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

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

yong keh soon

Thesis submitted in 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”):

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:

BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE  
Abstract

A review of financial news analytics and its applications on stock trading, is discussed in this survey paper. It is widely recognized that news does have it's influence on stock price movements. In the Internet age of today, the sources and volume of such news continues to grow. New technologies to enable automatic or semi-automatic news collection, extraction, aggregation and categorization are emerging. 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 is growing exponentially in Natural Language Processing (NLP) to develop models that can understand human generated text. In view of it's application in the financial industry, we review various textual preprocessing techniques, including text and semantic extraction, sentiment scoring, machine learning advancement, to determine the quantitative sentiment scores. We consider the various types of news available and how these can be processed to form inputs to financial models. We report on how news analytics are being applied in trading strategies for risk management and to enhance investment return.

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

BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE  
Abstrak

(Please delete this part): Abstract should not be more than 500 words in both versions; Bahasa Malaysia and English, typed in single paragraph.

Use *abstract title* for heading and *abstract text* for body of your abstract.

Keywords: Maximum of five (5) keywords.

Acknowledgements

This work is dedicated to my wife, Sabrina, who had never complained about countless weekends spent on projects and assignments for the past two years. And to my two kids 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 to enrich your knowledge and stay curious. I hope you have a similar attitude towards life with me as the role model, and it is not so tough.

Lastly to my late father. I remember attending this master program orientation in the afternoon, on the day of your funeral. Since young, you know I’m a curios kid who love silly science fiction. I always had your trust and support throughout my life that I’m deeply grateful for. Though you don’t talk much, this work is only possible because of your blessings.

Table of Contents

[BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE Abstract iii](#_Toc39354280)

[BUSINESS NEWS MINING AND IMPACT ON STOCK PRICE Abstrak iv](#_Toc39354281)

[Acknowledgements v](#_Toc39354282)

[Table of Contents vi](#_Toc39354283)

[List of Figures ix](#_Toc39354284)

[List of Symbols and Abbreviations xi](#_Toc39354285)

[List of Appendices xii](#_Toc39354286)

[CHAPTER 1: introduction 1](#_Toc39354287)

[1.1 Overview 1](#_Toc39354288)

[1.2 Background 1](#_Toc39354289)

[1.3 Problem Statement 1](#_Toc39354290)

[1.4 Objectives 1](#_Toc39354291)

[1.5 Research Questions 2](#_Toc39354292)

[1.6 Scope of Research 2](#_Toc39354293)

[CHAPTER 2: literature review 4](#_Toc39354294)

[2.1 Background 4](#_Toc39354295)

[2.2 Introduction 4](#_Toc39354296)

[2.3 Content Analysis 5](#_Toc39354297)

[2.3.1 News Semantic 6](#_Toc39354298)

[2.3.2 Feature Processing 9](#_Toc39354299)

[2.3.3 Dimension Reduction 10](#_Toc39354300)

[2.3.4 Feature Representation 11](#_Toc39354301)

[2.4 Prediction Model 12](#_Toc39354302)

[2.4.1 Dictionary Based Approach 12](#_Toc39354303)

[2.4.2 Machine Learning Based Approach 13](#_Toc39354304)

[2.5 Integrating News Categorization 14](#_Toc39354305)

[2.6 News Sources 16](#_Toc39354306)

[2.7 Conclusion 19](#_Toc39354307)

[CHAPTER 3: Research Methodology 20](#_Toc39354308)

[3.1 Terminology Used 20](#_Toc39354309)

[3.1.1 Financial Terminology 20](#_Toc39354310)

[3.1.2 Data Science Terminology 20](#_Toc39354311)

[3.2 Data Sources and Acquisition 21](#_Toc39354312)

[3.2.1 News Acquisition (Text) 21](#_Toc39354313)

[3.2.2 Stock Price Acquisition (Numerical) 23](#_Toc39354314)

[3.3 System And Data Management Design 25](#_Toc39354315)

[3.3.1 Database Schema 26](#_Toc39354316)

