, events extraction, transformation and classification. For example, information providers such as Bloomberg and Thomson Reuters have emerged into providing text analytics on top of their traditional news services.

Innovation and research grow exponentially in Natural Language Processing (NLP) to develop models that can understand human generated text. Cloud providers like Azure, Amazon, IBM and Google has long pioneered in machine learning and NLP toolsets for researchers and developers. Today, such advance tools are available publicly through open source, such as Python and NLTK library used in this research.

Although studies on news relation to stock prices had branched into many advance topics in recent years, the coverage are mainly on US market. This research focused on Kuala Lumpur Stock Exchange, which is a more active exchanges in South East Asia with about 1000 companies listed. Total KLSE market capitalization is USD 455b, ranked at the fourth place in South East Asia after Singapore, Thailand and Indonesia. This project covered two active sectors namely financial services and telecommunication & media. Six out of twenty stocks studied were top 10 largest company in KLSE.

This research demonstrated data mining process on long-term news data. It is hoped that sentiment analysis methodologies used here can uncover potential relation between news and share price direction.

## Problem Statement

Current research on news sentiment and its relation to stock price has the below limitation:

1. Lack of quantitative research on Kuala Lumpur Stock Exchange (KLSE)
2. Two research on KLSE has limited research period:

* 6 months period, only (Im, San, On, Alfred, & Anthony, 2014)
* Dividend period only (Hor, Goh, & Ow, 2018)

1. Impact of different news category (sector news and specific stock news) not explored

## Objectives

There are two key objectives for this research, summarized below:

1. To find out how well news are able to predict short-term stock price direction
2. To investigate which news categories (sector news, stock specific news) shows greater impact on stock price direction

## Research Questions

This research seeks for the answers to the below questions:

1. Is there any significant relation between news sentiment and short-term stock price direction?
2. Between sector news and stock specific news, is there any difference in their impact on stock price and index direction?

## Scope of Research

Targeted stocks in this research are the ten (10) largest market capital counters listed in Kuala Lumpur Stock Exchange (KLSE), for each sector below:

* Bursa Financial Services (here onwards referred as FSI)
* Bursa Telecommunication & Media (here onwards refereed as TELCO)

The table below shows the selected stocks for each sector. The market capital is based on closing price at 31st December 2019. Note that Astro Malaysia (ASTR.KL) which is the 5th largest stock in TELCO were relisted in KLSE in year 2012. OCK Group (OCKG.KL) which is the 8th largest stock in TELCO was listed in year 2011. As the accumulative of bottom four (4) stocks contributes less than 10% for its sector market capital, their price movement is not expected to have significant influence on overall sector index.

Table .: Top 10 Largest Capital Stocks (Financial Services)



Table .: Top 10 Largest Capital Stocks (Telecommunication and Media)



As for news provider, two local Malaysia news sources had been selected for this research:

* **The Star** – A dominant English-language newspaper in Malaysia, publishing since 1971. This research focused on “Business” section of this publisher.
* **The Edge** – A leading business and financial publisher in Malaysia and Singapore since 1994. The Edge Malaysia is known as the country’s best financial publication in providing insights and in-dept reports on Malaysia business with loyal followers from corporate leaders.

This research is conducted on ten (10) years data from 1st January 2010 till 31st December 2019.

# literature review

## Background

Many research studies have proven that news have impact on direction of stock prices. Financial services particularly investment banking and funds are complex and competitive industry. Institutions making use of alternative data sources such as news, blogs and tweets are becoming increasingly important as an edge to construct and maintain profitable equity portfolio. It is becoming more common for such information edge to be included as part of equity strategies (Xing, Cambria, & Welsch, 2018). Fundamental analysis backed by NLP, is now a critical indicator to understand how financial analysts and masses interpret the market.

News data is one of the fastest growing alternative data available to public in general. Finance domain knowledge and principles of data science are both key to uncover investment insights. In (Fisher, Garnsey, & Hughes, 2016), the authors refer to several NLP based research efforts that have gained exponential growth in the academic world in the last three decades. The authors have categorized a total of 266 monographs into Accounting, Auditing and Finance. Among the three sub domains, Finance is found to be the most heavily researched.

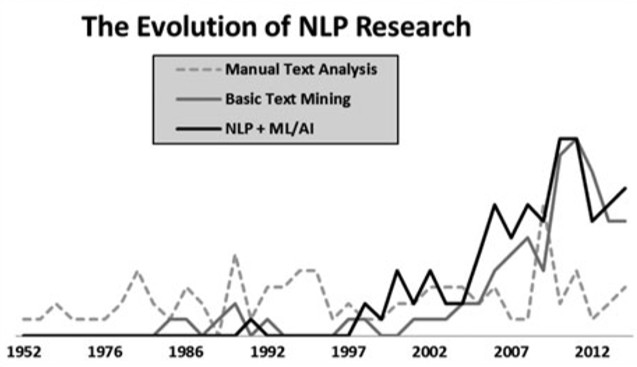


Figure .: Evolution of NLP Research

## Introduction

This review starts by providing a general overview of data pipelines and tools for news data extraction. Different commonly used finance related text data sources and various feature processing techniques, including extraction, reduction and representation are explored. Techniques that could aid in performing the sentiment analysis and share various findings from different techniques is included in this review. The weaknesses and strengths of various techniques are also discussed.

In more advanced implementations by commercial products, this review provides an understanding of fine-grained approaches in extracting the degree of news relevance of the target stock to improve financial news-based forecasting. We reviewed some common event extraction techniques and text data transformation for quantitative modeling. In addition, we identify common research issues and methodologies for the evaluation of such an event extraction system. Lastly, we reviewed a few papers where the researchers had extended their news-to-price movement findings and tested them against trading strategies. From these sources, we generalized a framework of integrating technical or fundamental indicators into news prediction.

Business news sentiment has significant correlation to stock price movement. Such predictive model can predict abnormal gain and used as a profitable trading strategy.

## Content Analysis

We shall discuss some NLP techniques commonly used in analyzing text and understand it's potential underlying meaning that could move the financial market. A summary of the papers that have been reviewed for this section can be found in below.

Table .: Paper Summary Of Content Analysis Method

|  | Author | Data Source | News Filtering | Features | Features Selection and Reduction | Feature Representation | Prediction | Machine Learning | Findings, Strengths/Weaknesses |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Tetlock (2007) | * WSJ Abreast of the Market column * 1984-1999 | None | * Unigram * Volumes | * Harvard IV Dictionary, mapped to 77 categories * PCA to derive pessimism score. “Weak” and “Negative” category explains 57% of variation | Article Sentiment Score based on Ratio of Negative Words over Total words | Dow Jones Industrial Return based on SMB (small-minus-big) | VAR -vector auto regression | * High pessimism predicts stock price downward, followed by reversion to fundamental. * Coefficient at 5% significant level * Harvard IV categories is built on physiology data, may not correctly reflect tone in finance |
| 2 | Tetlock, Saar-Tsechansky, and Macskassy (2008) | * Dow Jones News Service * Wall Street Journal * S&P 500 stocks * 1980-2004 | Firm names appear in fist 25 words | * Unigram * Fundamental variables such as earning and expected return | Harvard IV Dictionary, mapped into only Positive/Negative Category | Article Sentiment Score based on Ratio of Negative Words over Total words | Firms’ stock return | VAR -vector auto regression | * Fraction of negative words in firm-specific news forecast low earnings * Firms’ stock prices briefly underreact to the information embedded in negative words; and * Fundamental related news has the largest predictability power on firm’s earning and stock price returns * Stronger prediction model which considered fundamentals indicators |
| 3 | Loughran and McDonald (2011) | * 10K Filing * 1994-2008 | Data already tagged to stock in source | Unigram | Harvard IV Dictionary | TF-IDF | Firms’ stock return | Correlation | * Concluded that 73.8% negative words in HVD-4 dictionary are not negative in Finance domain * Manual negative word list which more correctly reflect negative tone in financial text |
| 4 | Hagenau, Liebmann, Hedwig, and Neumann (2012) | * German and UK Corporate Announcement * Pricing data from Datastream * 1998-2010 | * News source already tagged to company * 50 words minimum in article | * 2-Gram * 2-Words Combo | * 2-Gram * 2-Words Combo * Noun Phrase removal * Select high value features based Chi-square and Bi-Normal separation with exogenous market feedback | Log of word frequency in a message | Stock Direction | SVM | * Test Accuracies: Unigram (55.6%), 2-Gram (58.9%), Noun Phrase (61.2%) and 2-Word Combo (63%) * Chi-square feature selection improved accuracy for all features type, approx. 5% * Feature selection also reduced overfittings shown by reduction in training accuracy but improve in test accuracy * Feature selection incorporated market feedback, in feature selection produce better result * 50 minimum words filter may limit potential useful observations |
| 5 | Schumaker and Chen (2009) | * S&P 500 Stocks * Yahoo Finance * Oct-Nov 2005 | News already tagged to stock by news provider | Unigram | * Bag of words * Noun Phrases * Proper Noun * Named Entity * Minimum three occurrences per article | Binary | Stock Direction | SVM | * Average 57.1% directional accuracy on different feature types. Best accuracy at 58% when using Proper Noun feature * Simulated return average at 2.06 * First attempt to explore using Noun as feature * Using Noun as predictor is counter intuitive, active result could be overfitting with just one month of data |
| 6 | Nassirtoussi, Aghabozorgi, Wah, and Ngo (2015) | * EUR/USD currency pair from FXCM * News from MarketWatch (header only) * 2008-2011 | None | Unigram | * Stop words Removal * WordNet Dictionary * Hypernyms with WordNet * Sentiment Score using SentiWordnet | SumScore (combined positive and negative score) weighted by TF-IDF | Forex Pair Direction | SVM, remodeling every 2 hours | * Accuracy of 83.33% on test sample * 2-hours sliding window modeling uses most recent data for prediction, given very good result * Multi-layered dimension reduction method reduced modeling time * Uncommon way to measure sentiment by magnitude rather than direction * Only single source of news limits the number of observations |
| 7 | Li, Xie, Chen, Wang, and Deng (2014) | * 50 Stocks in Hong Kong Stock Exchange * News from FINET * 2003-2008 | News already tagged to stock by news provider | Unigram | * Use Harvard IV (HVD), Loughran McDonald (LMD) and Sentinet (SN) Dictionary to construct sentiment space * Stop words removal | Sentiment Score | Stock Price Direction | SVM | * Minor differences in accuracy on different dictionary used * Dictionary based approach achieve better result then bag of words * Sentiment polarity alone does not perform well in all test (<50% accuracy, with exception of Financial sector stocks) * Defined a range Neutral return, rather than absolute zero * Study covers both individual stock, and Sector * Manually annotated dictionary is not scalable |
| 8 | Kim, Jeong, and Ghani (2014) | * KOSPI Index (Korea) * News from two media company * 2011 | None | Unigram | * Stop words removal * Noun removal * Develop automated wordlist (based on stock price direction) from most frequent 700, to derive word sentiment score using TF-IDF | Article Sentiment Score calculation not mentioned | Stock Direction | None | * Best tuned accuracy of 65.2% in prediction price direction, F1 score of 62.6% * Automated annotation is fast and scalable * Result could be bias when automated annotation is based on only just one year timeframe |
| 9 | Im et al. (2014) | * BERNAMA, TheStar, Sun Daily and NST * KLSE Stocks * 2013 (6 months) | None | Uni-gram | * Stop words removal * Stemming * Develop manual word list of 8000 terms into positive/negative | Article Sentiment Score based on ratio of Positive over Negative word counts, applied on Headers and Body with different weightage | Stock Direction | None | * Accuracy of 90.8% when sentiment on both article and body are equally important. * Apply stemming improved accuracy by 5% * Experiment design considered weighted impact of header and article, which mimic human reading pattern * High accuracy could be a result of data mining because data period is only 6 months * Manual annotated word list is hard to scale |
| 10 | Hor et al. (2018) | * KLSE top 10 large capital stocks from Trading Service sector * The Edge news * 2012-2017 (dividend periods only) | Entity match-ing using Azure Text Analy-tics | Uni-gram | Stop words removal | Article Sentiment Score provided Azure Text Analysis service (pre-trained model) | Stock Price Direction | None | * Majority news were classified as Neutral by Azure, hence correlation to price direction were nearly null * Experiment design and deployment is simple using cloud * Scope on dividend period limiting data points for stock sentiment analysis * Alternative sentiment measure and tuning not explored |

### News Semantic

Business related text data sources could potentially contain relevant and valuable insights that could be captured and interpreted by organization that want to stay at the competitive edge. below shows an example of two business news about Apple and Google respectively. Google's stock price fell after a poor earnings news was announced. Apple's stock price fell the day after news about former CEO Steve Jobs passing.



Figure .: Sample Business News, (Ding, Zhang, Liu, & Duan, 2014)

Capturing the information is the easiest part. The most difficult task would be to understand the underlying meaning of what had been written and scaling the process with high speed and volume. Meanings need to be concluded through the text mining process for fund managers to make investment decisions (). In general, the process of giving meaning to text is called semantic analysis.

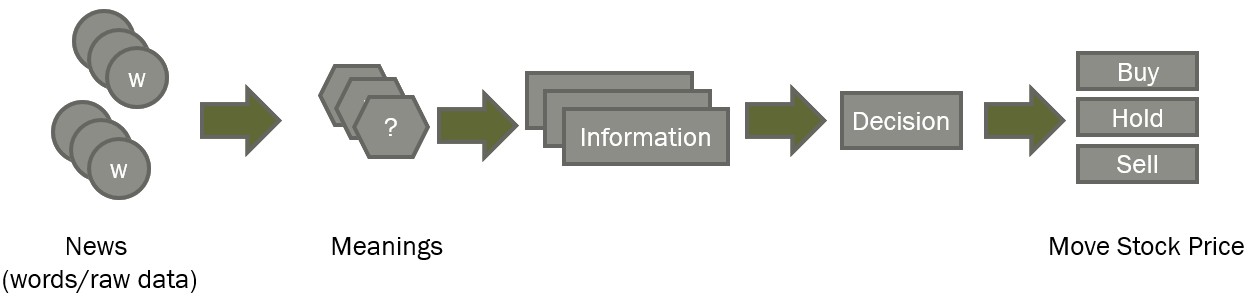


Figure .: From News To Decision

Humans understanding of what has being said is an unconscious process. This was built upon years of experience on the language learned together with the concepts of the text where they were used. However, this process is extremely complicated for machines. One of the approaches used by machines, is based on keywords and its statistics. A common example is machine learning uses keywords frequency count as “clue” to the meaning of text. A common application of this approach is Sentiment Analysis, which provides scoring of positive or negative sentiment to text data. This simple sentiment feature does not identify key events of the financial news, hence even more challenging to interpret its impact on stock market. As an example, the event “Apple has sued Samsung for copying its patented ‘look-and-feel’ of its tablet and smartphone”. With only term-based features {“Apple”, “sued”, “Samsung”, “copying”, ...} could inaccurately predict or represent the sentiment of the news and its impact to Apple stock price movements to both Apple and Samsung. These terms were not structured in a way where actor and object can be identified.

In (Ding et al., 2014), the authors proposed using structured information in representing news event and developed a prediction model to uncover potential relations between the stock price and news events. A two steps approach is taken, first to develop a scheme to extract the events, second to generalize them. The scheme represents the event as a (O1, O2, P, T), where object O1 is the actor and O2 is the object being acted upon. P is the action and T is the time taken for the action to be performed. As an example, the news headline “Sep 3, 2013 - Microsoft agrees to buy Nokia’s mobile phone business for $7.2 billion.” is mapped to (Actor, Action, Object, Date) tuple with values (Microsoft, buy, Nokia’s mobile phone business, Time = 3rd Sep 2013). Event tuple extraction is performed using the Open Information Extraction technique. For each event phrase, a verb is identified as the action (P) and the object left to the verb is the actor (O1). To the right of the verb is the object being acted upon (O2).

In another example, consider the sentence “Apple sues Samsung for 1 billion dollars.” The predicate verb is “sues”, object and subject are “Samsung” and “Apple” respectively. The representation of event is tuple of (Apple, Samsung, sues).

A more advanced approach is to process the logical sentence structures in the text to identify relevant elements with objective to understand the discussion topics. These techniques aimed to build relationships between different concepts discovered. For example, a model would understand that the text is related to “politics” and “economics” even without the actual words, basing on similar concepts to “election,” “Democrat,” “speaker of the house,” or “budget,” “tax” or “inflation”. Using the same concept, authors in (Wang, Zhe, Kang, Wang, & Chen, 2008) built a knowledge-base to casually relate financial articles to stocks, in a way knowledge is captured and preserved.

Information extraction can be a costly task and usually requires predefined grammar and domain specific detection rules. For this reason, the simpler approach using bag-of-words representation was found to be the most commonly used method. Whichever method that is used to analyze the text and sentences, behavioral scientists are becoming more skillful at uncovering the evidence of sentiment hidden in text.

### Feature Processing

Before the availability of advance computing hardware and software, text data was manually analyzed by team of experts. Considering the current big data era, large volume of news data generated daily had made human processing infeasible. We have summarized three common aspects of the pre-processing steps from our reviewed works, namely feature selection, representation and dimension reduction. illustrates an overview of the processes.

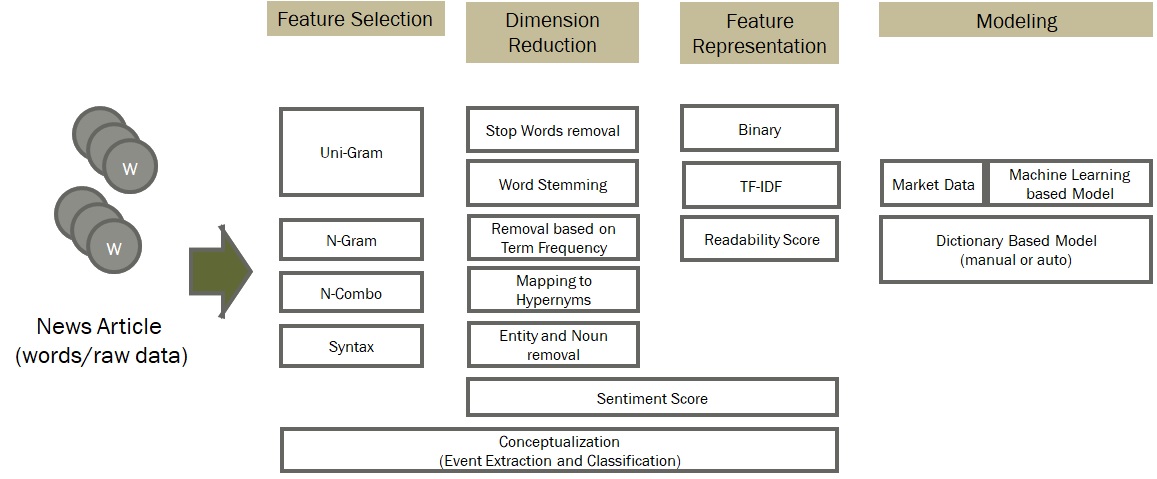


Figure .: News Feature Processing

A variation of N-Grams called the "N-Combo" is used in (Hagenau et al., 2012), where the authors chose two non-adjacent words as their feature, as the basis to form a readability score.

Another feature selection is based on grammar syntax. Using this technique, sentences are parsed into Part Of Speech (POS) as additional variables, as described in (Bharathi & Geetha, 2017). In the paper, nouns and adverbs were picked up by the POS tagger and used as part of the features to derive the sentiment score.

### Dimension Reduction

Text data are sparse when it is mapped into a feature matrix of word and frequency per article. This is simply because English is a language rich in words and their usage is usually scattered across an article. In a large corpus, this may result in a large number of features which is inefficient for computer processing. Using all features would not only slow down processing, but also produce less desirable results due to the noise.

One common way to reduce word features is to simply discard non-meaningful words, for example "the", "is" and "am". These so called "stop words" are known English terms that do not contain any information that may influence news sentiments. Work by authors in (Li et al., 2014) and (Kim et al., 2014) had used and deployed this method. Further to known stop words, authors of (Schumaker & Chen, 2009) and (Kim et al., 2014) also discarded words that have appeared as English Nouns using Part of Speech tagging. The authors deemed that the entity name detected as a Noun do not contain any sentiment value. Authors in (Schumaker & Chen, 2009) reported the best accuracy by deploying Noun removal. It is a convenient method, as POS tagging can be performed automatically by a program.

Another method is to generalize terms into their hypernyms by using a thesaurus. Work in (Nassirtoussi et al.) used the WordNet Synset to map multiple terms into a single common word. For example, "trillion", "billion" and "million" are mapped to "large numbers".

### Feature Representation

The most fundamental way to represent words as features is to use 0 and 1 (binary) to indicate the presence of the word in an article. Alternatively, word occurrences in articles that are used as frequency count is also known as Term Frequency (TF). Feature representation could happen at the feature selection and dimension reduction phase. In (Kim et al., 2014; Nassirtoussi et al., 2015; Schumaker & Chen, 2009), authors select features based on term frequencies filtered with minimum of occurrences and top-N most frequent terms.

In 1972, authors in (Sparck Jones, 1972) published a method of measurement called Inverse Document Frequency (IDF). For a term being searched, this method measures the frequency of keyword occurrence among the documents within a collection. Intuitively, it means terms that occurred many documents is not ideal discriminator compared to terms that occurs in fewer documents, hence should be given less weight. Coupled with TF, a well-known measurement of TF\*IDF, is a big improvement in information retrieval and had proved to be extremely robust and hard to compete (Robertson, 2004).

The result from previous discussed features representation, can further be used to derive new features that are closer to practical application. News text data are more meaningful when it represents sentiment. Sentiment score is typically represented with a number ranging from -1 to 1 or 0 to 1, where the lowest boundary represents extreme pessimism and highest boundary represents extreme optimism. Deriving a sentiment scoring system was a major focus in many financial NLP analysis, as seen in (Bharathi & Geetha, 2017; Kim et al., 2014; Li et al., 2014; Loughran & McDonald, 2011; Nassirtoussi et al., 2015; Tetlock, 2007; Tetlock et al., 2008).

One interesting representation of text data is the Readability Score, which measure the difficulty level of text being analyzed according to an education level. Authors in (Butler & Kešelj, 2009) deployed a readability analysis to predict the stock return and achieved better precision than N-Grams. The technique used had demonstrated that that computer programs had the ability to partially “understand” the annual financial report and derive prediction for future performance. The authors used programming packages to calculate the readability score based on Flesch, Flesch-Kincaid and Fog Index.

## Prediction Model

### Dictionary Based Approach

This is a computer-based, generalized content analysis system which as an automated code mapper. In such a dictionary, words are read by a computer program and assigned manually by hand, to one or more categories. This is commonly known as the "bag-of-words" approach in NLP. below shows a generic process of using the said approach to extract text sentiment text.

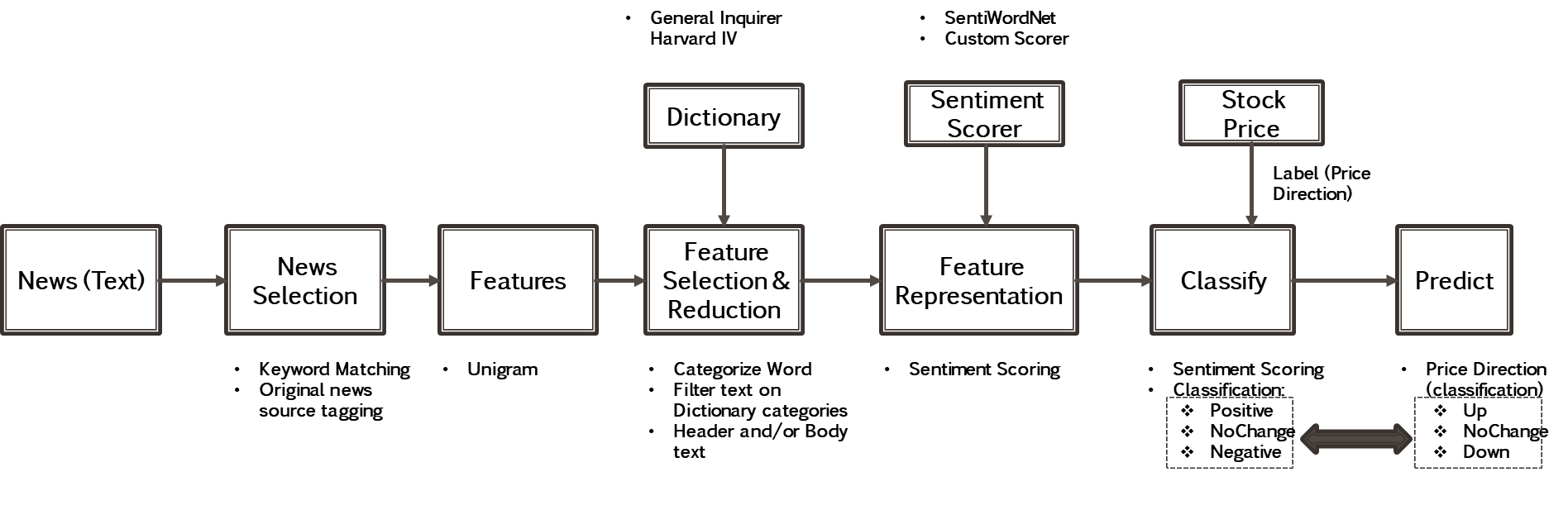


Figure .: Dictionary Based Approach

More than fifty years ago in Stone, Dunphy, Ogilvie, and Smith (1966) the author developed a system called General Inquirer (GI), which uses the Harvard Psychosocial Dictionary to parse and classify text. Suicide notes written by total of 66 men (33 actual, 33 simulated) were tested by system. Till today, General Inquirer remain as the popular dictionary for text sentiment analysis.

One of the best known early NLP work in Finance that uses General Inquirer is (Tetlock, 2007). The authors employ the GI with the Harvard-IV dictionary to analyze news text in the "Abreast of the Market" column of Wall Street Journal, using frequency of negative terms as the nominator and positive terms as denominator. Since (Tetlock, 2007) various work deployed similar approach in measuring positive versus negative toned words within finance domain. In particular, authors of (Loughran & McDonald, 2011) fine-tuned the Harvard-IV dictionary in the context of finance accounting when classifying terms into positive and negative. They had performed regression analysis on US 10K filing data with variations of stock data, such as returns and size. They concluded that 73% of the negative terms in Harvard-IV dictionary typically are not negative finance terms. As a contribution, the word list from this research is open for future use.

Most studies employ proportional weighting, which treats every word in the dictionary to be equally weighted. For example, work in (Tetlock, 2007) implicitly assumed that all positive and negative words in the work list are of equally strength within the two categories, which means no weighting is applied. In another work from (Hagenau, Liebmann, & Neumann, 2013; Loughran & McDonald, 2011), the weight of each term is inversely proportional to its document frequency (TF-IDF). In (Jegadeesh & Wu, 2013), the authors argue that term weighting is an important accurate compilation of words in a dictionary.

### Machine Learning Based Approach

Machine learning relies on statistical methods to infer input features and classify them based on statistical inferences. Once features are finalized, a portion of the data would be used to train a model against a known class (or label). A trained model could then be reused to predict new data into its class. For financial news analysis, a label could be generated automatically or manually. below shows the overview of the process.

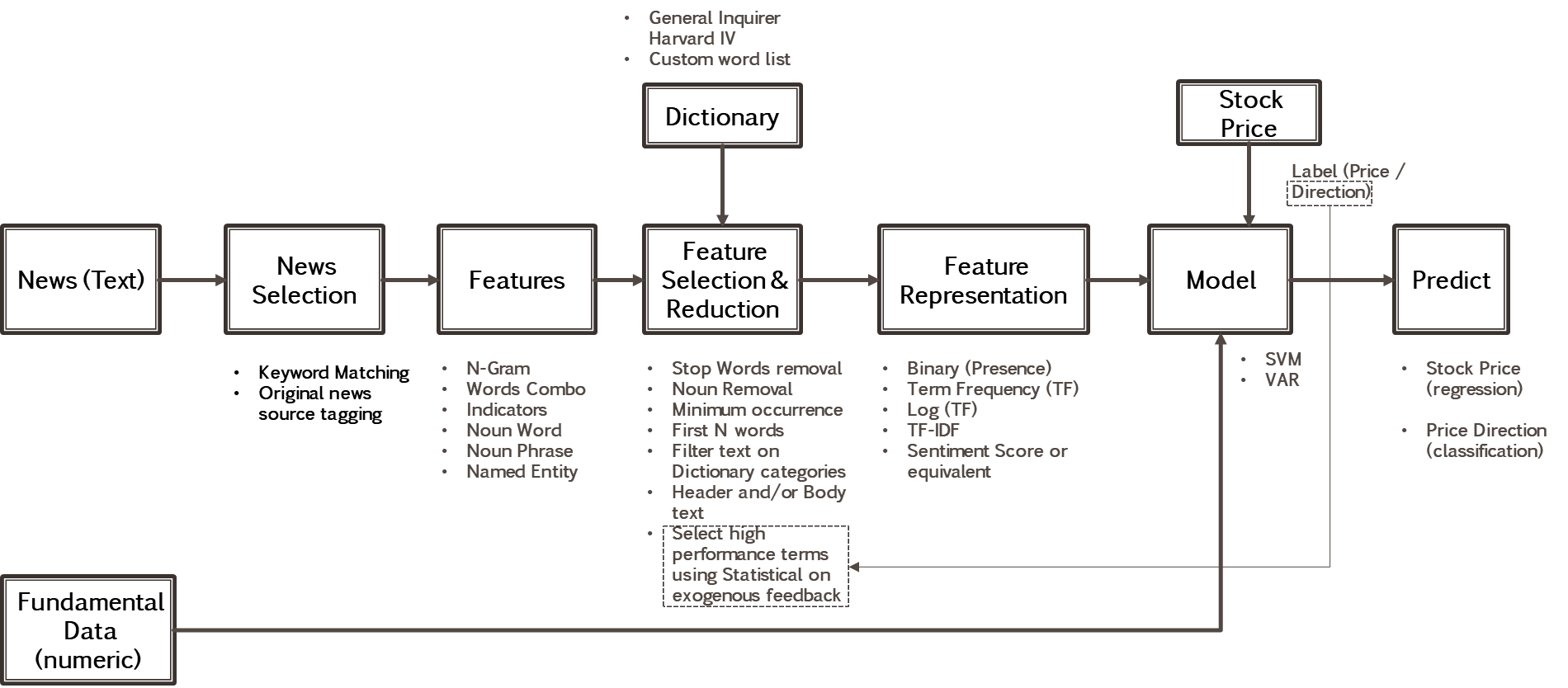


Figure .: Machine Learning Based Approach

In (Hagenau et al., 2012), a label is generated automatically based on market data, such as "UP" when the returns are positive and "DOWN" when the returns are negative. Using a SVM model on selected words as a feature, the authors achieved a 65.1% accuracy in the validation dataset. (Li et al., 2014) also deployed a SVM model on a automated label using a similar method, however only when compared to scenarios using raw words as a feature and only in sentiment measure. The authors found that using a purely sentiment measure does not perform well.

## Integrating News Categorization

Authors in (Shynkevich, McGinnity, Coleman, & Belatreche, 2016) investigate if financial related news has any influence over targeted stock, and if that can provide an edge for stock price forecasting. Stocks are classified to five news categories using Global Industry Classification Standard (GICS). This well know classification consists of five subcategories of individual stock, sector, industry, subindustry and industry group. Hence a total of five subsets for relevance analysis.

shows a prediction model proposed by (Shynkevich et al., 2016). News articles are downloaded from LexisNexis. LexisNexis offers news relevance scoring for its relevant companies, as a supplement to its news data. The news article that has a relevance score of more than 85% is considered as for the corresponding subset. Authors in (Shynkevich et al., 2016) classify news articles as positive or negative in following trading day, whether the stock price went up or down after the news publication. Each news header and body are combined into a pool of words and word stems are extracted to form a list of unique features. Association by Chi-Square is performed to sort the 500 unique features in descending order according to the scores. Subsequently, the author engineered a new feature to represent each news article as a vector of based on 500 \* TF \* IDF. This results in a sparse matrix size of [number of data observations] \* 500. This procedure was applied separately to other subsets. Each observation is labeled as ‘Up’ or ‘Down’ for machine learning. This methodology is based on the first principle without using any pre-defined dictionary for preprocessing (Shynkevich et al., 2016).

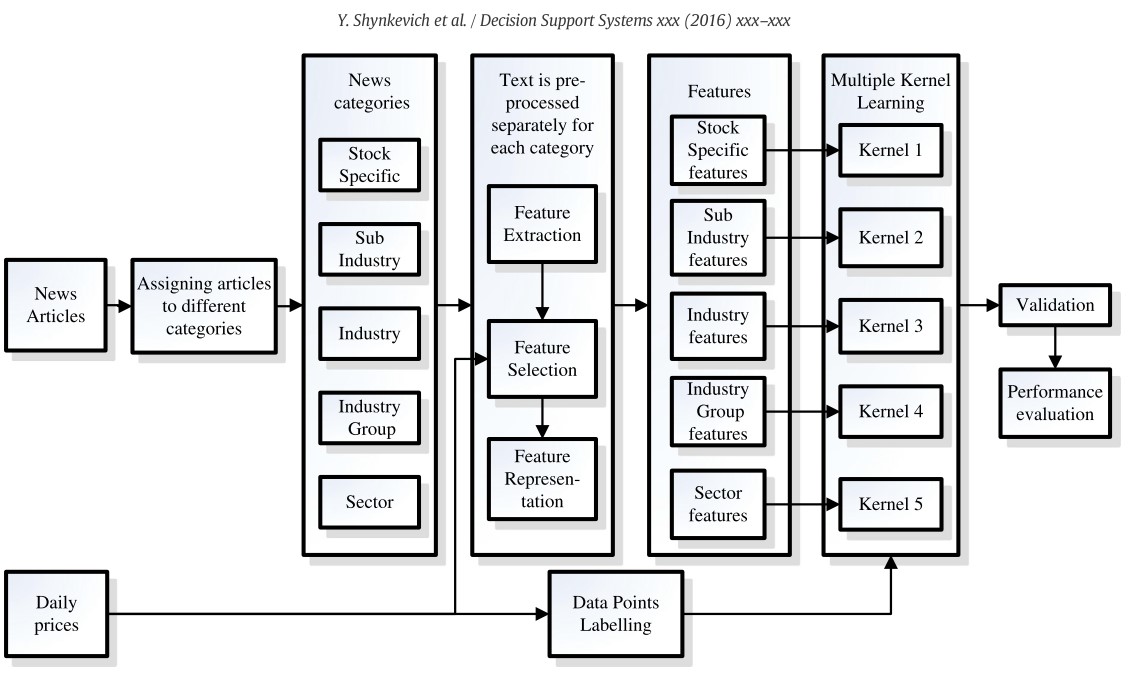
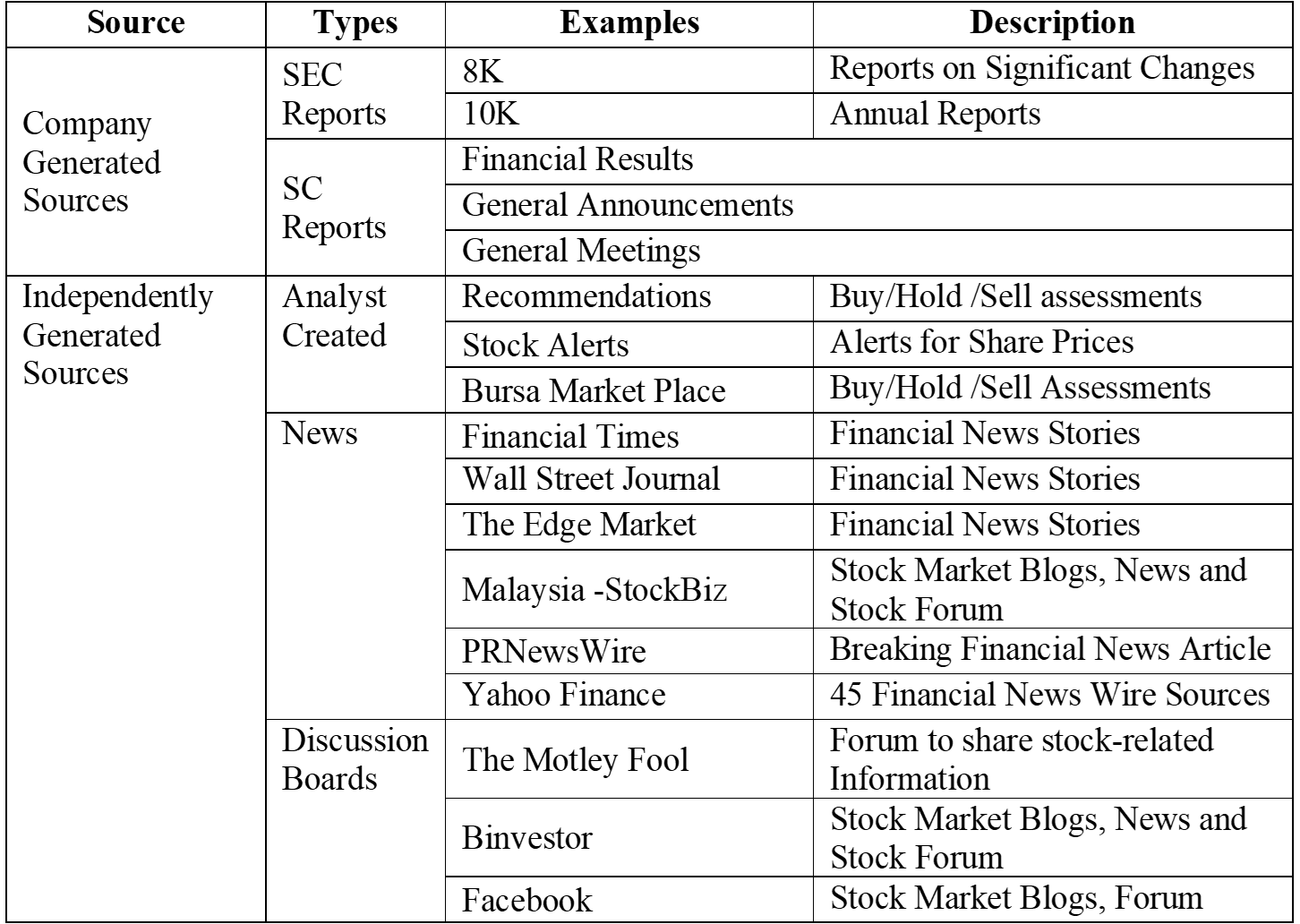


Figure .: Proposed Predictive System, adopted from (Shynkevich et al., 2016)

## News Sources

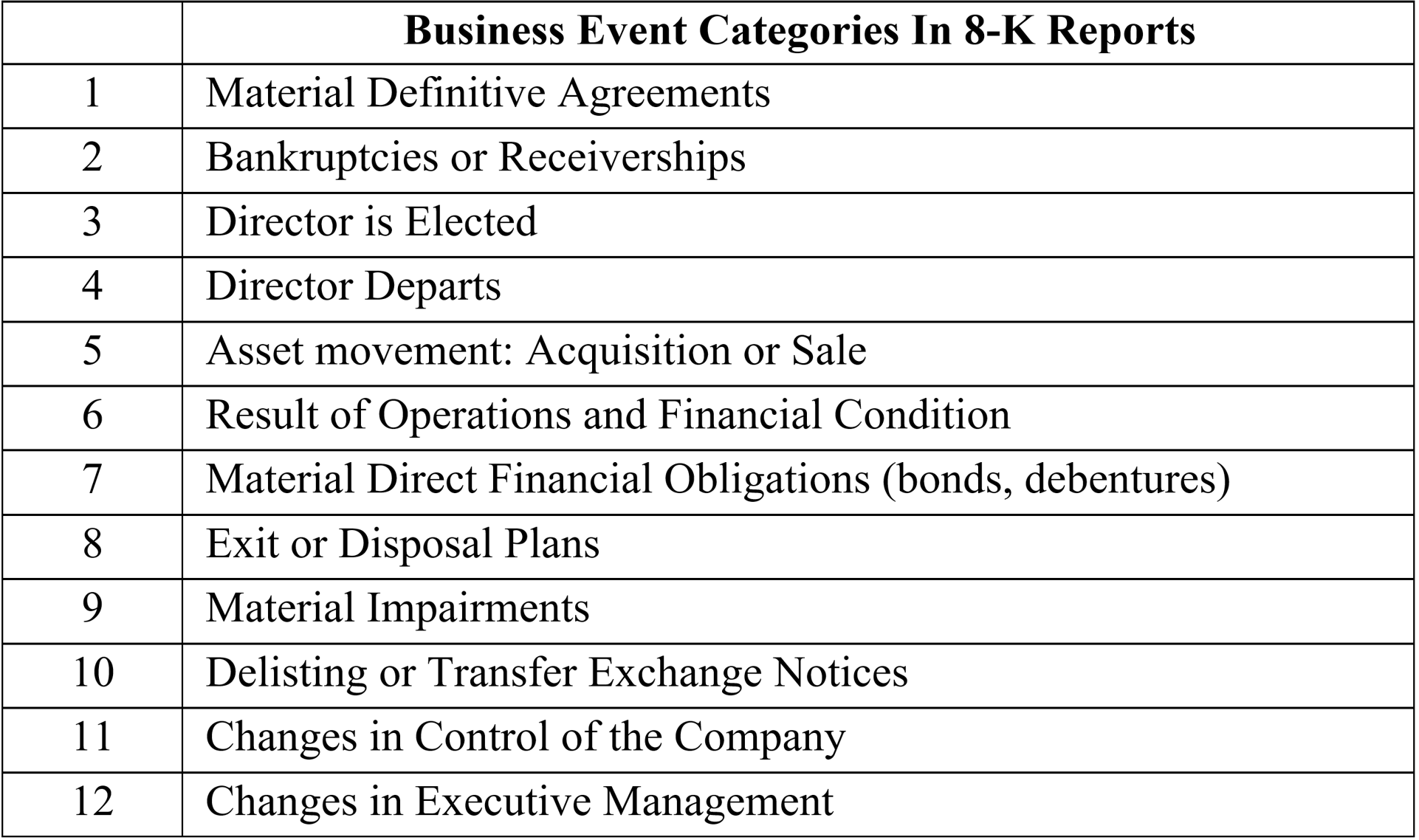
Multiple sources of textual data can be analyzed for stock price prediction, as resented below (). Authors in (Schumaker & Chen, 2009) state that there are two sources of textual data: computer-generated and independently generated sources. Textual information in a computer-generated source (quarterly and annual reports), is not explicitly expressed by financial ratios, but rather encapsulated with meaning for discovery. Analyst reports, news, and discussion boards can provide a more generalized view of a company (Schumaker & Chen, 2009). These textual data can be classified into categories which are relevant to stock market price prediction.