[3.4 Data Processing 27](#_Toc39354317)

[3.4.1 Processing Missing Data and Errors 27](#_Toc39354318)

[3.4.1.1 Missing Numerical Data and Errors 27](#_Toc39354319)

[3.4.1.2 Missing News Data and Errors 28](#_Toc39354320)

[3.4.2 Feature Engineering 28](#_Toc39354321)

[3.4.2.1 Stock Price Return 28](#_Toc39354322)

[3.4.2.2 Sector Index Price Return 30](#_Toc39354323)

[3.4.2.3 News Sentiment 32](#_Toc39354324)

[3.4.3 News To Stock Entity Mapping 41](#_Toc39354325)

[3.4.4 Data Aggregation 42](#_Toc39354326)

[3.4.5 Returns Labelling 43](#_Toc39354327)

[CHAPTER 4: Data Interpetation 45](#_Toc39354328)

[4.1 News 45](#_Toc39354329)

[4.1.1 Overall News 45](#_Toc39354330)

[4.1.2 Sector News 46](#_Toc39354331)

[4.2 Share Price 46](#_Toc39354332)

[4.2.1 Share Price Returns 47](#_Toc39354333)

[4.2.2 Returns Label 48](#_Toc39354334)

[CHAPTER 5: Data Modeling 50](#_Toc39354335)

[5.1.1 The Model 50](#_Toc39354336)

[5.1.2 Performance Metrics 51](#_Toc39354337)

[5.1.3 Performance Baseline 52](#_Toc39354338)

[CHAPTER 6: RESULTS 54](#_Toc39354339)

[CHAPTER 7: CONCLUSION 56](#_Toc39354340)

[CHAPTER 8: Future Works 57](#_Toc39354341)

[CHAPTER 9: References 58](#_Toc39354342)

[Appendix A: newsdatabase code 61](#_Toc39354343)

List of Figures

[Figure 2.1: Evolution of NLP Research 4](#_Toc39353520)

[Figure 2.2: Sample Business News, (Ding, Zhang, Liu, & Duan, 2014) 6](#_Toc39353521)

[Figure 2.3: From News To Decision 7](#_Toc39353522)

[Figure 2.4: News Feature Processing 9](#_Toc39353523)

[Figure 2.5: Dictionary Based Approach 12](#_Toc39353524)

[Figure 2.6: Machine Learning Based Approach 14](#_Toc39353525)

[Figure 2.7: Proposed Predictive System, adopted from (Shynkevich et al., 2016) 15](#_Toc39353526)

[Figure 3.1: Web Page For News Searching 23](#_Toc39353527)

[Figure 3.2: Web Article Web Page 23](#_Toc39353528)

[Figure 3.3: Eikon API Call For Pricing Data In Python 24](#_Toc39353529)

[Figure 3.4: CIMB Price Chart 25](#_Toc39353530)

[Figure 3.5: Overall System Design 26](#_Toc39353531)

[Figure 3.6: Database Schema 26](#_Toc39353532)

[Figure 3.7: MBBM.KL Stock Price Return Illustration 29](#_Toc39353533)

[Figure 3.8: Graphical Representation of Senti WordNet Opinion Related Properties (Esuli & Sebastiani, 2006) 35](#_Toc39353534)

[Figure 3.9: Sentiment Scoring and Label For News Article 36](#_Toc39353535)

[Figure 3.10: Sentiment Score Processing Flowchart 39](#_Toc39353536)

[Figure 3.11: Example Sentiment Scores DataFrame 40](#_Toc39353537)

[Figure 3.12: Example News To Stock Entity Mapping DataFrame 41](#_Toc39353538)

[Figure 3.13: News Sentiment To Entity DataFrame Example 42](#_Toc39353539)

[Figure 3.14: News Sentiment To Entity DataFrame Example 42](#_Toc39353540)

[Figure 3.15: Share Price Labels Example 44](#_Toc39353541)

[Figure 4.1: Overall News By Sources 45](#_Toc39353542)