**Table 2.2: News Event** Classification **In 8-K Reports (Schumaker & Chen, 2009)**



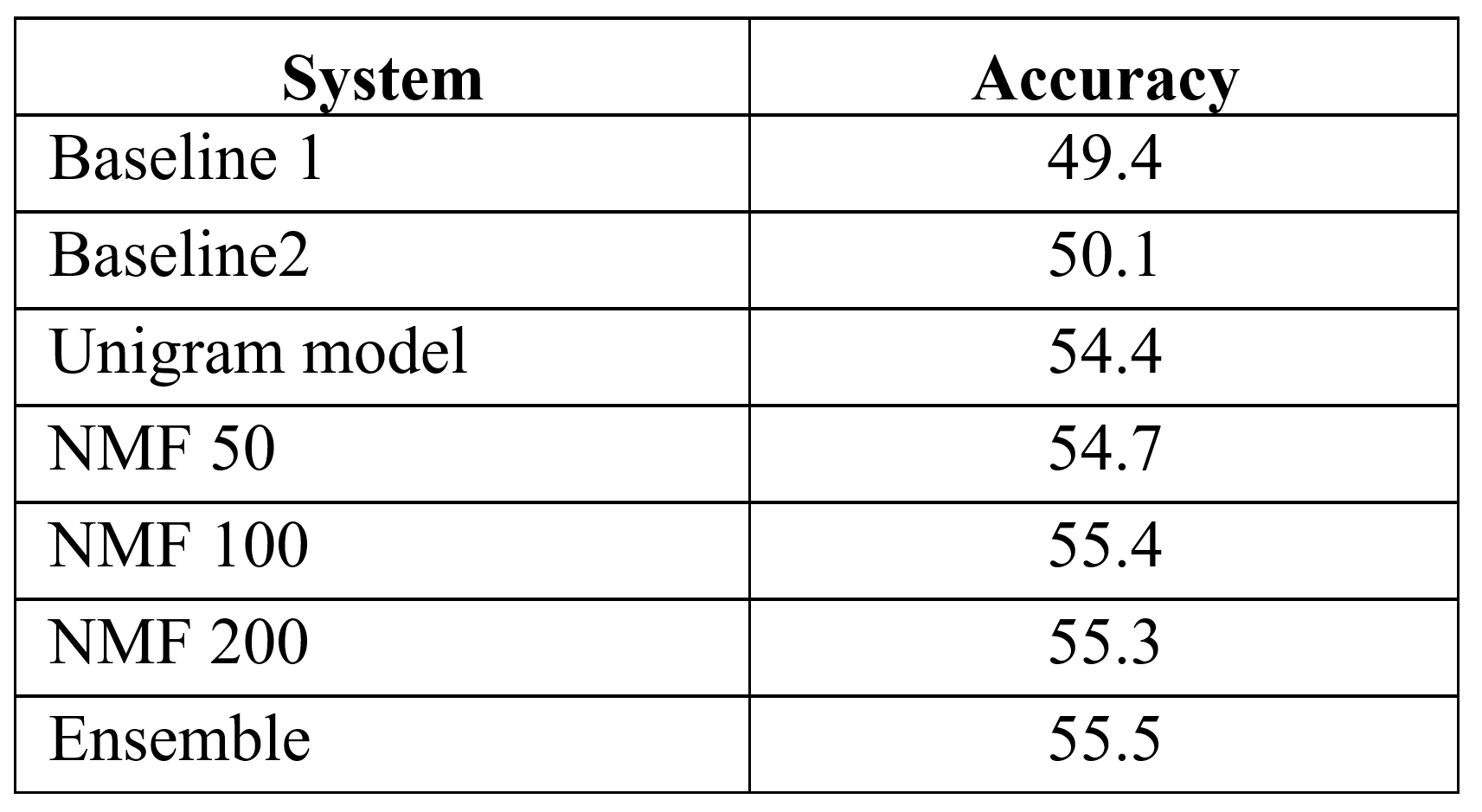
Authors in (Lee, Surdeanu, MacCartney, & Jurafsky, 2014) mention that public listed U.S companies are required to file 8-K reports whenever there is a significant business event, including layoffs, bankruptcies, layoffs, the election of a director, etc (Table 2.3).

Table .: List Of Financial Event Types In 8-K Reports (Lee et al., 2014)



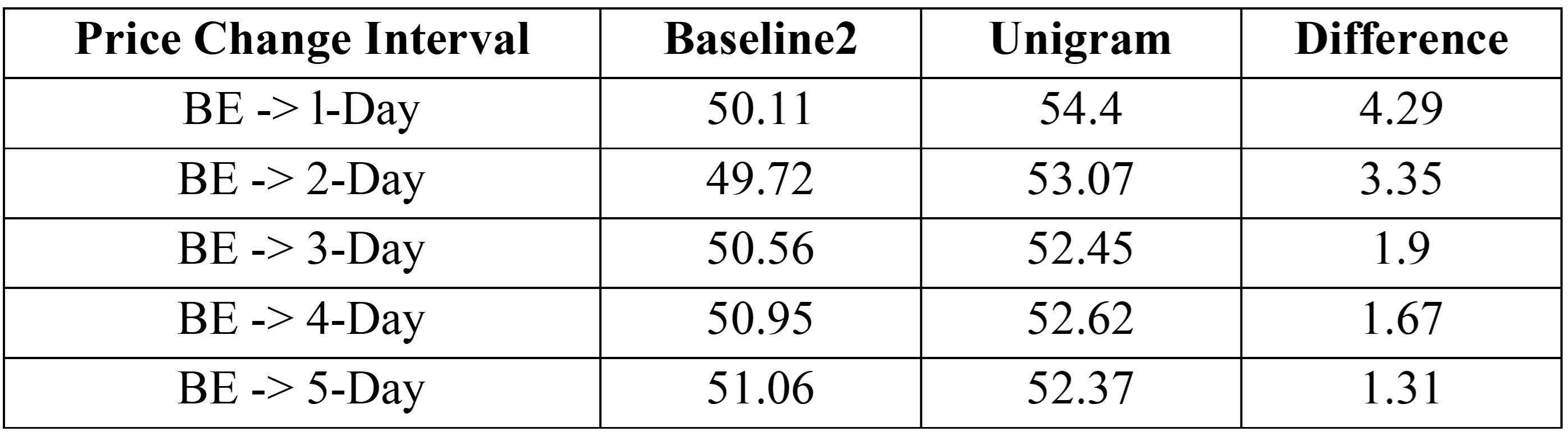
They build a corpus that aligns the description of financial events as reported in 8-K documents and aligned with the corresponding stock price. Company stock price difference before and after the publishing date of the report is calculated for each filed document, which is then converted into an Up/Down/Stay label. Authors in (Lee et al., 2014) then generated event categorical features with four non-linguistic feature types, including the volatility index, stock price difference, earning surprise and event category. For the linguistic features, unigrams and non-negative matrix factorization (NMF) are used. The features used in each model for the experiment are shown in Table 2.4. The features are used to forecast stock price movement (UP, STAY and DOWN), corresponding to an event. The features and the labels are then used for model training (Lee et al., 2014). summarizes the accuracies. All non-language based features achieve around 50% accuracies, whereas unigrams with negation heuristics based approach improve the accuracies by 4%. More improvement is observed in linguistic features compared to non-linguistic features (Lee et al., 2014).

Table .: Results Of Proposed Model (Lee et al., 2014))



below from the same paper shows the prediction performance of linguistic and non-linguistic features as it shifted farther to future from the event date. The linguistic feature’s predictive power is higher in the short term and diminishes quickly with time. The non-linguistic feature (event etc) exhibits a relatively flat trend of predictive power for varying time intervals and is somewhat better than linguistic features in the sense that its predictive power is retained after some time.

Table .: Result of Temporal Aspect Model (Lee et al., 2014)



Similar research conducted by (Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2013) also affirmed the findings. The authors quantify news articles by extracting relevant events using proprietary application. They concluded that the assigned expert-defined impacts of the extracted events has positive correlation with results from events driven models. Authors in (Mitra & Mitra, 2011) also concluded positively on using a news-driven investment strategy.

## Conclusion

A general framework on feature processing and transformation is discussed, which are common techniques based on dictionary and machine learning. The idea of dictionary is to use sentiment categories based on pre-trained dictionary to derive sentiment score. Machine learning on the other hand automatically “learn” the sentiments based on data and market feedback. We also seen that dictionary-based technique can also be incorporated into machine learning approach. We had seen researchers used news categorization and event extraction methods to influence news “relevance” scores. This results in fine grain features allowing higher accuracy in stock price prediction. This leads to future research direction where news must be analyzed through different concept in order to model the complex finance ontology more accurately. Based on techniques discussed in Content Analysis section, we then explored how researcher integrates their model with different news sources. News sources of different types may contain useful information beyond basic sentiment based purely on words without concept. A paper reviewed using US 8-K filing data, which is structured in a way complying to US regulatory requirement. Having a structure in the content allows researcher to exploit different ways to derive meanings through event extraction and categorization.

# Research Methodology

## Terminology Used

### Financial Terminology

Table .: Financial Terminology

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Market Capital | Total dollar value of outstanding shares. It is calculated as total number of shared outstanding multiply by the current market price of a share. It is commonly used as one of the metrics to select company of different sizes for portfolio diversification. |
| Out Standing Shares | Refer to a company's stock currently held by all its shareholders, including share blocks held by institutional investors and restricted shares owned by the company’s officers and insiders. It is used in calculating key metrics such as a company’s market capitalization, as well as its earnings per share (EPS) |
| Open Price | First traded price that started the day |
| Close Price | Last traded price of the day |
| Reuters Instrument Code (RIC) | It is the short code used by Reuters to identify a company |
| Initial Public Offering (IPO) | The date when a company transition from private to a public company. It is also the date when the stock is listed in Stock Exchange. |
| Exchange Ticker | It is the short code where Stock Exchange used to represent the listed company |

### Data Science Terminology

Table .: Data Science Terminology

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Application Programming Interface (API) | An software component that bridge two applications together through an agreed communication parameters. |
| DataFrame | A Python data structure representing a 2-dimensional table with variables as column and rows representing records. |
| Natural Language Processing | Text processing in the context of natural language as automated by computer program. Typical application are speech and text related system. |
| Open Source Software | Software where source code are shared publicly which encourage modification and enhancement |
| Wordnet | WordNet is an English language based lexical database for conceptual and relationship lookup. It organizes words with the same similar concept and meanings together to describe the relationship, strength and distance. It is a project began by George A.Miller (Miller, 1995). |
| Senti-Wordnet | It extends WordNet synset with concept of sentiment score. (Baccianella, Esuli, & Sebastiani, 2010) |

## Data Sources and Acquisition

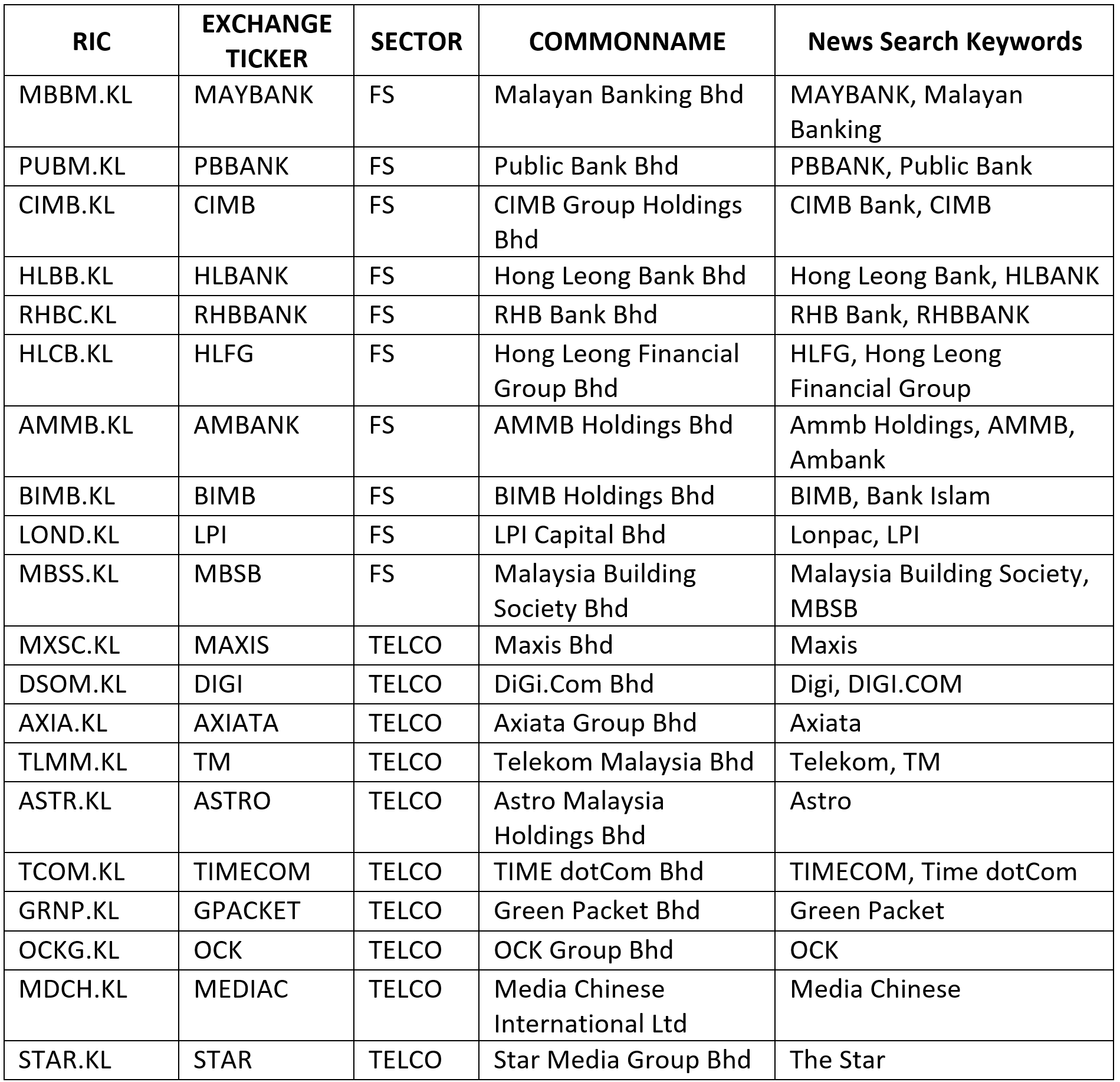
This section describe methods to acquire data from sources defined in Scope of Research (Section 1.5).

### News Acquisition (Text)

This research focused only on business and finance related news. To minimize noises from non business-related news, filters based on category and keywords are applied. News are acquired from provider’s web site. Due to the unstructured nature of HTML, this project deployed web crawling and text processing techniques to search and parse news of interest from the respective news provider website. A computer program took stocks keywords (Table 3.3) as the search parameters and loops through all stocks in scope. The algorithm is described in detail below:

1. Identify the URL of the news provider search engine
2. For each stock in scope (fetch the links):
   1. For each search keywords on the stock:
      1. Execute search on the keyword
      2. Obtain search result, which is re news summary, containing URL links to news detail news article
      3. Loop through every result page to scrap all URL links
   2. Compile all URL links for all keywords and remove any redundant URLs
3. For each unique URL (fetch the news headline and detail):
   1. Fetch HTML page for the news article detail
   2. Parse the HTML page to obtain News Date, Header and Body
4. Save the all fetched news into local disk

Table .: News Search Keywords



Screenshot (Figure 3.1) below illustrates a sample search page for keyword “Maybank” under “Business” section. Search returns news summary which contain links to article detail (Figure 3.2). Although the example shown is for TheStar, similar approach is applied to TheEdge Malaysia.

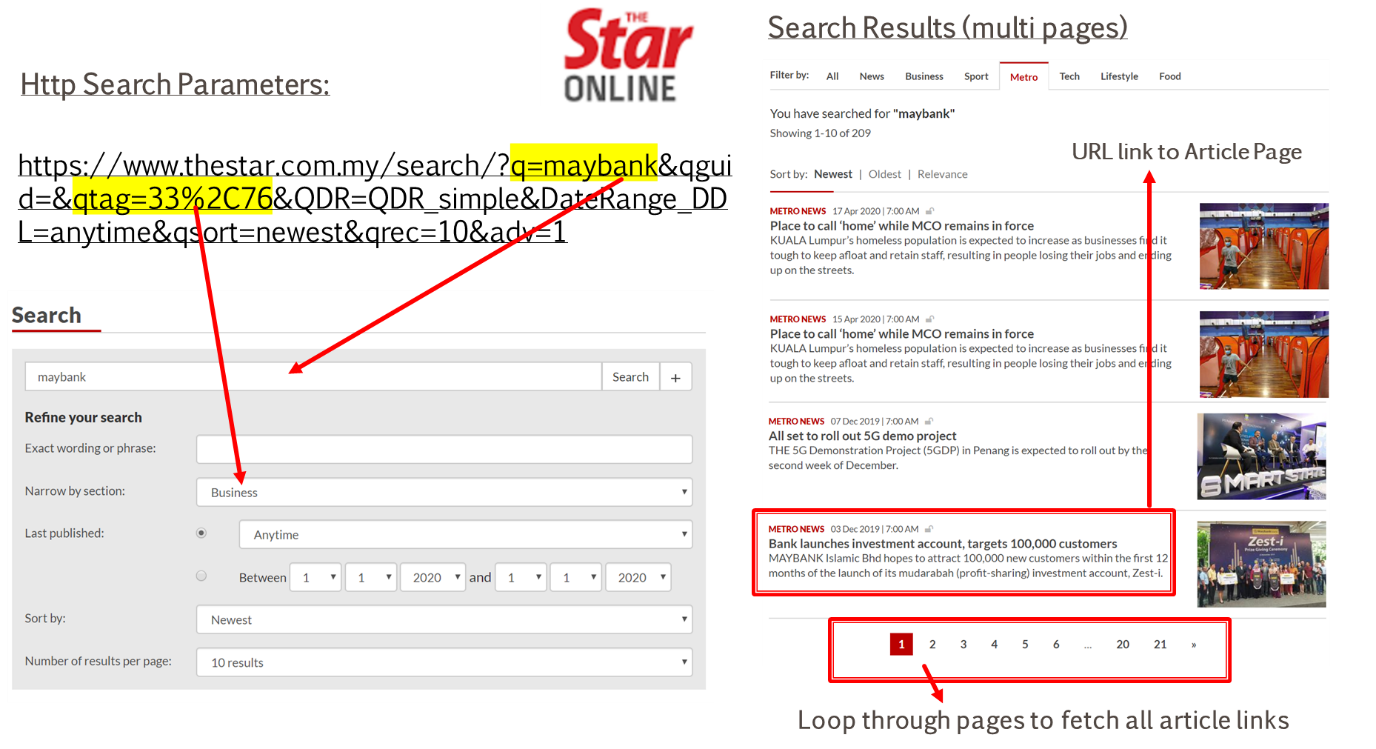


Figure .: News Provider Search Page

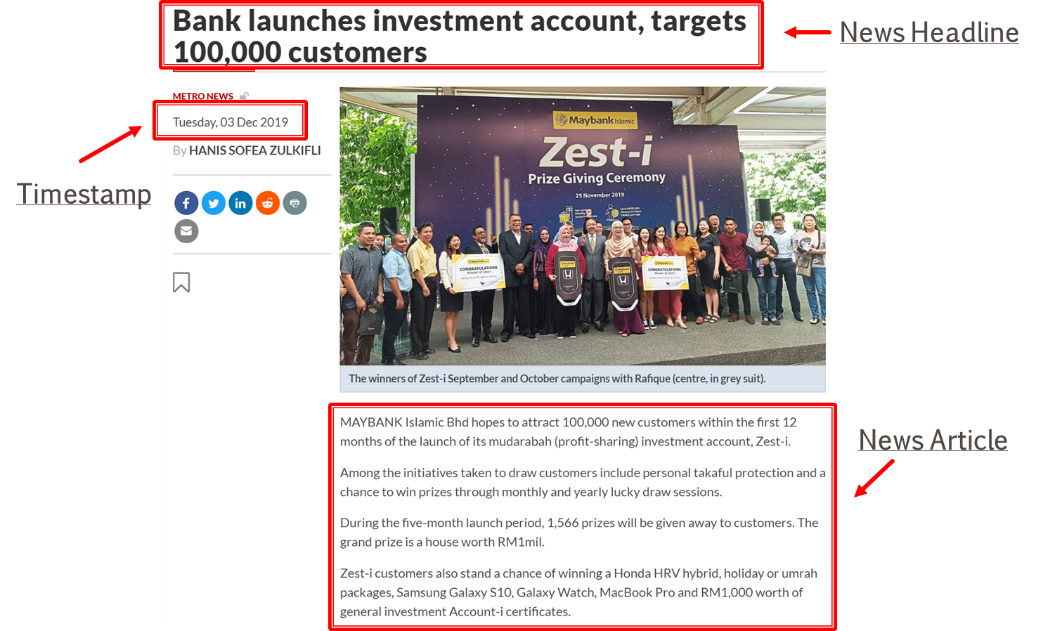


Figure .: News Article Web Page

### Stock Price Acquisition (Numerical)

Stock price data is relatively simpler as the data source is structured. This project utilized data service provided by Thomson Reuters Eikon platform. Eikon is a financial data platform which includes market, companies, and financial news data. Although Eikon support financial news, its coverage on KLSE is very limited (close to none for most stocks).

Many data platform had been explored (including Yahoo Finance, World Trading Data, I3 Investor, Alpha Vantage and Quandl), however, none of the alternatives provides sufficient data required for this research, specifically on KLSE. They either do not provide historical outstanding shares information or contain too many errors (wrong pricing, unavailability of adjustment, and lots of missing data).

This project used Eikon API (Figure 3.3) for three historical variables, namely OPENPRICE, CLOSEPRICE and OUTSTANDINGSHARES. Eikon imposed certain limits on frequency of API calls. Algorithm in this project minimizes requests sent to Eikon by making only single call per stock for all variables mentioned above. Below is the detail algorithm:

1. For Each Stock Within Scope:
   1. Eikon API Call To Get Variables/Column: [OpenPrice, ClosePrice, OutstandingShares]
   2. For Each Variable/Column:
      1. Remove rows with missing data
      2. Calculate new MarketCap variable by multiplying OpenPrice and OutstandingShares
   3. Align All Variables/Columns with Date as the Index Into A large 2-Dimensional Array
2. Consolidate Variables from All Stocks Into A Single 2-Dimensional DataFrame
3. Store the DataFrame into Local

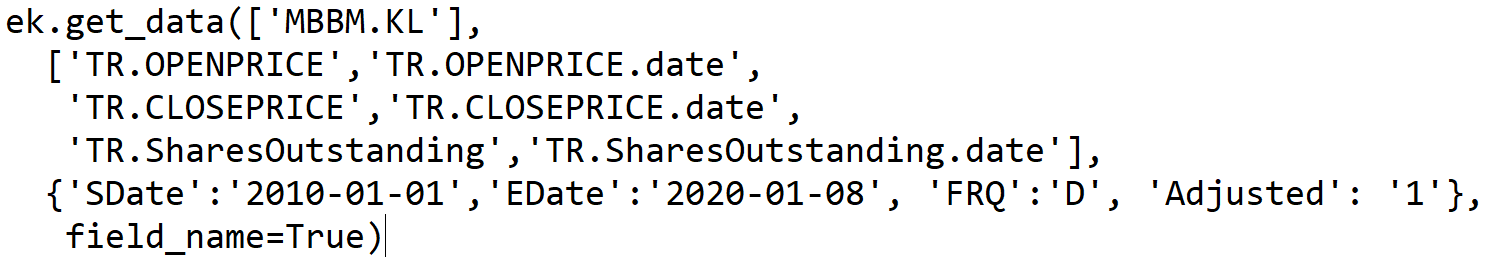


Figure .: Eikon API Call For Pricing Data In Python

It is important to note that “Adjusted” price and outstanding shares are captured for 10 years. Price adjustment reflects true value after the effect of corporate action, such as stock splits, consolidation and right issues. For example, when a stock split from one to two due to bonus issue, the share value will be reduced by half. This corporate action had occurred to CIMB.KL on 31st May 2010 (Figure 3.4 below). If the outstanding shares are not adjusted, market capitalization would be wrongly reported as halved overnight.

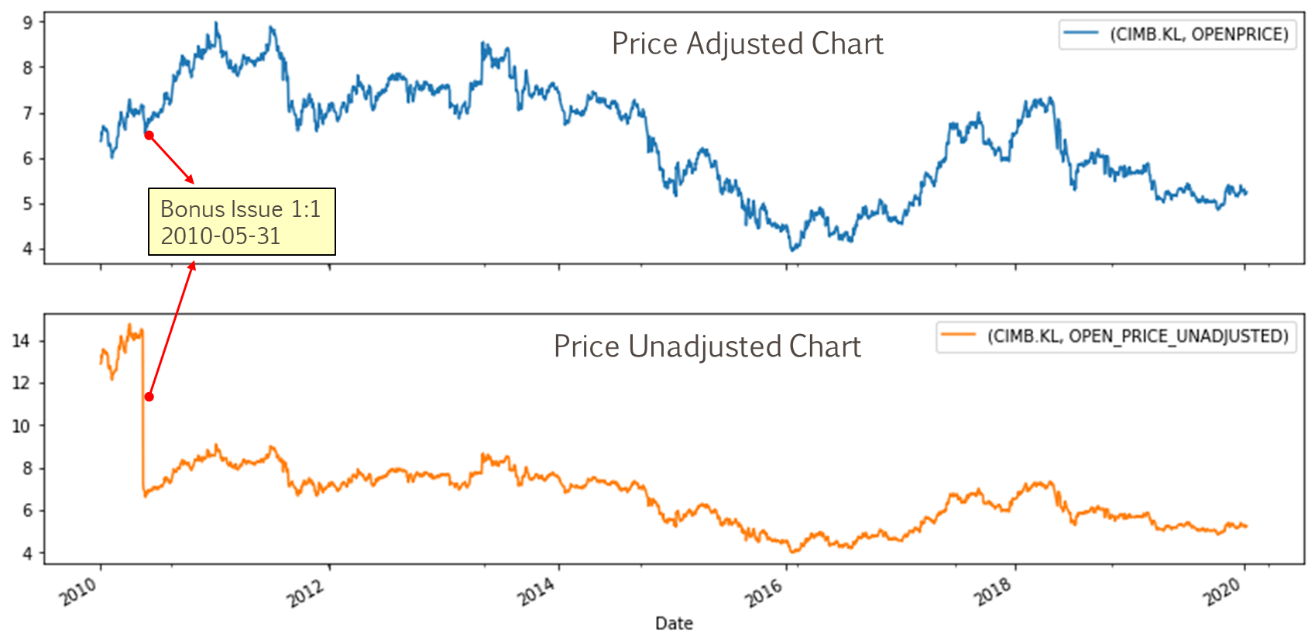


Figure .: CIMB Price Chart

## System Design

This project used Python as the only programming language for end to end data processing on a single computer. Except for Eikon which require license, all other components and libraries used are from Open Source. System diagram depicted in Figure 3.5 below shows the end to end data flow, from data acquisition to preprocessing, feature engineering, modeling and lastly generating results.

Two Python object library has been created as reusable components to acquire price and news from internet. These modules (EikonDatabase and NewsDatabase) supports incremental updates which can be used for future scope such as new timeframe and additional stocks.

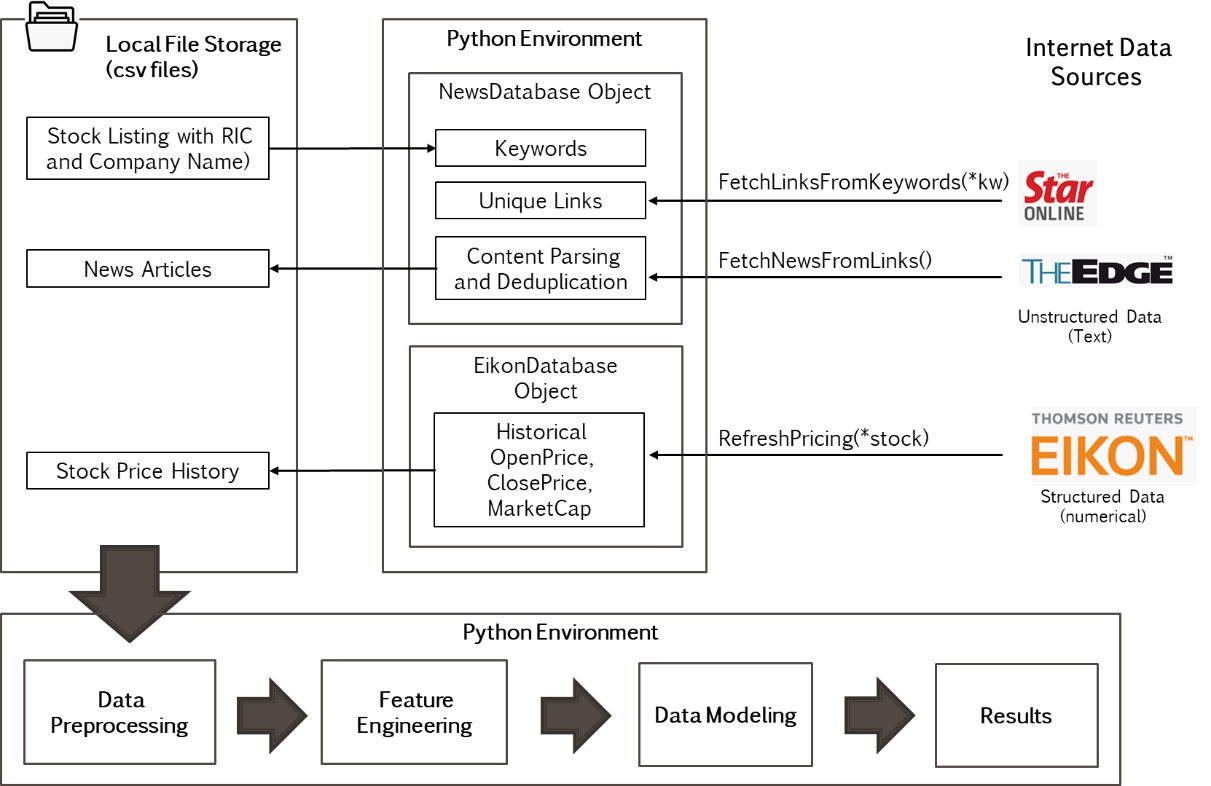


Figure .: Overall System Design

## Database Schema

This project deployed simple file-based database which consists of three (3) main tables shown below (Figure 3.6). \*RIC and \*ID are primary keys. Total data size is about 600MB.

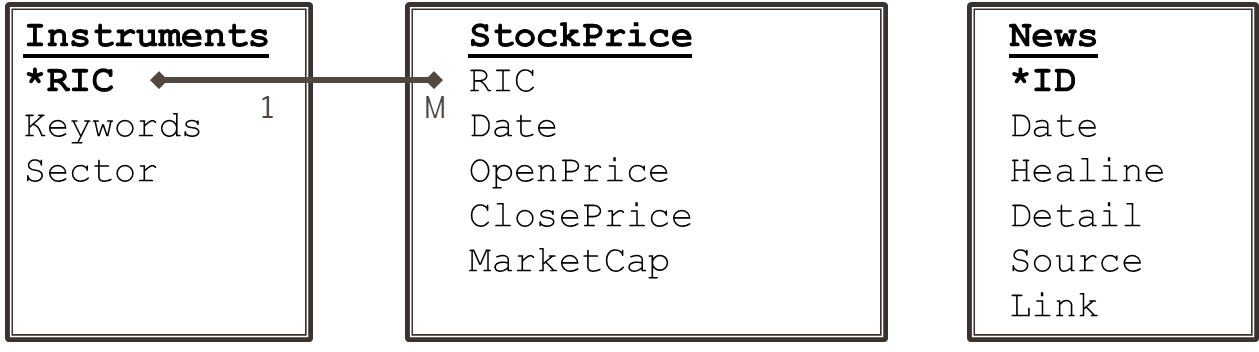


Figure .: Database Schema

* Instruments (Table 3.3, excel)

This file contains all in-scope stocks symbols and its industry sector. The primary key is RIC. Both Sector and RIC information are necessary to analyze model performance for each stock and overall sector.

* Stock Price (csv)

Open/Close Pricing and MarketCap data are stored for each stock in a time series manner. Difference between Open and Close Price is used to determine stock returns, whereas MarketCap is needed as a weight for calculating sector index.

* News (csv)

This file contains news id, published date, headline and its detail body content. Further stage of the data processing will use this text information to derive news sentiment scores (a numeric variable to quantify its polarity).

## Data Preprocessing

### Handling Missing Data and Errors

#### Missing Numerical Data and Errors

Average number of KLSE trading days in a calendar year is around 240, that is 2,400 days for a period of ten (10) years. As seen in Table 3.4 below, the quantity of data for all stocks are very good (around 2,400). The only exception are ASTR.KL and OCKG.KL due to their IPO date on year 2012. Data quality is revalidated by visually inspecting all price charts similar to Figure 3.4, to ensure no noticeable gaps on pricing and market capital.

Table .: Stock Price Data Count

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RIC** | **min\_date** | **max\_date** | **OPENPRICE** | **CLOSEPRICE** | **MARKETCAP** |
| AMMB.KL | 2010-01-04 | 2019-12-31 | 2451 | 2451 | 2451 |
| **ASTR.KL** | **2012-10-19** | **2019-12-31** | **1763** | **1763** | **1763** |
| AXIA.KL | 2010-01-04 | 2019-12-31 | 2452 | 2452 | 2452 |
| BIMB.KL | 2010-01-04 | 2019-12-31 | 2445 | 2445 | 2445 |
| CIMB.KL | 2010-01-04 | 2019-12-31 | 2450 | 2450 | 2450 |
| DSOM.KL | 2010-01-04 | 2019-12-31 | 2451 | 2451 | 2451 |
| GRNP.KL | 2010-01-04 | 2019-12-31 | 2449 | 2449 | 2449 |
| HLBB.KL | 2010-01-04 | 2019-12-31 | 2452 | 2452 | 2452 |
| HLCB.KL | 2010-01-04 | 2019-12-31 | 2451 | 2451 | 2451 |
| LOND.KL | 2010-01-04 | 2019-12-31 | 2442 | 2442 | 2442 |
| MBBM.KL | 2010-01-04 | 2019-12-31 | 2452 | 2452 | 2452 |
| MBSS.KL | 2010-01-04 | 2019-12-31 | 2450 | 2450 | 2450 |
| MDCH.KL | 2010-01-04 | 2019-12-31 | 2437 | 2437 | 2437 |
| MXSC.KL | 2010-01-04 | 2019-12-31 | 2452 | 2452 | 2452 |
| **OCKG.KL** | **2012-07-17** | **2019-12-31** | **1822** | **1822** | **1822** |
| PUBM.KL | 2010-01-04 | 2019-12-31 | 2453 | 2453 | 2453 |
| RHBC.KL | 2010-01-04 | 2019-12-31 | 2430 | 2430 | 2430 |
| STAR.KL | 2010-01-04 | 2019-12-31 | 2440 | 2440 | 2440 |
| TCOM.KL | 2010-01-04 | 2019-12-31 | 2446 | 2446 | 2446 |
| TLMM.KL | 2010-01-04 | 2019-12-31 | 2452 | 2452 | 2452 |

#### Missing News Data and Errors

Two areas of text data errors encountered during web page scrapping that were treatment:

1. Publish timestamp– Timestamp availability as when the article is published is inconsistent. This is seen mainly when the news is published before market open around 7am. Therefore, only the published date part is used throughout this analysis. When the article is current, the web page might just publish the time without date. Missing date are substituted with current date.
2. No headlines – Less than 0.02% of data has either missing news headline or detail, hence are omitted in this research.
3. Chinese News – About 2.2% of news are encoded in Chinese characters. They are found as translated copies of existing English article, hence were dropped in this analysis to avoid translation and skewing of result.

### Feature Engineering

#### Stock Return

This research is interested on share returns for current and next 5 days. Returns is measured as the proportion of Open-To-Close over the Open Price. The prediction label based on this measurement is different from (Li et al., 2014) as this project used fixed Open Price at current day rather than daily open. For example, Open Price for MBMM.KL on 11 Aug 2019 was 8.75. The same day Close Price is 8.83, which is 0.91% higher from 8.75. The hike of 0.91% is calculated by (8.83-8.75)/8.75 x100%. The concept is illustrated in Figure 3.7 below.

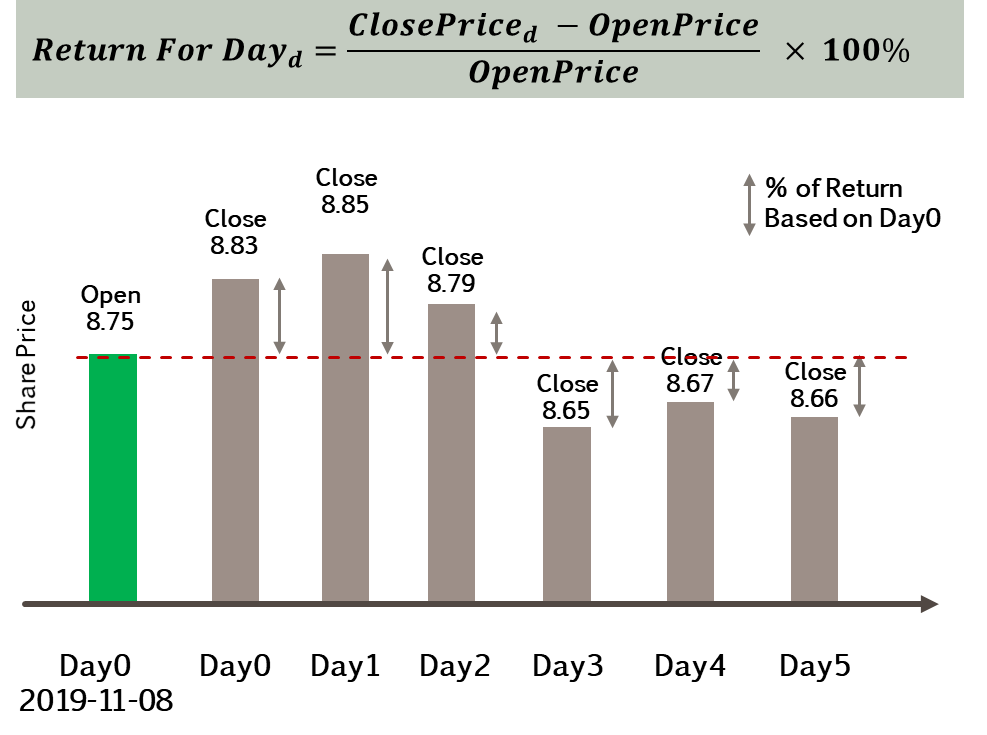


Figure .: MBBM.KL Stock Price Return Illustration

To compute the time series returns, intermediate share price DataFrame is created by shifting up a day for five (5) consecutive days in the future (see upper part of Table 3.5). Returns is then calculated by performing columnar arithmetic based on equation above (Figure 3.7) on the intermediate DataFrame. See the sample results for Maybank stock price below (lower part of Table 3.5). Note that shifting up of CLOSE PRICE results in blank values at the last few rows. To avoid this, the time series data till 2020-01-08 is obtained to ensure not returns till 2019-12-31.

Table .: MBBM.KL Stock Price Return Example

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CURRENT** | | **FUTURE** | | | | |
| **MBBM.KL** | **OPEN PRICE** | **CLOSE PRICE** | **CLOSE PRICE** | | | | |
|  | **DAY0** | **DAY0** | **DAY1** | **DAY2** | **DAY3** | **DAY4** | **DAY5** |
| 2019-11-08 | 8.75 | 8.83 | 8.85 | 8.79 | 8.65 | 8.67 | 8.66 |
| 2019-11-11 | 8.84 | 8.85 | 8.79 | 8.65 | 8.67 | 8.66 |  |
| 2019-11-12 | 8.84 | 8.79 | 8.65 | 8.67 | 8.66 |  |  |
| 2019-11-13 | 8.79 | 8.65 | 8.67 | 8.66 |  |  |  |
| 2019-11-14 | 8.66 | 8.67 | 8.66 |  |  |  |  |
| 2019-11-15 | 8.68 | 8.66 |  |  |  | Blank areas will be filled up with future data points |  |
|  |  |  |  |  |  |  |  |
| **MBBM.KL** |  | **CURRENT RETURN** | **FUTURE RETURN** | | | | |
|  |  | **DAY0** | **DAY1** | **DAY2** | **DAY3** | **DAY4** | **DAY5** |
| 2019-11-08 |  | 0.9143 | 1.1429 | 0.4571 | -1.1429 | -0.9143 | -1.0286 |
| 2019-11-11 |  | 0.1131 | -0.5656 | -0.0215 | -1.9231 | -2.0362 |  |
| 2019-11-12 |  | -0.5656 | -2.1493 | -0.0192 | -2.0362 |  |  |
| 2019-11-13 |  | -1.5927 | -1.3652 | -0.0148 |  |  |  |
| 2019-11-14 |  | 0.1155 | 0.0000 |  |  |  |  |
| 2019-11-15 |  | -0.2304 |  |  |  |  |  |

#### Sector Index Return

This project used capitalization-weighted method to calculate Sector Index. This is similar to method used for Kuala Lumpur Composite Index (KLCI), which is the main indicator for Malaysian economy and made up of thirty (30) largest stocks. Capitalization-weighted market indexing is aggregation weighted price for all component stock by is capital size, which is the stock price multiplied by number of outstanding shares.