[Figure 4.2: News and Market Capital Scatter Plot 46](#_Toc39353543)

[Figure 4.3: Sector News 46](#_Toc39353544)

[Figure 4.4: Sector Price Returns Overall Chart 47](#_Toc39353545)

[Figure 4.5: Shares Price Returns Chart By Stock 48](#_Toc39353546)

[Figure 4.6: Returns Label Distribution For Stocks 48](#_Toc39353547)

[Figure 4.7: Returns Label Distribution For Sector Index 49](#_Toc39353548)

[Figure 5.1: Prediction Model 50](#_Toc39353549)

[Figure 5.2: Three Categories Confusion Matrix 52](#_Toc39353550)

[Figure 5.3: Baselining Method 53](#_Toc39353551)

List of Tables

[Table 1.1: Top 10 Largest Capital Stocks (Financial Services) 3](#_Toc39353552)

[Table 1.2: Top 10 Largest Capital Stocks (Telecommunication and Media) 3](#_Toc39353553)

[Table 2.1: Paper Summary Of Content Analysis Method 6](#_Toc39353554)

[Table 2.2: Financial Event News Types In 8-K Reports (Schumaker & Chen, 2009) 16](#_Toc39353555)

[Table 2.3: List Of Financial Event Types In 8-K Reports ((Lee et al., 2014) 17](#_Toc39353556)

[Table 2.4: Results Of Proposed Model (Lee et al., 2014)) 18](#_Toc39353557)

[Table 2.5: Result of Temporal Aspect Model (Lee et al., 2014) 18](#_Toc39353558)

[Table 3.1: Financial Terminology 20](#_Toc39353559)

[Table 3.2: Data Science Terminology 20](#_Toc39353560)

[Table 3.3: News Search Keywords 22](#_Toc39353561)

[Table 3.4: Stock Price Data Count 27](#_Toc39353562)

[Table 3.5: MBBM.KL Stock Price Return Example 29](#_Toc39353563)

[Table 3.6: Sector Index Price Calculation 30](#_Toc39353564)

[Table 3.7: Stock MarketCap Weighing 31](#_Toc39353565)

[Table 3.8: Market Capitalization Weighted Financial Services Index Price 31](#_Toc39353566)

[Table 3.9: Structure of WordNet (Keselj, 2009) 33](#_Toc39353567)

[Table 3.10: Various Meaning of The Same Word 34](#_Toc39353568)

[Table 3.11: SWN Scores Example 35](#_Toc39353569)

[Table 3.12: News Sentiment Score (Positive Example 1) 37](#_Toc39353570)

[Table 3.13: News Sentiment Score (Positive Example 2) 37](#_Toc39353571)

[Table 3.14: News Sentiment Score (Negative Example 1) 38](#_Toc39353572)

[Table 3.15: Penn Treebank POS Tag Mapping to WordNet 40](#_Toc39353573)

[Table 5.1: Performance Baseline 53](#_Toc39353574)

[Table 6.1: Model Performance Per Group 55](#_Toc39353575)

[Table 6.2: Model Performance Per Stock 55](#_Toc39353576)

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: Example ……………………………………………………………... | 7 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# introduction

## Overview

First topic in each chapter should numbered with “chapter number”.1. Use Heading 2 or h2 for title and for table of content TOC3 must be used.

## Background

## 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 (Tan, Phang, Chin, 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 can be used 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 of the 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 does have an impact on stock prices. In an increasingly competitive and complex financial landscape, getting an informational edge is vital to constructing and maintaining a superior equity portfolio. Alternative data sources such as news, blogs and tweets are becoming increasingly important sources of insight, to be included as part of equity strategies (Xing, Cambria, & Welsch, 2018). Fundamental analysis assisted by NLP, is now a critical marriage of techniques to unlock the complete picture of how experts and the masses feel about the market.

Text data is one of the largest and fastest growing forms of alternative data. Uncovering investment insights requires not only domain knowledge of finance, but also a strong grasp of data science and machine learning principles. 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.