In addition to individual stock price, returns for both Financial Services (FS) and Telecommunication & Media (TELCO) sectors are measured by their Index price. Each index has time series prices (OPEN and CLOSE), which are calculated by summing up the weighted prices from all underlying stocks as shown in Table 3.6 below. Weightage for each stock is calculated as proportion of their market capital value over the combined market capital of all stocks within the index (Table 3.7).

Table .: Sector Index Price Calculation

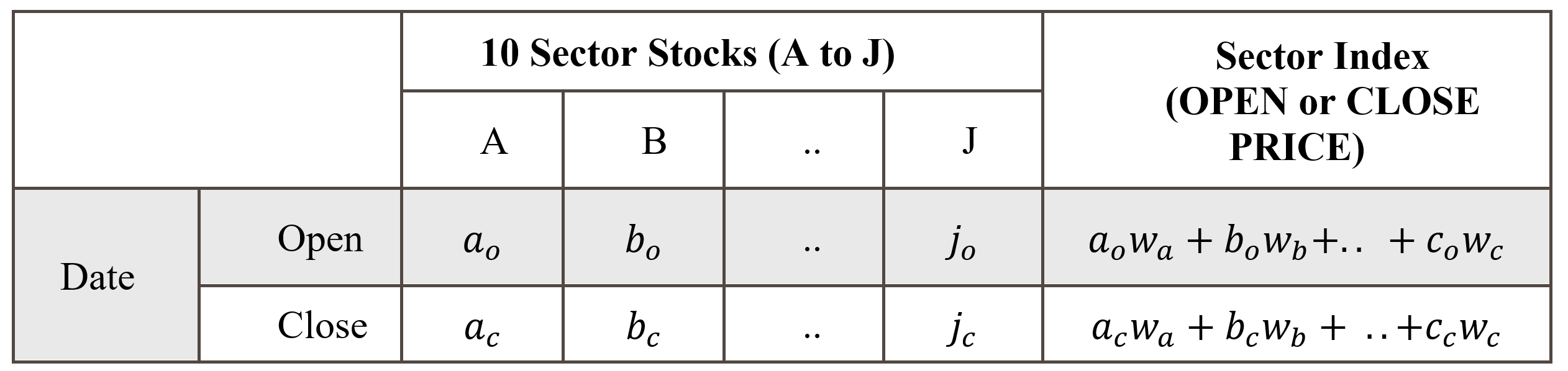
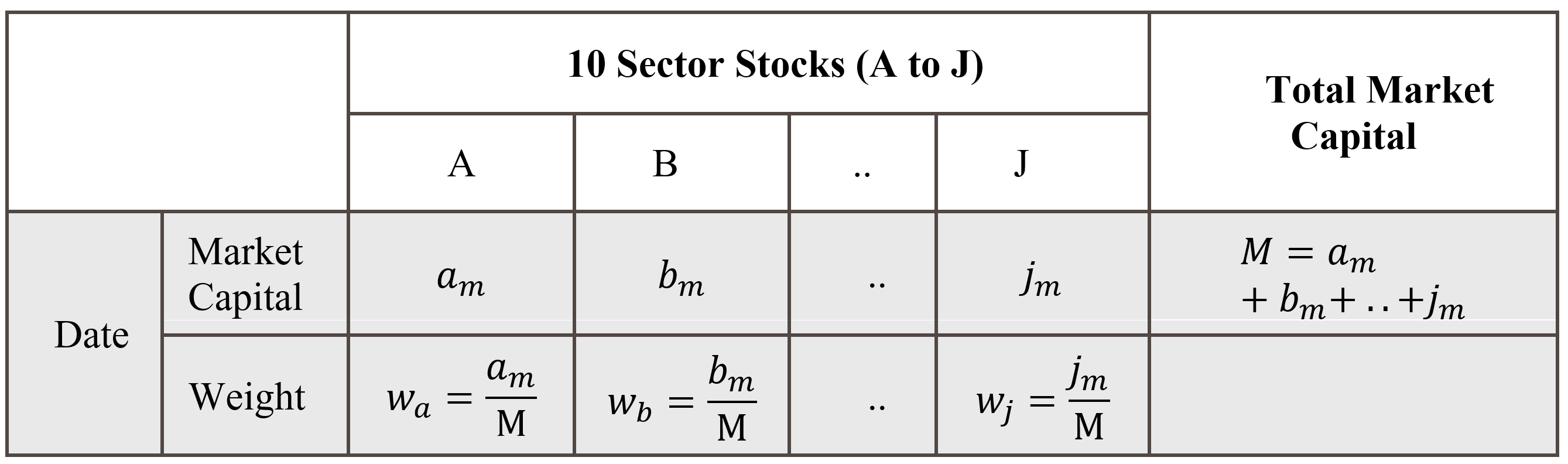


Table .: Stock Market Capital Weighing



As an example, market capital for all Financial Services (FS) stocks within scope are calculated as in Table 3.7 below. Maybank (MBBM.KL) weigh 0.2860 on 4th Jan 2010. When multiplied by its Open Price, it contributes to 1.9589 of Index Open Price. FS Index Open Price is the summation of all weighted open prices, which is 7.0256.

Table .: Market Capitalization Weighted Financial Services Index

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **2010-01-04** | | | |
| **Stocks** | **Stock**  **Market Capital** | **Stock**  **Weightage** | **Stock**  **Open Price** | **Stock**  **Weighted Open Price** |
| **MBBM.KL** | 48,554,963,380 | 0.2860 | 6.8500 | 1.9589 |
| **PUBM.KL** | 27,690,298,560 | 0.1631 | 10.7726 | 1.7569 |
| **CIMB.KL** | 44,895,201,928 | 0.2644 | 6.3658 | 1.6833 |
| **HLBB.KL** | 12,972,678,292 | 0.0764 | 7.5357 | 0.5758 |
| **RHBC.KL** | 8,730,826,345 | 0.0514 | 3.8676 | 0.1989 |
| **HLCB.KL** | 7,916,815,616 | 0.0466 | 7.3301 | 0.3418 |
| **AMMB.KL** | 15,161,350,550 | 0.0893 | 4.9900 | 0.4456 |
| **BIMB.KL** | 1,312,151,770 | 0.0077 | 1.1157 | 0.0086 |
| **LOND.KL** | 1,900,505,222 | 0.0112 | 4.7922 | 0.0536 |
| **MBSS.KL** | 651,801,094 | 0.0038 | 0.5800 | 0.0022 |
| **SUM** | **169,786,592,757** | **1.0000** |  | **7.0256** |
|  |  |  |  | **(Index Open Price)** |

Once daily sector Index Open and Close prices are derived from weighting method above, the same procedure defined in Section 3.5.2.1 above is applied to get Index return.

#### News Sentiment

This section describes the concept and technique used to transform news article into quantifiable sentiment polarity, which is a number between 0 and 1. The tools used are Senti-Wordnet dictionary, Python NLTK library for sentence tokenization and Part of Speech (POS) tagging. It is important to explain the fundamental concept of Wordnet and POS tagging prior to applying Senti-Wordnet.

1. Wordnet

WordNet is the most well-developed and widely used lexical database designed for efficient computer software programmatic operation. For each English word, it produces a set of synonyms (called Synset) based on its built-in dictionary and the POS tag of the word such as nouns, verbs, adjectives, and adverbs. Wordnet superficially resembles thesaurus in the aspect of providing synonyms, however, Wordnet measure how closely terms are related to each other. Measurement of distance and linkage with terms with similar concept are available for researcher.

From coverage point of view, WordNet’s goal differs from good standard dictionary in that WordNet uses word concept (here called word sense) which is a traditionally method for dictionaries writing. It connects its large lexical databased of English words by interlinking words of similar conceptual-semantic from the same POS. Various WN semantic relation and their support for POS as described in Table 3.9 below.

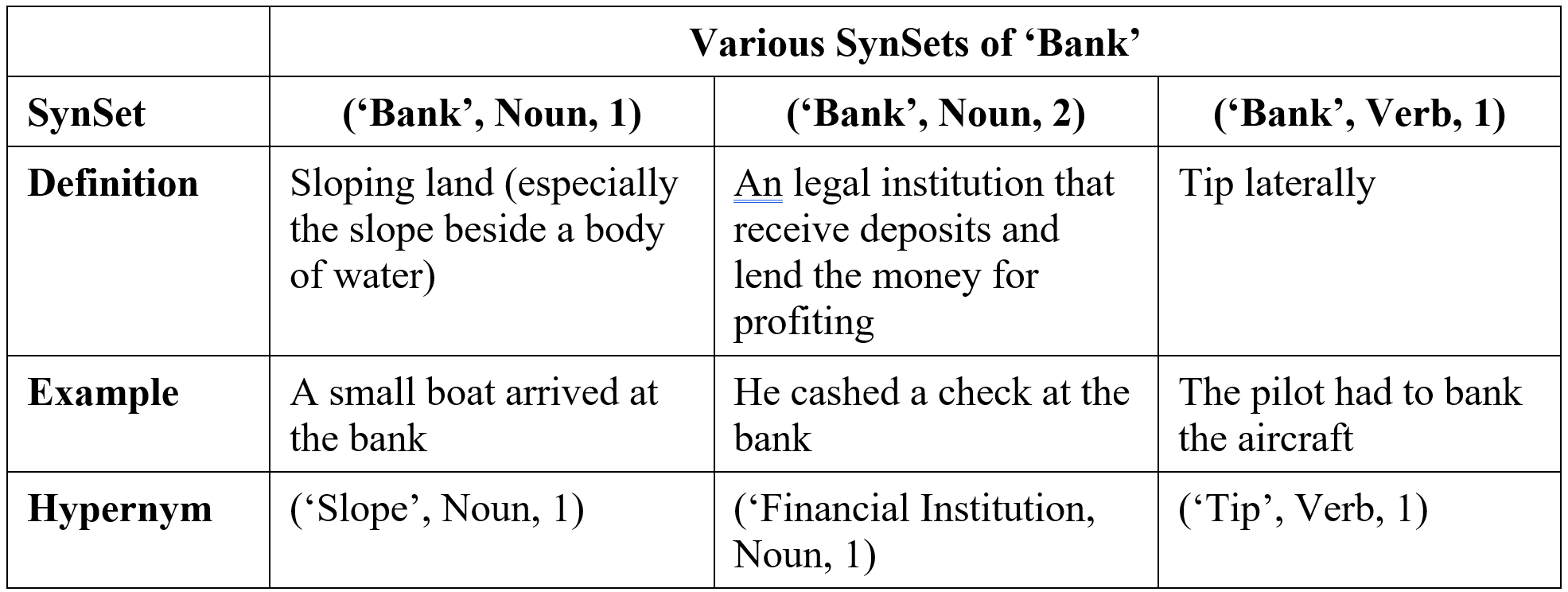
Table .: Structure of WordNet

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Semantic Relation (Definition)** | | **Part of Speech** | | | | **Example** |
| **N** | **V** | **Adj** | **Adv** |
| Synonym | A concept that means exactly or nearly the same as another | X | X | X | X | (sofa, couch, lounge) are all synonyms to each other |
| Antonym | A concept opposite in meaning to another | X | X | X | X | (love) 🡪(hate, detest)  (heavy) 🡪(light)  (quickly) 🡪(slowly) |
| Hypernym | * A concept whose meaning denotes a superordinate * From concept to superordinate | X | X |  |  | (dalmatian) 🡪 (dog)  (fly) 🡪 travel |
| Hyponym | * More specific meaning than a general or superordinate term applicable to meaning denotes a subordinate * From concept to subtypes | X | X |  |  | (house) 🡪 (beach house, guest house) |
| Meronym | From parts to whole | X |  |  |  | (course) 🡪 (meal)  (kitchen) 🡪 (home) |
| Holonym | From wholes to parts | X |  |  |  | (snow) 🡪 (flake, snowflake) |
| Entailments | An entailment is an implication. |  | X |  |  | (snore) 🡪 (sleep)  (looking) 🡪 (seeing) |

This research utilized WordNet 3.0 dictionary that contains 147,306 terms and 117,659 SynSets. Lexicographers in WordNet had organized the semantic relations with multiple rules such as symmetric, transitive and reflexive relations. The detail discussion on how the network graphs being constructed is beyond the discussion of this research. Based on the concept of semantic relation discussed so far, Part of Speech plays a role in determining the meaning of a word. In WordNet, a basic construct of such is called a SyncSet, which is a tuple of (Word, POS, Word Sense). Word Sense is a concept that differentiate the meaning of the word according to its context. As shown in example below (Table 3.10), the noun “Bank” has different meaning for Word Sense 1 and 2.

The understanding of SynSet is fundamental to sentiment analysis using Senti-Wordnet which is discussed in the following.

Table .: Various Meaning of The Same Word in WordNet



1. Senti-Wordnet

SentiWordNet (SWN) is a lexical resource for mining text data. It is a sentiment lexicon associating sentiment information to each WordNet Synset. In high level, SWN employs two learning algorithms (SVM and Rochhio) and four (4) training sets to yield eight (8) classifiers. As the result from the work in (Baccianella et al., 2010), Each SWN synset consist of three sentiment scores . Each scores (positive, negative and objective) are valued between 0 and 1. The relationship among the three sentiment scores is illustrated in Figure 3.8 below (Esuli & Sebastiani, 2006). The sum of the all three sentiment scores equals to one.

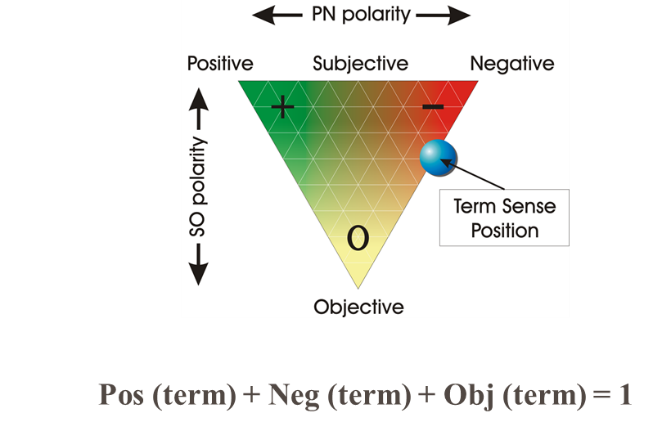


Figure .: Graphical Representation of Senti-WordNet Opinion Properties

Because SWN rides on SynSets from WordNet, it uses the same lexical resource which is essentially a synset-based sentiment measure with positive, negative and objective score. This means that a same term with different senses might have completely diﬀerent scoring. As shown in Table 3.11 below, different senses of word “Happy” results different sentiment associations. Ideally, SWN should be integrated with Word Sense Disambiguation algorithm for identifying words with closest context. However, WSD is beyond the focus of this research. This project always uses first sense as the most common scenario.

Table .: Senti-WordNet SynSet Scores Example

|  |  |  |  |
| --- | --- | --- | --- |
| **SynSet** | **(‘Happy, Adj, 1)** | **(‘Happy, Adj, 4)** | **(‘Bankrupt’, Noun, 1)** |
| **Pos(term)** | 0.875 | 0.125 | 0 |
| **Neg(term)** | 0 | 0 | 0.625 |
| **Obj(term)** | 0.125 | 0.875 | 0.375 |
| **Term Label** | Positive, because  Pos > Neg | Positive, because  Pos > Neg | Negative, because  Neg > Pos |

1. Sentiment Scoring

This research had considered both news header and detail for sentiment analysis. Sentiment are evaluated separately for header and detail, then later combined to get an average score that represent the overall sentiment of the article. This two steps evaluation is illustrated in Figure 3.9 below.

Firstly, SWN sentiment score is evaluated for each word using Equation 3.1. The word is labeled as POSITIVE if SWN Pos score is greater than Neg score. On the opposite, it will be labeled as NEGATIVE. Words without Neg and Pos scores (Obj=1) are considered OBJECTIVE. Similar labeling method was also found in work (Hor et al., 2018; Li et al., 2014; Nassirtoussi et al., 2015).

Secondly, sentiment score for the entire text (article header or detail) is calculated as number of positive term as numerator, and total positive and negative terms as denominator (see Equation 3.2). Similar denominator was mentioned in work (Zhang & Skiena, 2010). The value one (1) indicates extreme POSITIVE sentiment, whereas 0 represents extreme NEGATIVE sentiment. Positive or negative terms may not present in short text (especially short headers), hence Equation 3.2 encoded the sentiment score to 0.5 as a neutral point.

As each article comprises of news Header and Detail, average sentiment score is calculated as the final sentiment score for the article. Figure 3.9 below gives a simplified view of the above-mentioned workflow.

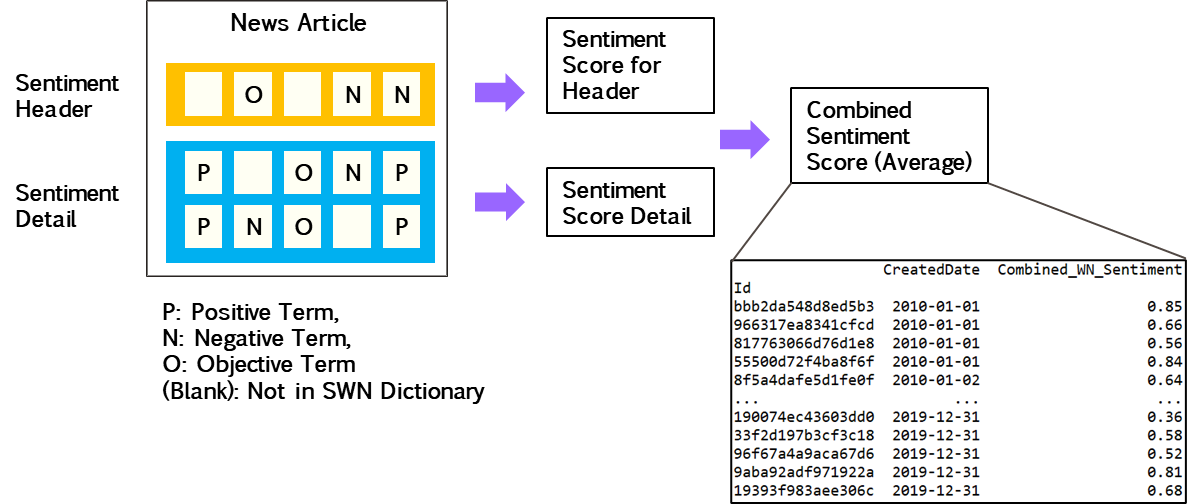


Figure .: Sentiment Scoring For News Article and Term Labeling

Equation .: Term Sentiment Label

Equation .: News Header/Detail Sentiment Score

Two sample news articles and their sentiment score calculations are illustrated below.

Table .: News Sentiment Score Example 1

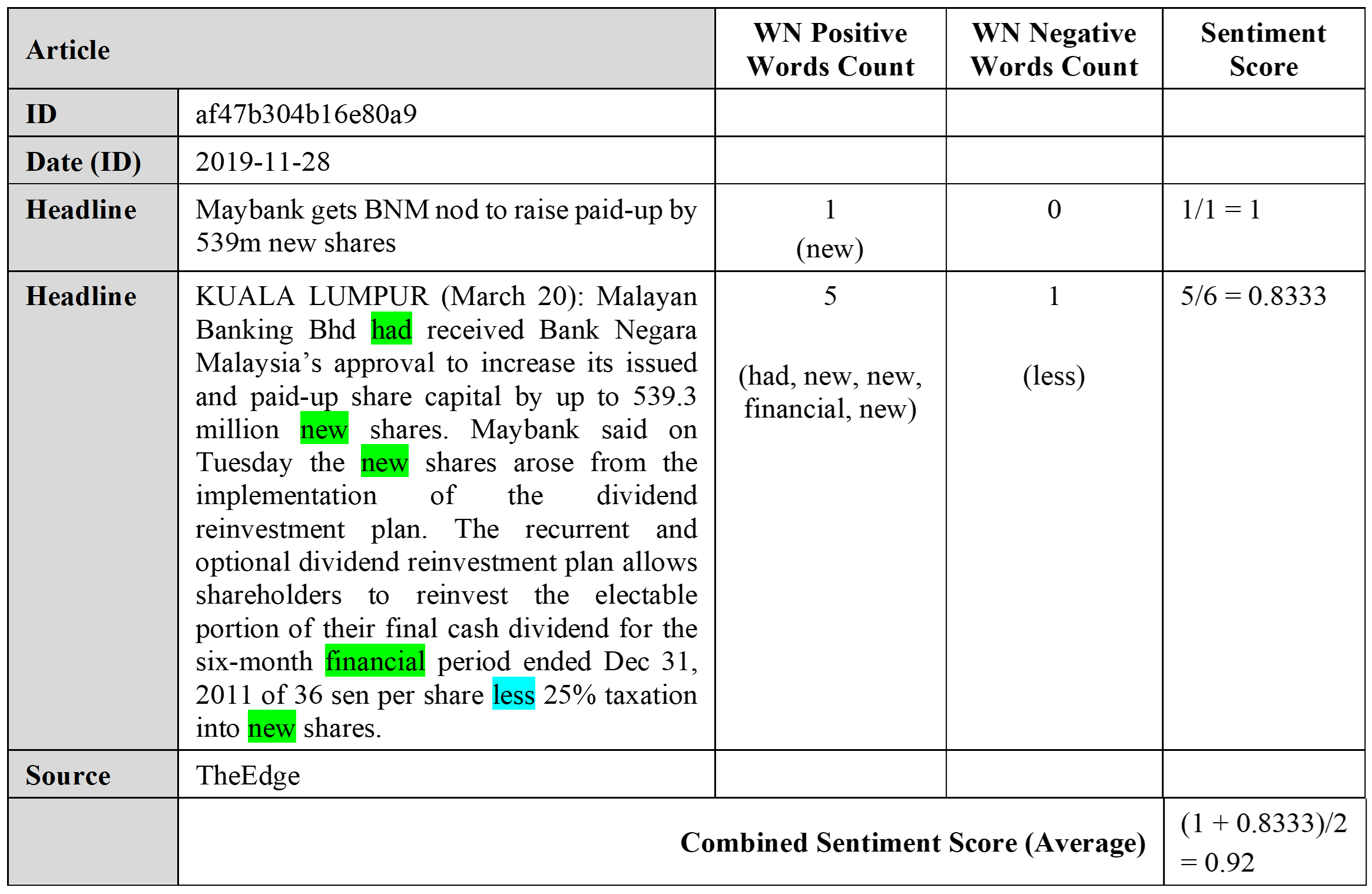


Table .: News Sentiment Score Example 2



1. Text Processing Implementation

Every news article stored in News database (refer system design in Section 3.3 above) is processed by computer program row-wise to determine its sentiment score. The program executes various calculation following equations in previous section. NLTK is chosen as the python library to implement NLP operations as it comes with convenient toolset such as tokenizers, POS taggers and SWN integration. The program first replace null values with empty string to avoid Python errors during NLTK operations. There are about two (2) percent of news encoded in Chinese character which is not supported by SWN. As it is a small amount, Chinese news are discarded. The complete workflow is depicted in Figure 3.10.

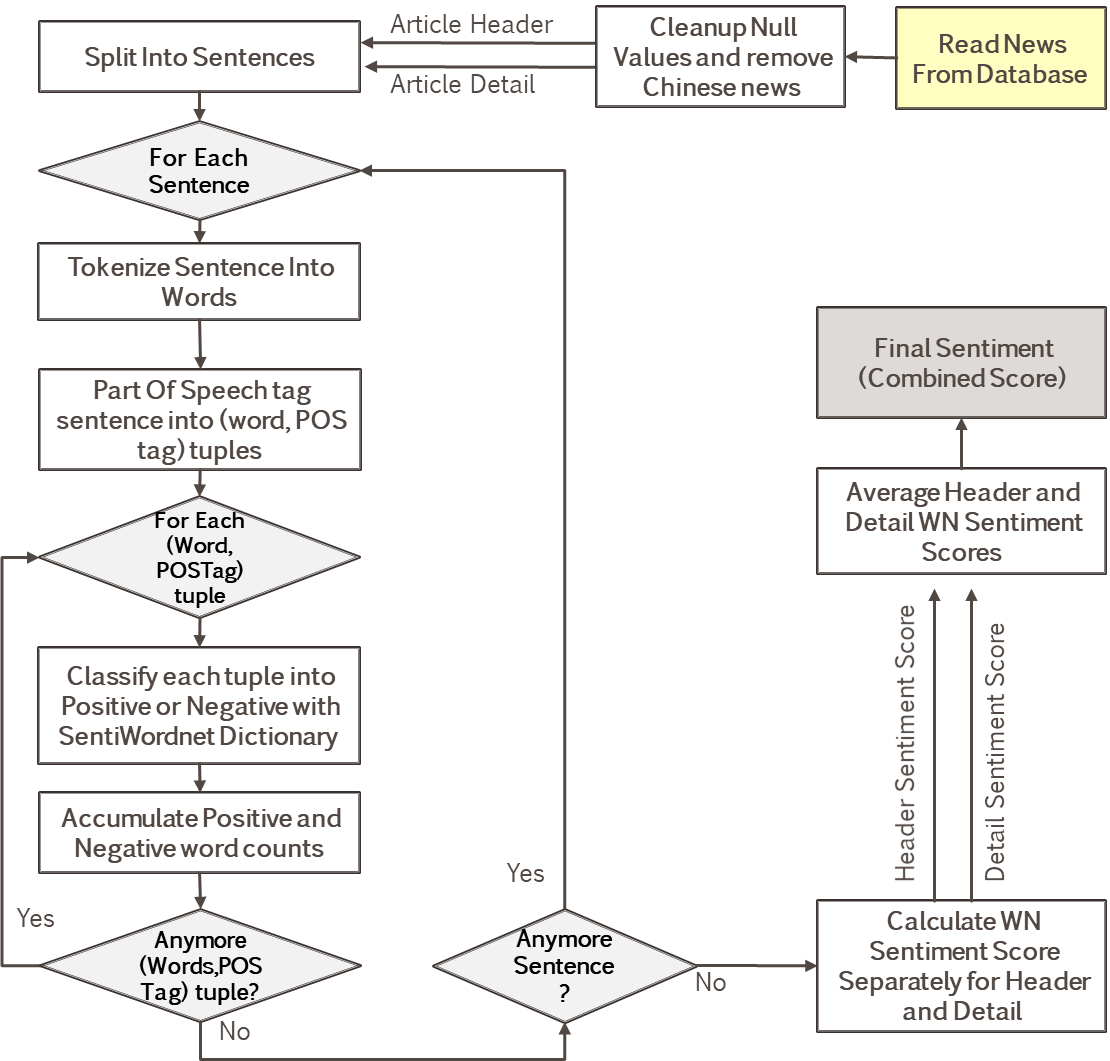
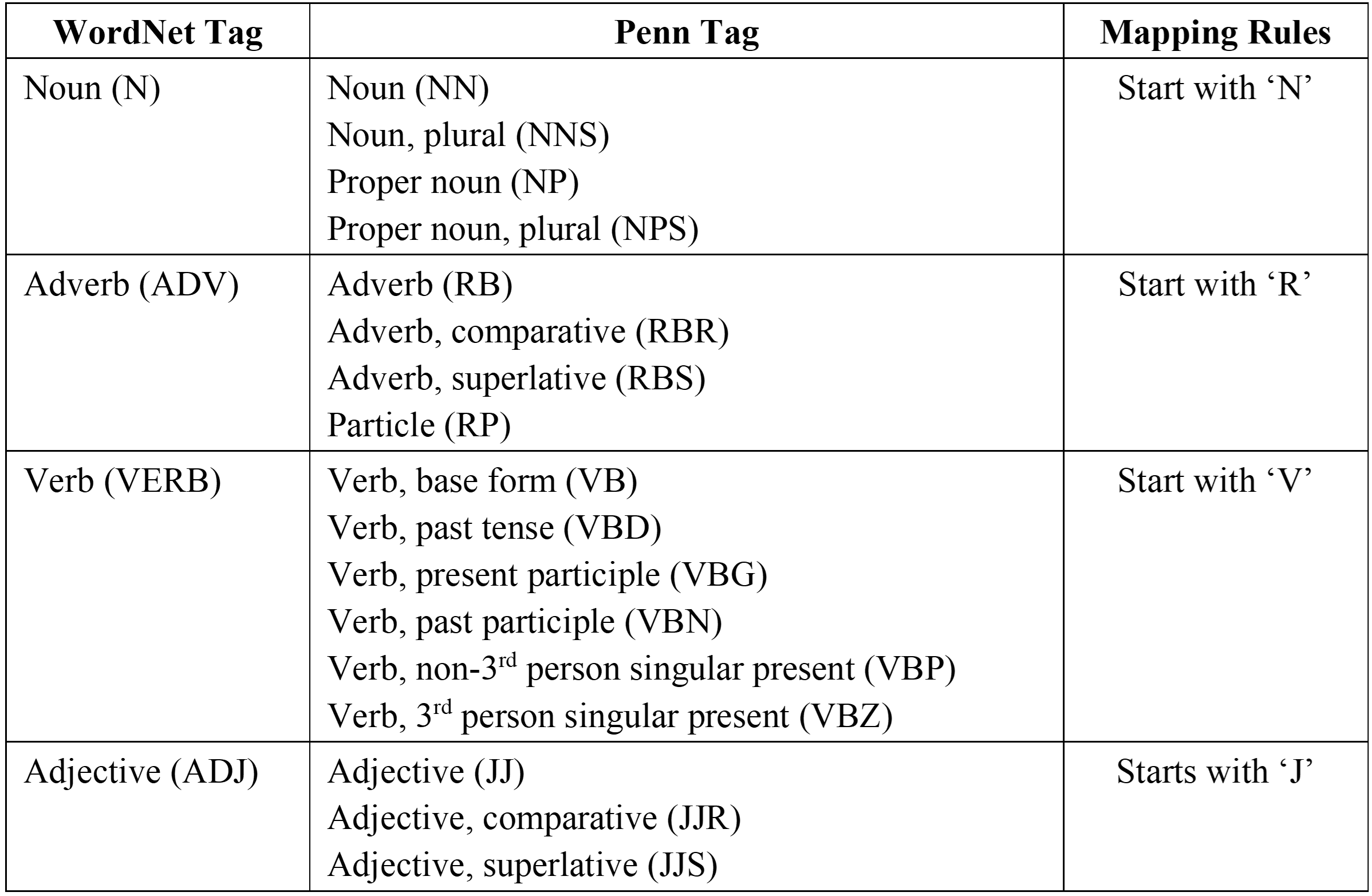


Figure .: Sentiment Score Processing Flowchart

As oppose to Machine Learning based NLP implementation, dictionary-based model used in this research does not require text stemming. This is because terms that does not present in WN dictionary were automatically filtered. Any stemming prior to WN evaluation will result in invalid words (unfound in WN dictionary).

As explained earlier, SWN requires that terms being POS-tagged. NLTK conveniently came with pretrained POS tagger and Sentence Tokenizer. NLTK tokenizer split text into sentences so that words within sentences can be POS-tagged. NLTK included POS tagger implements tagsets from Penn Treebank Project (Santorini, 1990), which consists of thirty-six (36) POS tags. Since WN supports only four (4) POS tags, mapping specified in Table 3.14 below is necessary. Simple mapping rules that matches first character of Penn tag is sufficient in this process.

Table .: Penn Treebank POS Tag Mapping to WordNet



The output from this stage is an intermediate news DataFrame consisting key columns of Id, Created Date and Combined WordNet sentiment score. Other columns are kept here only for validation purpose. At the aggregation phase (discussed later), sentiment scores are averaged out daily as the final sentiment score of that day.

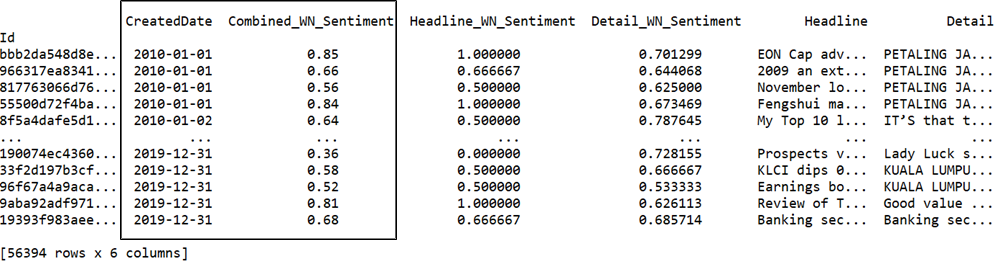


Figure .: Example Sentiment Scores DataFrame

### News To Stock Entity Mapping

This research studies news impact to Financial Service and Telecommunications & Media sectors. Sentiment scores produced from step above must be tagged individual stocks prior to further analysis. Mapping relevant news to stocks uses keyword search in news header and detail. These are the same keywords used for scrapping relevant news from provider’s website (refer to Section 3.2.1).

News to entity mapping is represented in a DataFrame with RICs in columns and news id as row index. The cells with Boolean True indicates relevance of news id to stock. It is possible that news articles contain keywords from multiple stocks. In such case, multiple True values in a single row will be observed. Figure 3.12 below shows a snapshot of news relation to Financial Services sector stocks. The last news id (bottom row) is related to both PUBM.KL and HLBB.KL.

Sector news relevance is represented with two new columns, namely ‘FSI\_INDEX’ and ‘TELCO\_INDEX’. The news is considered relevant to sector news if it is related to any of the index constituents (underlying stocks of the index). In computer program, this is simply derived from logical operator OR applied column wise across all its constituents. The final DataFrame at this stage has eleven (11) columns for each sector, ten (10) component stocks and one sector Index. It is structured this way for the convenience of filtering and aggregation in subsequent phase.

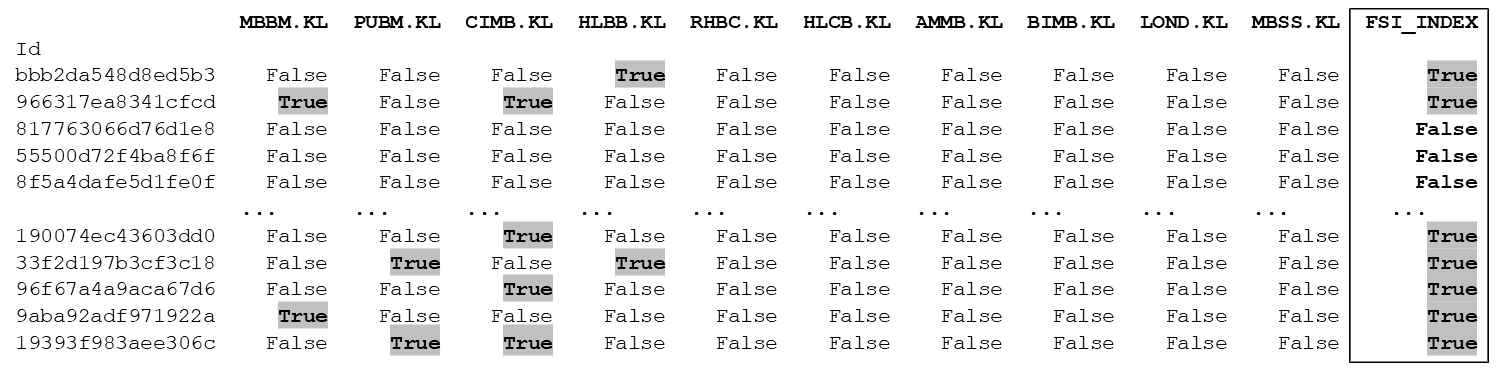


Figure .: Example News To Stock Entity Mapping DataFrame

### Data Aggregation

This stage simply joins data from News Entity DataFrame (output from Section 3.5.2.3) and News Sentiment DataFrame (output from Section 3.5.3) using news Id as the common key. The join produced a neat DataFrame that maps all news to its entities (including Sector Index) together with news sentiment scores. Shown in Figure 3.13 below is the snippet of this intermediate output filtered on Financial Services stocks.

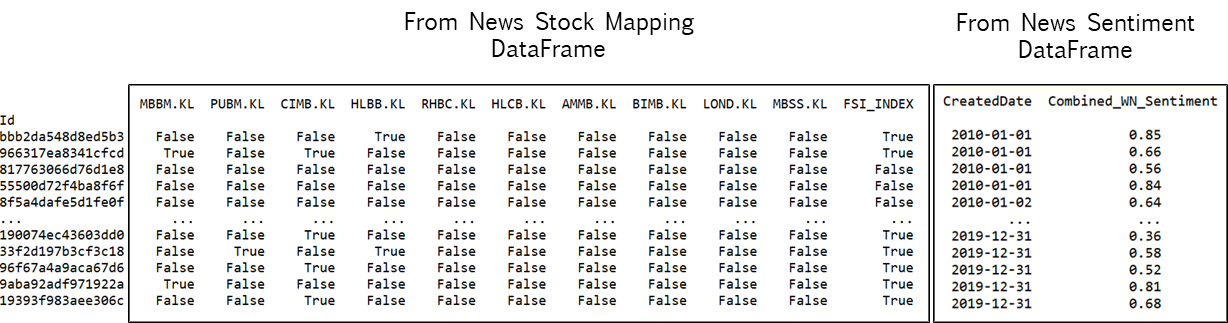


Figure .: Example News Sentiment To Entity Mapping

Stock price analysis in this research focus on daily returns and modeled with news sentiment as predictor. Hence, news sentiment score must be aligned at daily basis to match stock returns. The alignment procedure involves looping through every stock to filter its related news sentiment scores and average them on daily basis. The average daily sentiment scores for each stock are concatenated back into single DataFrame show in example snapshot below (Figure 3.14). Value NaN (null) represents unavailability of news on that day. At this point, news sentiment features are fully ready for modeling.

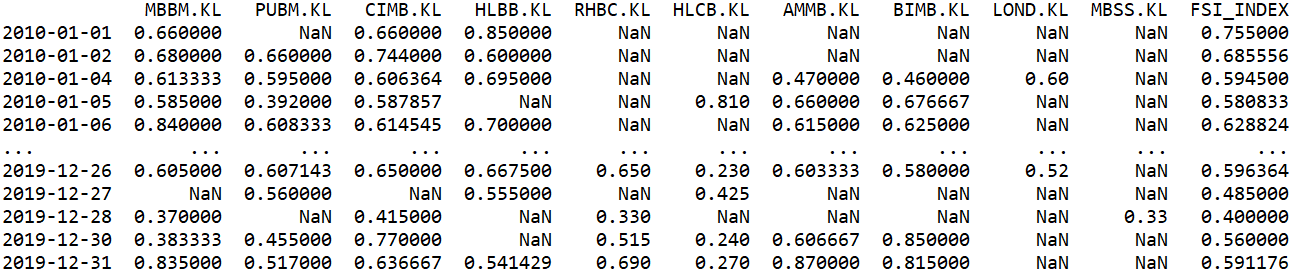


Figure .: Example News Sentiment Scores

### Returns Labelling

This research differs from most work found in literature review by its categorization of price movement. While most researches worked on just UP and DOWN labels, this research included ‘STAY’ as the third label. The idea of these three categories is to mimic human decision as a reaction to external influence such as news sentiment.

Other than news sentiment, another factor which may influence investor decision to trade is the cost incurred in buying and selling shares. There are three (3) types of trading cost involved for every transaction in KLSE. These costs can be summarized as below:

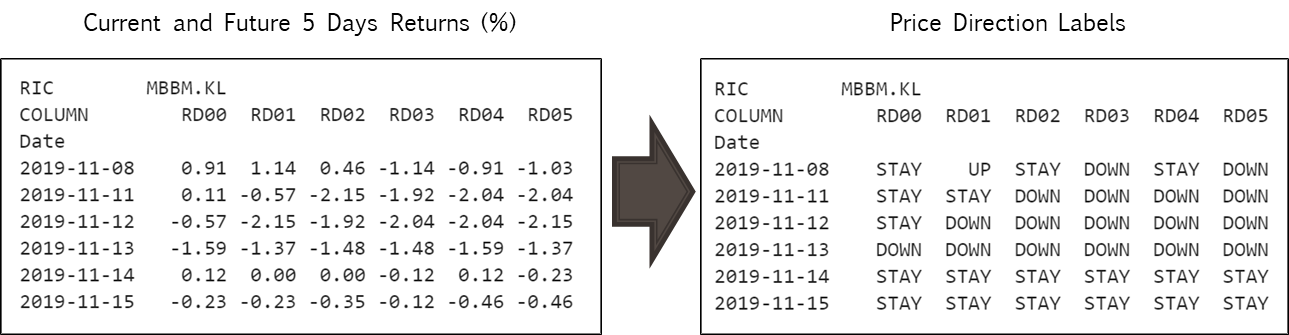
* Broker’s Commission - charges depending on brokers, generally 0.1% for cash and 0.3% for contra account, applicable to buyer and seller.
* Stamp Duty - charged at RM1 for every Rm1000 transacted.
* Clearing Fees - charged at 0.03% and capped at RM1000 per transaction applicable to seller and buyer.