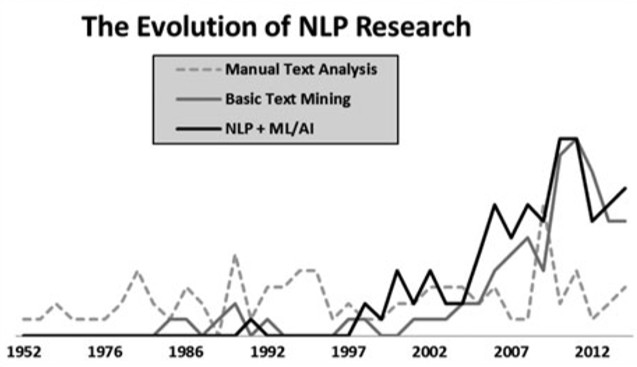


Figure .: Evolution of NLP Research

## 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.

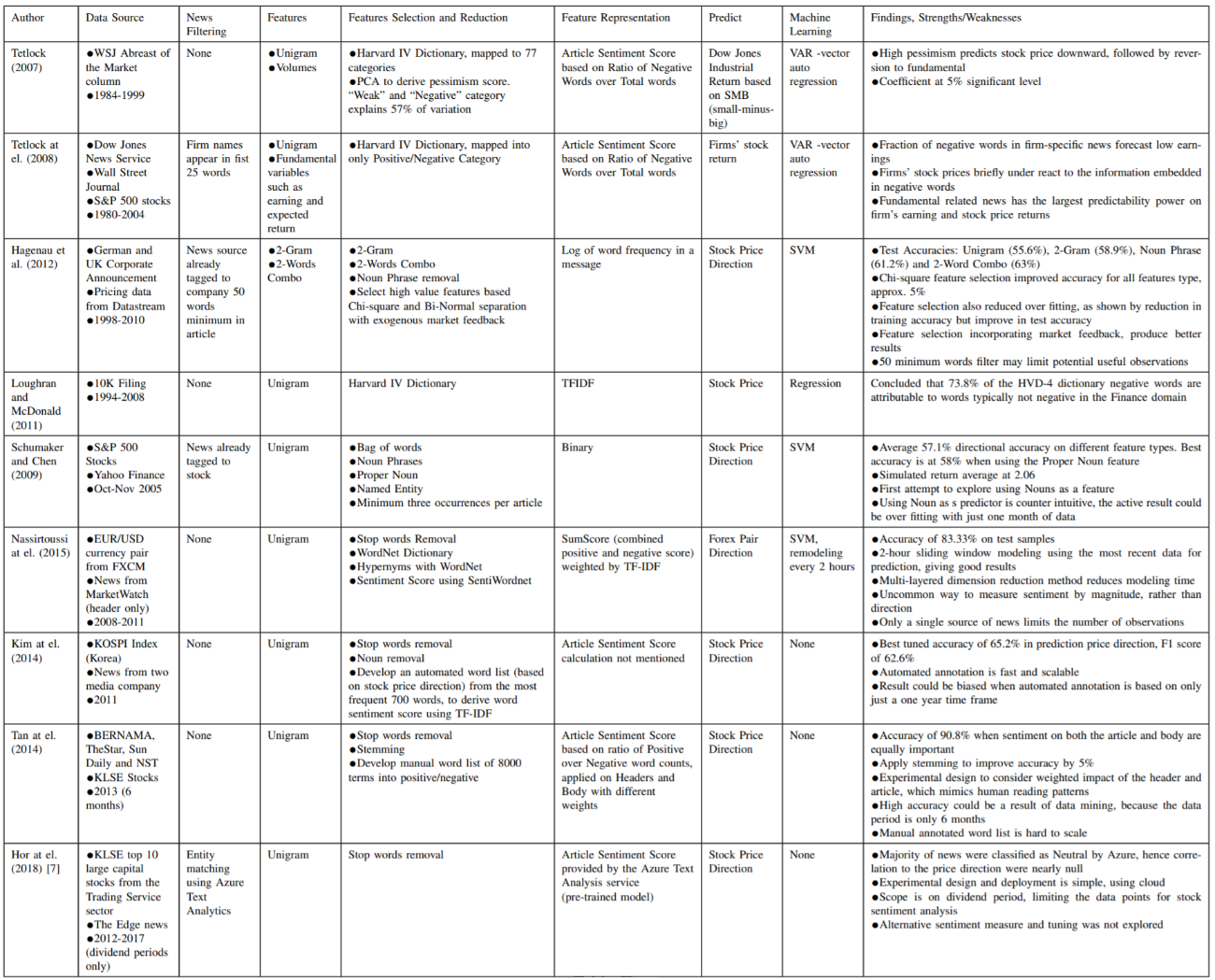
In more advanced implementations by commercial products, this review provides an understanding of fine-grained approaches in extracting the degree of news relevance of the target stock to improve financial news-based forecasting. We reviewed the most common event extraction techniques for textual data through a multi-dimensional, qualitative evaluation. 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