This research considered a cost threshold of 0.5%, covering the cost for both way trades (buy and sell). This is based on thinking process that an investor will buy only if he believes the stock price will gain beyond this threshold (UP). On the other hand, sell the stock if he thinks its price will drop beyond the threshold (DOWN). In other words, the investor will not trade if he predicts the price movement is within the threshold (STAY). Similar three categories of price direction is found in work (Li et al., 2014). The author explored various threshold levels instead of fixed 0.5% in this study. The threshold used in this research can be justified with analysis in Section 4.2.1, where most the average returns fell within this boundary. This also means this research is investigating on abnormal returns through UP and DOWN events.

Determination of labels with 0.5% threshold is implemented using Equation 3.3 below. Figure 3.15 illustrates transformation of stock returns into price direction labels. At this stage, targets have been defined and ready for model building.

Equation .: Share Returns Labeling

Figure .: Example Share Returns Label



# Data Interpetation

## News

### Overall News

Total amount of 147,930 news was scrapped from provider websites from year 2010 to 2019. Out of this number, thirty-eight (38%) are related to stocks under this research scope. This equals to 56,394 news, spreading over 10 years with data points ranging from 4,300 to 7,800 articles yearly. TheEdge contributed to majority of the news data (66.1%). This is expected as TheEdge is a major business news provider in Malaysia

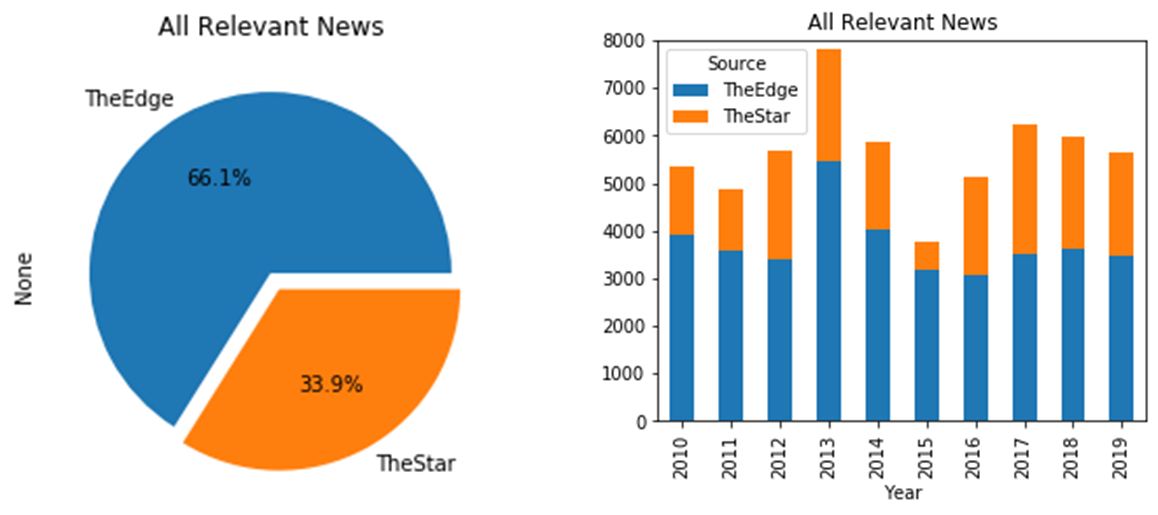


Figure .: News Count By Sources

In generally, large capital stocks attracted more interest to the news publisher. Correlation between News Count and company Market Capital is illustrated in scatter plot below (Figure 4.2). However, there were two outliers. Maybank (MBBM.KL) being the largest Malaysian bank (over RM80b) has only 5,375 news, whereas CIMB.KL as the fourth largest stock within the scope (over Rm36b) attracted the most interest from publisher (22,715 news).

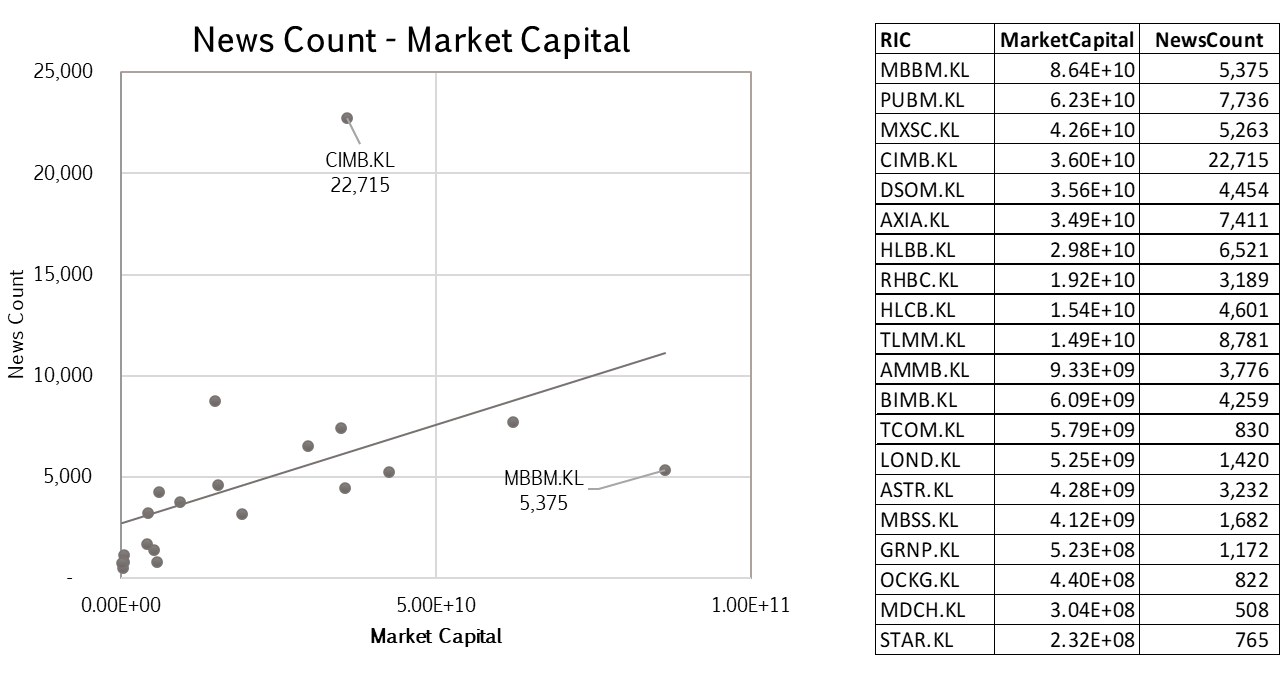


Figure .: News and Market Capital Scatter Plot

### Sector News

Financial Services (FS) sector news are more active as compared to Telecommunication & Media sector (TELCO). It is observed that FS stocks has 1.8x times more news than TELCO, in other words 64% of all news were related to FS sector. From the perspective of market capital, total FS market capital about RM335b, compared to RM142b in TELCO. This means total market capital of FS stocks is 70% of FS and TELCO combined. Financial services is a dominant sector in KLSE, hence undoubtedly it has the most new coverage.

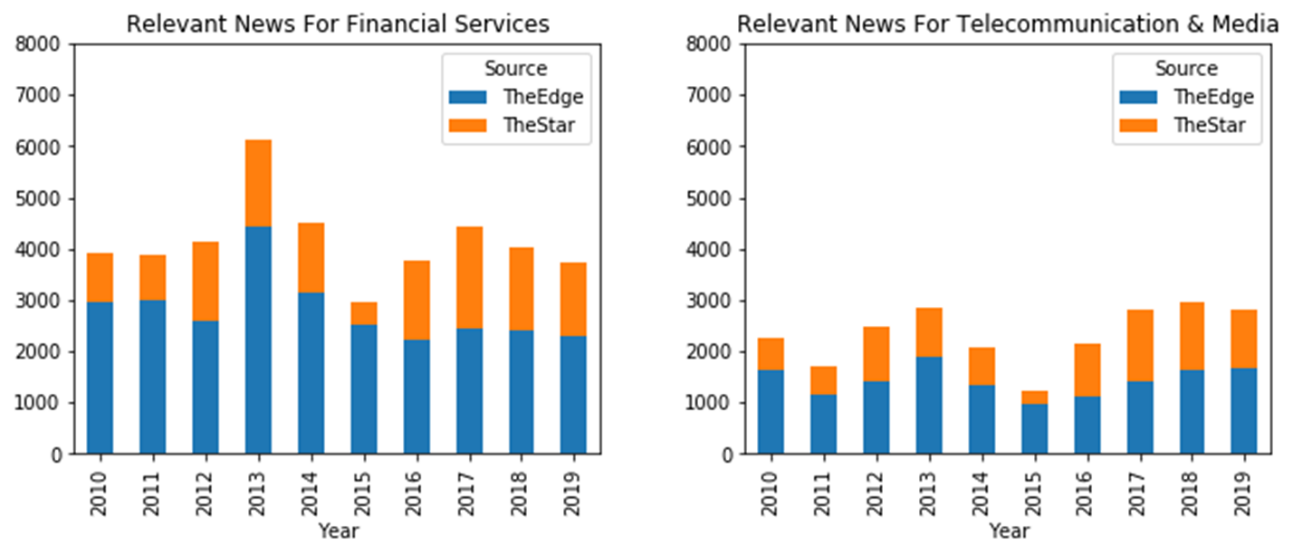


Figure .: Sector News Count

## Share Price

This section analyzes share returns (Section 3.5.2.1) and its directional labels produced in Section 3.5.5. Returns are projection of gaps between Open and Close price for current day and near term (subsequent five days).

### Share Returns

Share returns for all stocks within each sector are averaged to get an overall view of short-term trends. As seen in Figure 4.4 below, stocks in Financial Services (FS) sector are generally trending up within range of 0.40%. Most stocks have very little variability within the current day (RD0), but gradually trend up in the next five (5) days (RD01 to RD05). Average returns of FS stocks trend from 0.02 in RD00 to 0.13 (RD05). The boxplot can be interpreted together with individual stock breakdown in Figure 4.5. Only two stocks (MBBS.KL and CIMB.KL) had dipped into negative returns in average.

Compared to FS, TECO stocks are more volatile in short terms. This can be visualized clearly in the heteroscedastic plots of individual TELCO stocks performance (Figure 4.5). Trend on underlying stocks varies highly within -0.4 to +0.8. Three stocks (STAR.KL, ASTR.KL, MDCH.KL) has negative trending and two stocks (TCOM.KL, OCKG.KL) has steepest gain in short term. However, when combined all TELCO stocks, its average return trends up from 0.06 (RD00) to 0.15 (RD05).

In general, both FS and TELCO sectors have positive trend. For comparison purpose, below FS and TELCO charts are arranged side by side with the same scale.

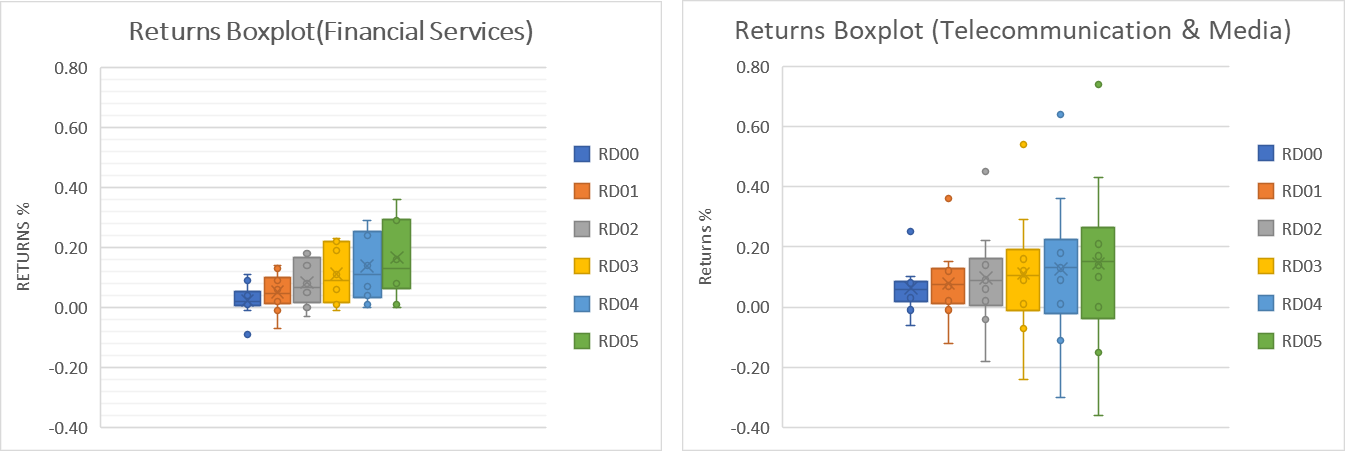


Figure .: Sector Price Returns Overall Chart

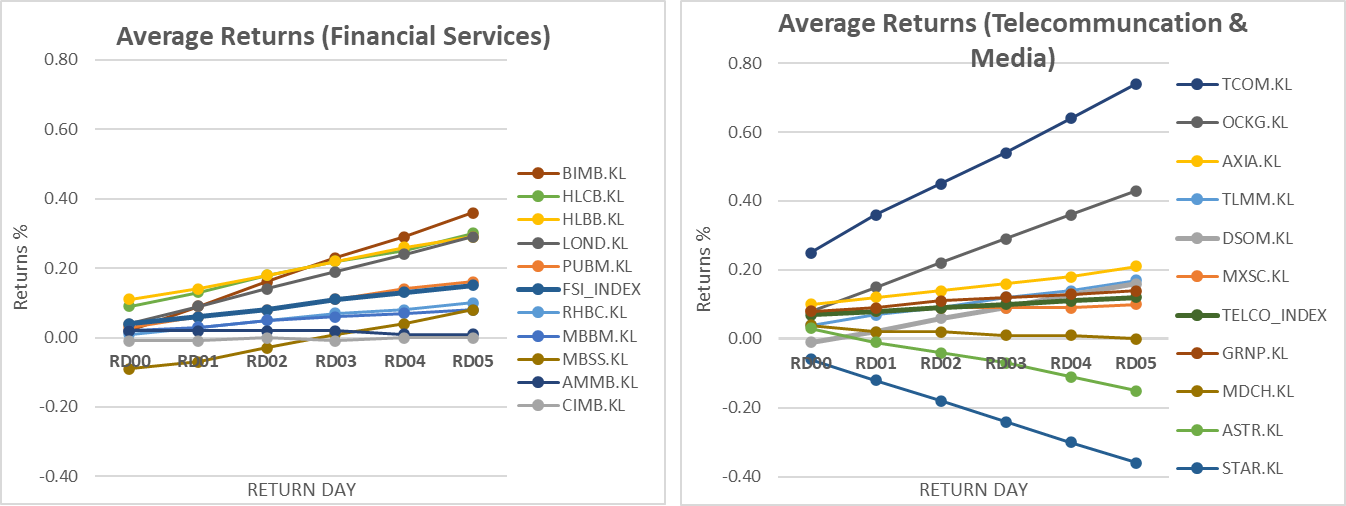


Figure .: Shares Price Returns Chart By Stock

### Returns Label

Share returns labels were analyzed in the next chart below (Figure 4.6). Percentage of the labels distribution (UP/STAY/DOWN) is visualized across six days (RD00 to RD05).

Both FS and TELCO sectors shows similar pattern where majority of share price movement are within STAY boundary on current day (RD0). Share returns moved away from STAY gradually after few days in both UP and DOWN directions. Although analysis from previous section suggested up trends on returns, the magnitude of hike is too subtle to push up the number of UP labels (recalled that the STAY boundary is 0.5%). This is shown in Figure 4.6 where the shift is well balanced over time in both UP and DOWN direction.

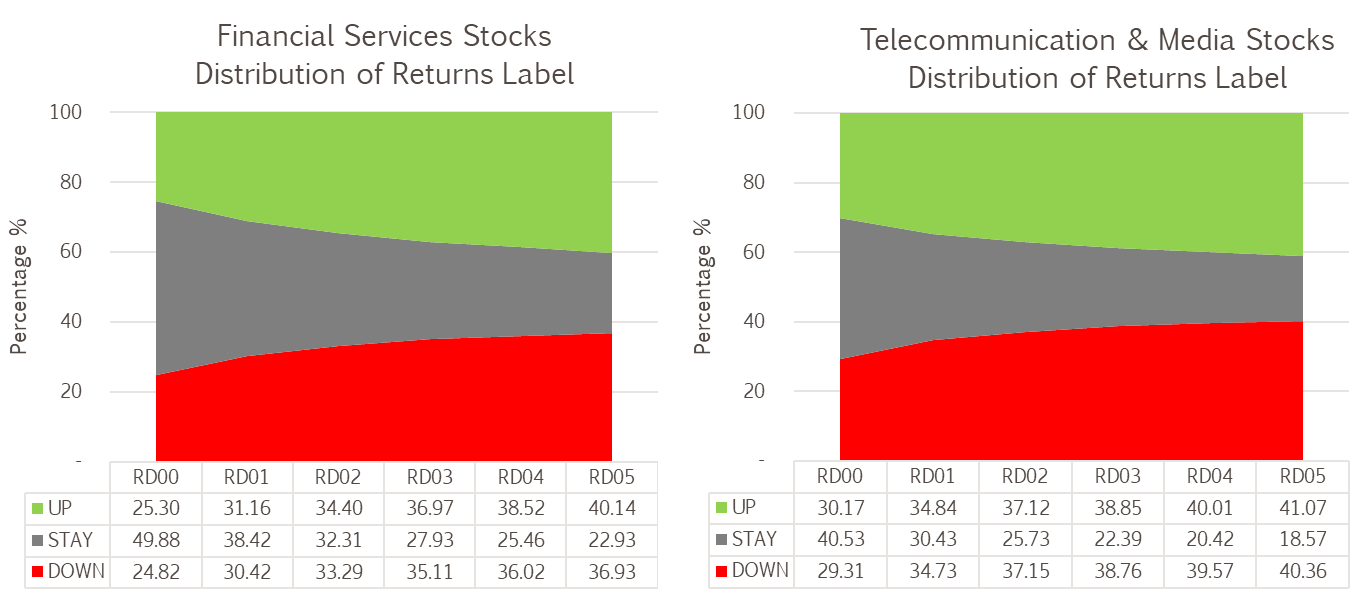


Figure .: Returns Label Distribution For Stocks

Label distribution on both sector index shows similar pattern (Figure 4.7 below). However, FS Index has a “fatter” STAY labels compared to TELCO. This is consistent to the understanding that FS stocks are less volatile compared to TELCO.

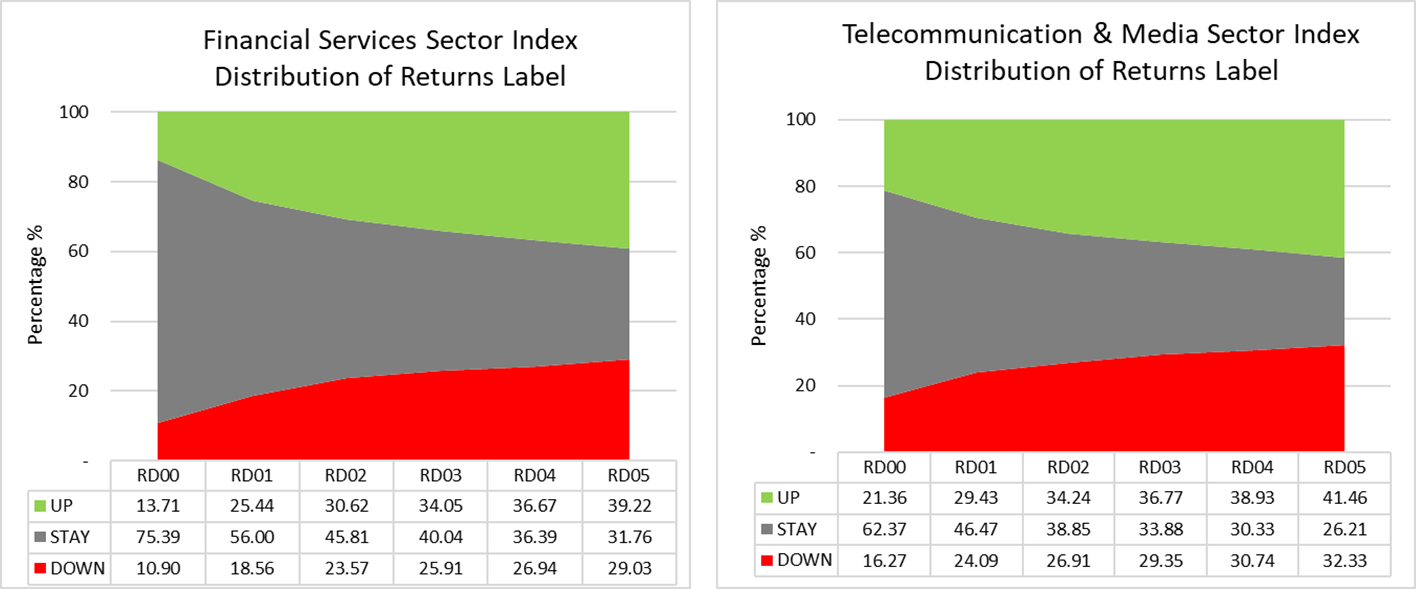


Figure .: Returns Label Distribution For Sector Index

# Data Modeling

## The Model

Modeling of sentiment scores to predict returns involved multiple steps depicted in Figure 5.1 below. Sentiment scores are mapped into DOWN, STAY and UP labels by segmenting the scores into boundaries set by thresholds. The centre and threshold (here called parameters) dictates the upper and lower boundary. Scores below lower boundaries are mapped to DOWN, above upper boundary are mapped to UP, whereas scores within the boundaries are mapped to STAY. A grid search is performed on the centre and threshold to discover the optimum parameters for different objectives.