In this section, we shall discuss some of the common NLP techniques that are 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



### News Semantic

Business related text data sources contain potentially relevant and valuable information that could be captured and understood by companies 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.

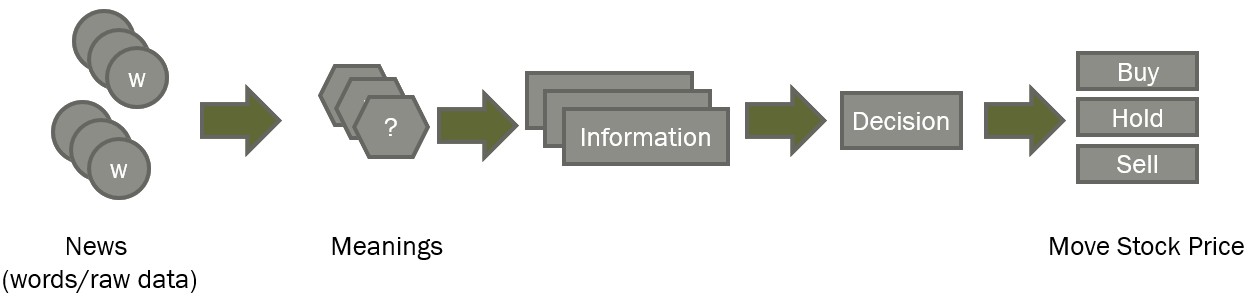


Figure .: From News To Decision

The way humans understand what’s being said is almost an unconscious process. We rely on what we already know about language itself and about the concepts present in a body of text. However, this process is extremely complicated for machines. One of the approaches used by machines, is based on keywords or statistics, or even pure machine learning using a matching or frequency technique for clues as to what the text is “about”. A common application of this approach is Sentiment Analysis, which provides scoring of positive or negative sentiment to text data. This simple sentiment feature is difficult to identify key events embedded in financial news, and even more difficult to model the impact of events on stock market prediction. As an example, the event “Apple has sued Samsung for copying its patented ‘look-and-feel’ of its tablet and smartphone”. using term-level features {“Apple”, “sued”, “Samsung”, “copying”, ...} alone could inaccurately predict or represent the sentiment of the news and its impact to Apple stock price movements to both Apple and Samsung. These unstructured terms cannot indicate the actor and object of the event.

In (Ding et al., 2014), the authors proposed using structured information to represent events and develop a prediction model to analyze the relationship between the events and the stock market. 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 tuple of (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. For example, the event “Sep 3, 2013 - Microsoft agrees to buy Nokia’s mobile phone business for $7.2 billion.” is modeled as: (Actor = Microsoft, Action = buy, Object = 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 identified as “sues”, and its subject and object is “Apple” and “Samsung”, respectively. The event is then represented as (Apple, Samsung, sues).

A more advanced approach is to process the logical structure of sentences in the text to identify the most relevant elements and understand the topic being discussed. It would also understand the relationships between different concepts in the text. For example, it understands that the text is about “politics” and “economics” even if it does not contain the the actual words, but related concepts such as “election,” “Democrat,” “speaker of the house,” or “budget,” “tax” or “inflation”. Using the same concept, authors in (Wang, Zhe, Kang, Wang, & Chen, 2008) had an ontology to casually relate news to financial instruments, in a way knowledge is captured by mapping instruments to news categories.

Information extraction can be a costly task and usually requires predefined grammar and domain specific detection rules. Due to this, the bag-of-words-based document representation has been the mainstream method for a long time. Whichever method that is used to analyze the raw data of words and sentences, behavioral scientists have become increasingly skilled at uncovering the evidence of sentiment or it's effect within text.

### Feature Processing

In the past, the volume and velocity of textual data were manageable enough to be manually analyzed by teams of human experts. But given the volume of text data being currently produced daily, it is now an untenable task for even a large team of researchers to wade through it all. We have summarized three common aspects of the pre-processing steps from our reviewed body of works, namely: feature selection, dimension reduction and feature representation. illustrates an overview of the processes.

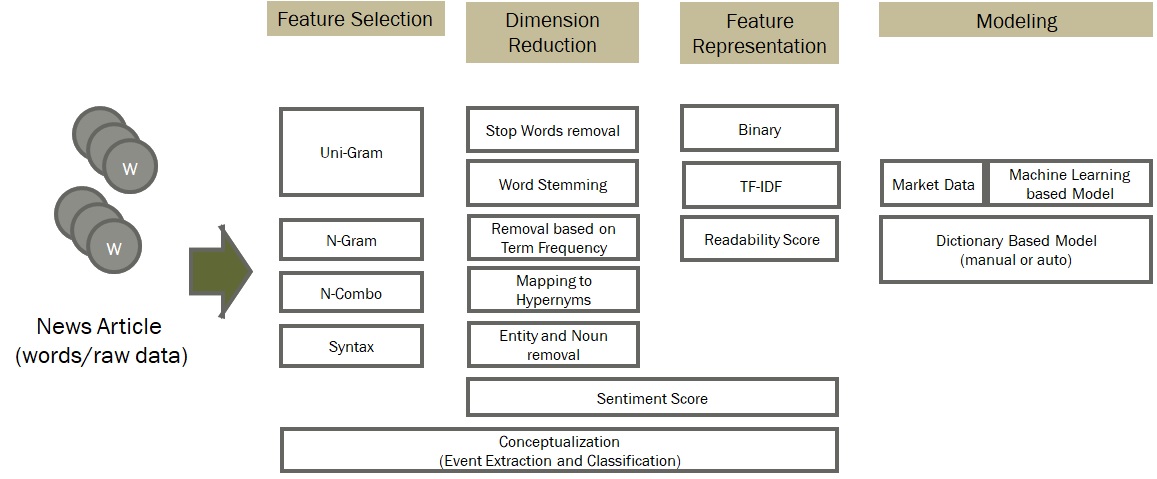


Figure .: News Feature Processing

A variation of N-Grams called the "N-Combo" is used in (Hagenau, Liebmann, Hedwig, & Neumann, 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, such as "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, Xie, Chen, Wang, & Deng, 2014) and (Kim, Jeong, & Ghani, 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, Aghabozorgi, Wah, & Ngo) 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 are also used to count the frequency of the words in an article, this 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 by the minimum number 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). It is based on counting the number of documents in the collection, being searched, that contains (or are indexed by) the term in question. The intuition was that a term which occurs in many documents is not a good discriminator and should be given less weight than the one which occurs in a few documents, the measure was a heuristic implementation of this intuition. Coupled with TF, a well-known measurement of TF\*IDF, is a giant leap in information retrieval and had proved to be extraordinarily robust and difficult to beat (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, Saar-Tsechansky, & Macskassy, 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 the ability to partly understand the annual report and learn what it means for future performance. The authors used programming packages available by (Fast, 1999) to calculate the readability score based on Flesch, Flesch-Kincaid and Fog Index.