This result from the mapping is used for prediction. The model is evaluated against actual labels for current and next five (5) days – RD00 to RD05. This evaluation used three (3) categories confusion matrix for accuracy, precision, recall and F1 means measurement (see Section 5.2 later).

Results from the evaluation are averaged to have overall representation of the model performance. The averaged result is compared to averaged baseline derived from labels using approach discussed in the following section. This evaluation process is repeated for various group of stocks and sector index within the scope of this research.

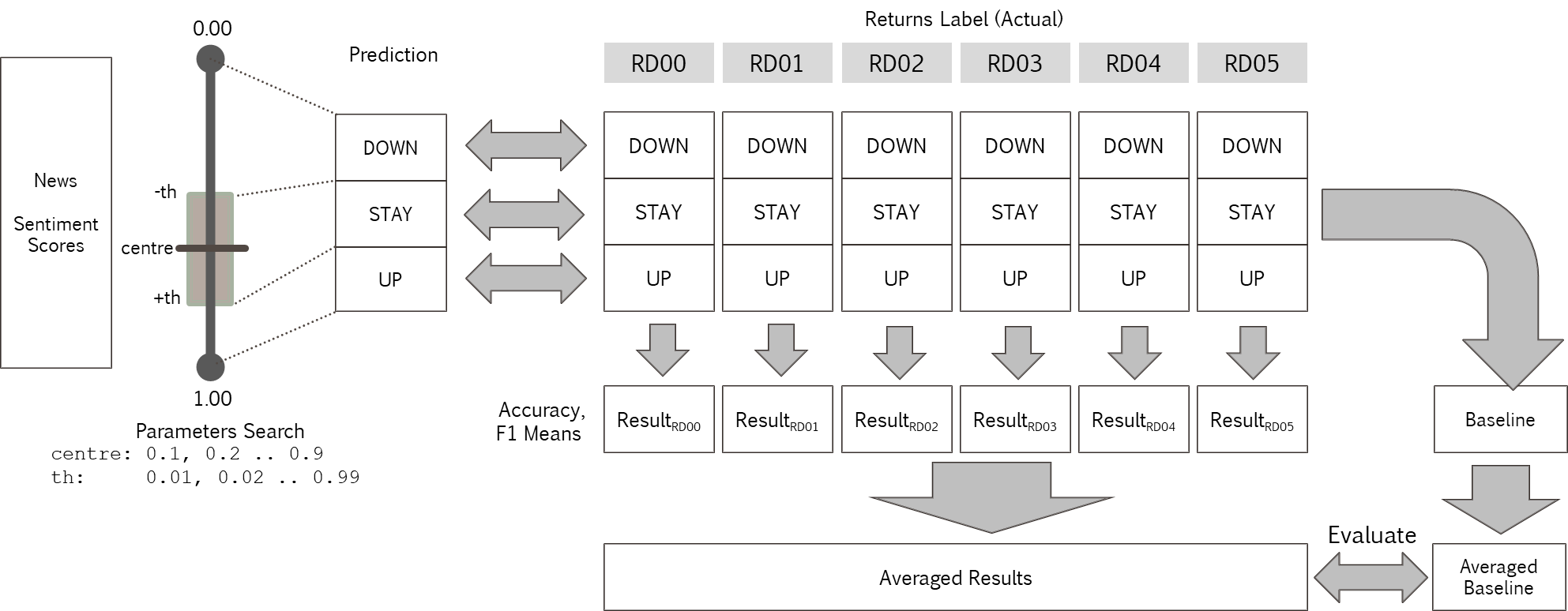


Figure .: Prediction Model

## Performance Metrics

Performance of three (3) categories classification is derived from confusion table, which is a 3x3 matrix of actual and predicted count of each label. Figure 5.2 shows two ways a confusion matrix can be interpreted to evaluate precision and recall. Few terminologies are required to further understand the evaluation:

* True Positive (TP) – It is true and correctly predicted
* False Negative (FN) – It is true, but was not predicted
* False Positive (FP) – It is not true, but incorrectly predicted.

Precision can be calculated by looking at the confusion matrix in a columnar manner (illustrated in upper part of Figure 5.2), in which quantity of TP and FP were placed. Precision for a specific label is evaluated as TP over total prediction (TP+FP). It gives an idea how precise labels were predicted. In other words, how many predictions were correct.

When the confusion matrix is read row-wise, recall will then be a meaningful measure. Recall is calculated as TP over total labels (TP+FP). It tells how many observations for a category were correctly predicted.

Accuracy is the only metric discussed here that evaluated from all cells in the confusion matrix. It tells overall how many observations were correctly predicted. Accuracy may not be the best or only performance metric when there are imbalance labels. High accuracy attributed by overwhelmed TP in one category may looks promising on surface but may not be significant. For example, a high accuracy due to overwhelming STAY labels does not motivate investor to trade. For this reason, precision and recall measurement on each label will be a better objective. Instead of reporting both recall and precision, this research optimizes on F1 means, which is a harmonic means between recall and precision. Equations for precision, recall, accuracy and F1 means are included in Figure 5.2 below. Accuracy and F1 means are the final metrics reported in this research.

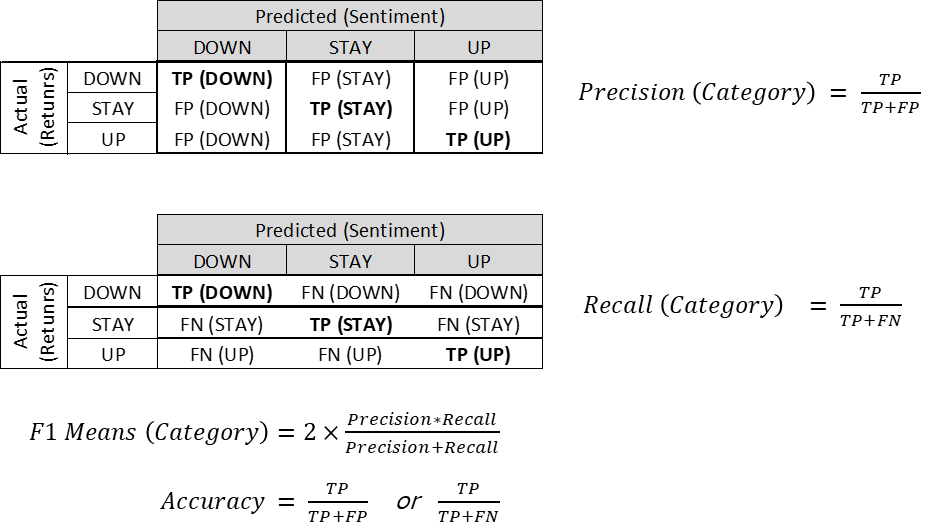


Figure .: Three Categories Confusion Matrix

## Performance Baseline

There are two possible ways to baseline performance. One is random guessing, which places prediction equally across all three (3) labels. This simply means 33% as accuracy and F1 means baseline. A better baselining method uses known distribution of the labels. This approach has higher benchmark than random guessing, hence was adopted as the baseline. As there are multiple sets of labels (RD00 to RD05), average metrics are used for baseline comparison.

In baselining accuracy, the adopted approach simply predicts labels with the highest (max) probability for each set of labels (RD00 till RD05). From example shown in Figure 5.3 below, one can place ‘all-in’ with STAY when predicting next day return (RD01), and this would yielded 38.42% of accuracy. Max accuracies from each column (RD00 to RD05) are averaged to get the final accuracy baseline.

As for F1 means, the probability of each labels simply represents the F1 means value. It is from the same idea that if one placed ‘all-in’ to a specific label, precision and recall for that label are be equivalent to the probability of label. Since precision and recall are the same value, their harmonic means, F1 will also be the same value based on equation in Figure 5.2 earlier. The label probability for each category are averaged up as the F1 means baseline.

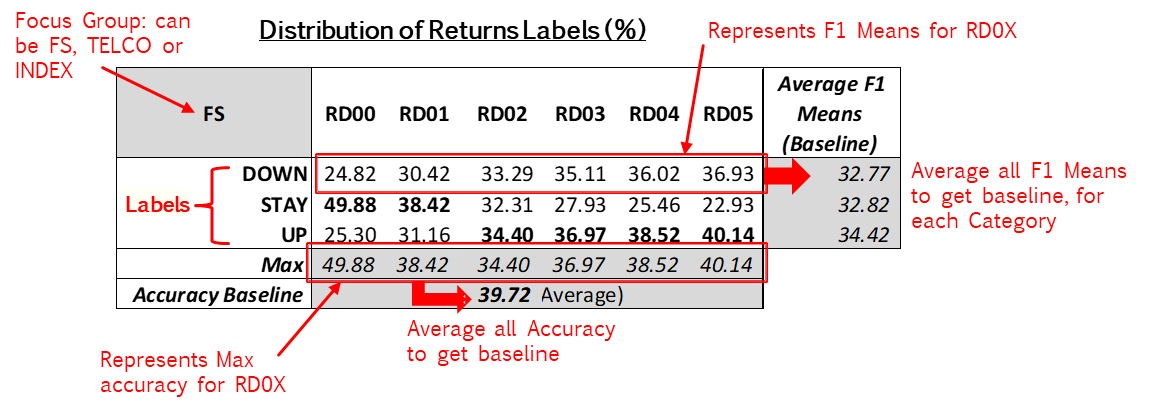


Figure .: Baselining Method

Using the baselining method discussed above, four (4) baseline groups have been calculated (Table 5.1 below) to evaluate model fitness in four aspects:

* Financial Services (FS) – all returns label from stocks within FS sectors are aggregated.
* Telecommunication & Media (TELCO) – all returns label from stocks within TELCO sectors are aggregated.
* FS INDEX – index returns based on weighted stock prices from constituent stocks.
* TELCO INDEX –index returns based on weighted TELCO stocks prices from constituent stocks.

Table .: Performance Baseline



# RESULTS

This section presents the result from model evaluation discussed previously. Baseline and model performance are summarized side by side for comparison, together with the parameters that optimized the performance metrics.

## Results By Group of Stocks and Index

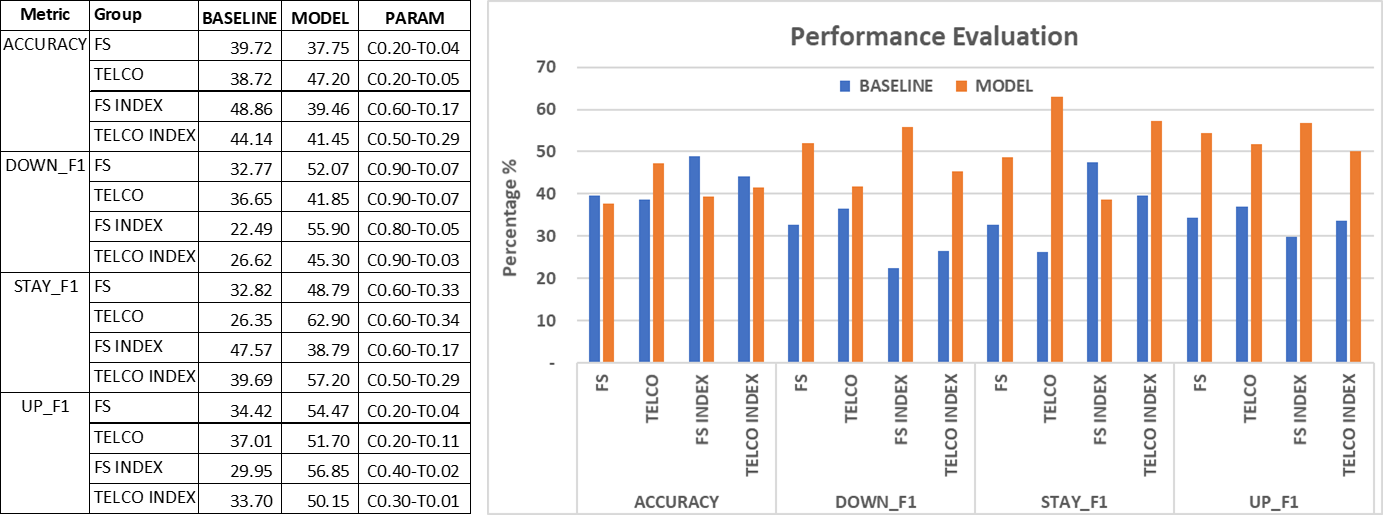
On accuracy, the model performs better than baseline only in TELCO group at 47.20%, which is 8% margin above baseline. All other groups performed 2% to 10% below baseline. Although the model did not outperform the baseline, predicted accuracies are all well above random guessing of 33%.

In general, the model F1 means outperform baselines in almost all categories (UP, STAY, DOWN). For UP category, model F1 means performs 14% to 27% better than baseline, and all F1 means are well above 50%. The optimum centre parameter was lower for this category (between 0.2 and 0.4). This can be understood as allowing more predictions on UP labels due to its increasing probability from RD00 to RD05.

On the other hand, the optimum center parameter is higher for F1 means prediction on DOWN category (between 0.80 and 0.90). Opposed to the trend of UP labels, probability of DOWN labels decreased from RD00 to RD05 too. The model adjusted to a higher centre to allow more predictions on DOWN label. F1 means for DOWN category performed 5% to 33% better than baseline.

As for STAY, only FS INDEX group did not perform well, which is 8% lower than baseline. The other groups outperform baseline within range of 16% to 36%. As STAY is the zone of ‘neutral’ opinion, its centers well balanced within 0.5 and 0.6. The result is summarized with table and chart in Table 6.1 below. The optimum parameters in the table are interpreted with “C” for center and “T” for threshold.

Table .: Model Performance Per Group



## Individual Stocks Results

In addition to group analysis, individual stocks were analyzed using the same methodology. Referring to Table 6.2, the average model accuracy is very close the baseline (less than 1% below), while F1 means outperform baseline in the range of 15% to 19%. As a whole, the model did well against baseline.

Table .: Model Performance Per Stock



## Comparison To Other Research

Similar approach of three categories Senti-WordNet based prediction in (Li et al., 2014) reported accuracy range between 10% to 40% for finance stocks in Hong Kong Stock Exchange. It is quite low compared to this research.

A different study approached a two categories prediction (UP/DOWN) conducted on KLSE revealed impressive 90% accuracies, 88% to 98% precision and recall. The authors (Im et al., 2014) built a manually annotated dictionary to better classify finance terms as opposed to Senti-Wordnet used in this research.

Authors in work (Kim et al., 2014) researched on prediction model for Korean stocks using machine-learning approach. Although not directly comparable with dictionary-based model, they reported 60% accuracy and 76% F1 score.

# CONCLUSION

The results obtained from this research answers the below questions set forth earlier in this research project:

1. Is there any significant relation between news sentiment and short-term stock price direction?

Answer: Quantitative analysis using sentiment scores and modeling techniques deployed in this research generally outperformed baselines. The sentiment-based model can predict well with F1 means between 45% to 53% across all categories, while outperforming baseline by 15% to 19% margin. Overall accuracy at 42.7% is also well within 1% margin from baseline. This performance indicates that news sentiment was indeed related to short term stock price direction and its relation is not due to randomness.

1. Between sector news and stock specific news, is there any difference in their impact on stock price and index direction?

Answer: For sector index group, model performance against F1 means baseline ranged by -8% to 33.41% margin. For stocks group, the performance is better with 5% to 36% exceeding baseline. As for accuracy, both sector index group performed badly with 2% to 10% below baseline. It can be concluded generally that the impact of sector news to its index direction is not as strong as stocks groups.

# Future Work

The outcome of this research had sparked several ideas. Leveraging on studies by other researchers, the methodology presented in this research can be expanded, integrated or completely reworked with ideas below:

1. Perform the same methodology using other sentiment scoring methods. This includes using different ratios, or libraries such as TextBlob and Vader.
2. Explore the use of different dictionary, such as LMD developed by (Loughran & McDonald, 2011), which specialized in finance and accounting domain.
3. Feature fine tuning and data quality improvement. There is no short of ideas how NLP features can be enhanced. Such includes considering only news body (discarding headers that are too short), counting the length article, better filters when news mapping to stocks.
4. Machine learning based model with word stemming or filters by dictionary. The main challenge is feature engineering and fine tune bag of words that best fit a model.
5. Machine learning based approach can be also be coupled with features generated by dictionary. For example, sentiment scores can be used together with bag of words in a multi-tier modeling.

# References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.* Paper presented at the Lrec.

Bharathi, S., & Geetha, A. (2017). Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems, 10*(3), 146-154.

Butler, M., & Kešelj, V. (2009). *Financial forecasting using character n-gram analysis and readability scores of annual reports.* Paper presented at the Canadian Conference on Artificial Intelligence.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014). *Using structured events to predict stock price movement: An empirical investigation.* Paper presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Esuli, A., & Sebastiani, F. (2006). *Sentiwordnet: A publicly available lexical resource for opinion mining.* Paper presented at the LREC.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management, 23*(3), 157-214.

Hagenau, M., Liebmann, M., Hedwig, M., & Neumann, D. (2012). *Automated news reading: Stock price prediction based on financial news using context-specific features.* Paper presented at the 2012 45th Hawaii International Conference on System Sciences.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.

Hor, W. L., Goh, W. X., & Ow, S. H. (2018). Impact of News Sentiment Analysis on Market Price Movement During the Dividend Period. *International Journal of Engineering \& Technology, 7*(4.11), 168-173.

Im, T. L., San, P. W., On, C. K., Alfred, R., & Anthony, P. (2014). Impact of Financial News Headline and Content to Market Sentiment. *International Journal of Machine Learning and Computing, 4*(3).

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics, 110*(3), 712-729.

Kim, Y., Jeong, S. R., & Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. *Int. J. Advance. Soft Comput. Appl, 6*(1), 2074-8523.

Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014). *On the Importance of Text Analysis for Stock Price Prediction.* Paper presented at the LREC.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14-23.

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.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Mitra, G., & Mitra, L. (2011). *The handbook of news analytics in finance* (Vol. 596): John Wiley & Sons.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & 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.

Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2013). An automated framework for incorporating news into stock trading strategies. *IEEE transactions on knowledge and data engineering, 26*(4), 823-835.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation, 60*(5), 503-520.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, 570.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS), 27*(2), 12.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems, 85*, 74-83.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11-21.

Stone, P. J., Dunphy, D. C., Ogilvie, D. M., & Smith, M. S. (1966). *The General Inquirer: a Computer Approach to Content Analysis*: Cambridge, Mass., & London.

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.

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.

Wang, S., Zhe, Z., Kang, Y., Wang, H., & Chen, X. (2008). An ontology for causal relationships between news and financial instruments. *Expert Systems with Applications, 35*(3), 569-580.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review, 50*(1), 49-73.

Zhang, W., & Skiena, S. (2010). *Trading strategies to exploit blog and news sentiment.* Paper presented at the Fourth international aAAI conference on weblogs and social media.

Appendix A: CODES AND DATA (GITHUB)

The data, codes and report for this research is publicly available at:

<https://github.com/yongks/news>

The github repository is arranged in following structure:

|  |  |
| --- | --- |
| Folders/File | Description |
| Modules | NewsDatabase and EikonDatabase organize in reusable modules. The modules can be re-used for future news and price update  Updating share price with EikonDatabase require Eikon account which is not included here. However, this module is still usable to retrieve price data previously downloaded for the scope of this project. |
| database | Data used by Modules |
| processing | Five phases of data processing in python scripts  Phase 1 – Transform News to various NLP features  Phase 2 – Mapping News to Entity (including Sector)  Phase 3 – Merging NLP features (Phase1) and Entity (Phase2) for final sentiment scores  Phase 4 – Calculating Price returns for current and next 5 days  Phase 5 – Modeling, Evaluation, Baselining and Results |
| data | Intermediate data generated from Phase1 to Phase 4 |
| results | Results generated from Phase 5 |
| reports | PowerPoint presentation and this research project |
| settings.cfg | Definition of various path and API keys |

Note:

This repository utilized Github LFS (large file system) for all CSV files due to news data exceeded 100MB per file limit.