## Prediction Model

### Dictionary Based Approach

Computer-based, general-purpose content analysis dictionaries are a kind of automated codebook. 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" model in NLP. Below () illustrates a generic process of using a dictionary-based approach to extract the sentiment from text.

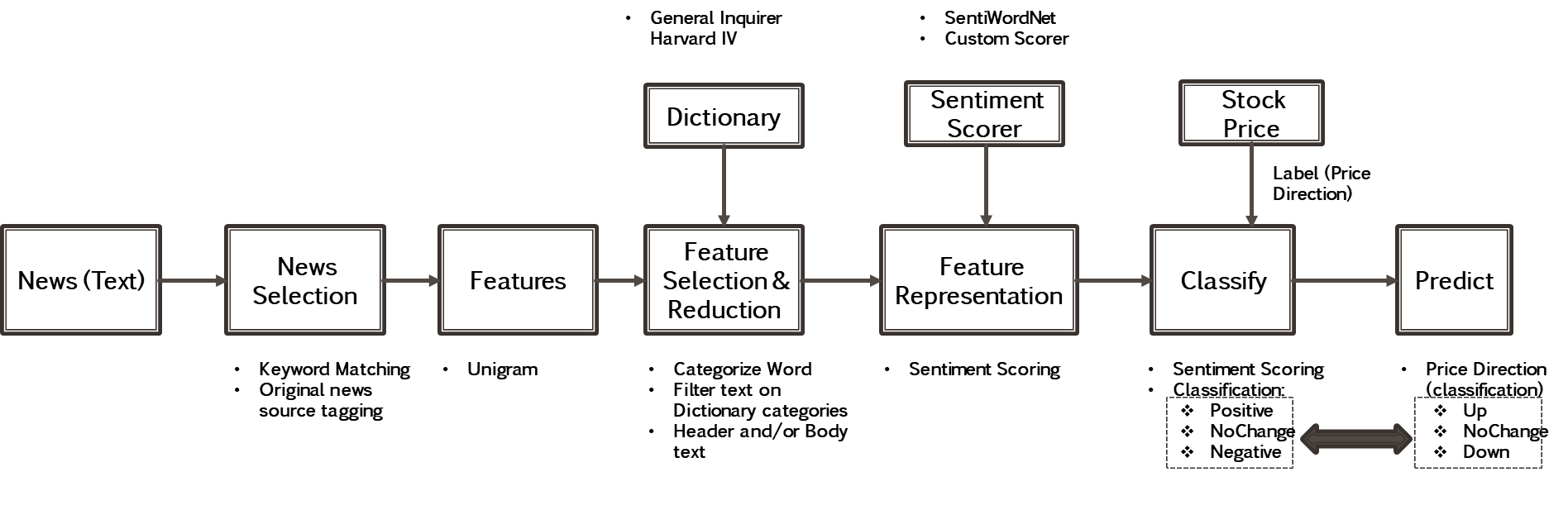


Figure .: Dictionary Based Approach

More than half a century ago in (Stone, Dunphy, & Smith, 1966), the author developed a system called General Inquirer (GI), which uses the Harvard Psychosocial Dictionary in parsing and categorizing text. The system was tested on 66 suicide notes — 33 written by men who had taken their own lives and 33 written by men who were asked to produce simulated suicide notes. 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 General Inquirer alongside the Harvard-IV dictionary to calculate the fraction of negative words in the "Abreast of the Market" Wall Street Journal column. A plethora of papers post (Tetlock, 2007) apply a similar methodology to measure the positive versus negative tone of news across a wide variety of finance and accounting applications. In particular, authors of (Loughran & McDonald, 2011) refined the Harvard-IV dictionary to more finance-centric definitions of positive and negative words. They had performed regression analysis on US 10K filing data with variations of stock data, such as returns and size. Their conclusion is that 73.8% of the negative words in the Harvard-IV dictionary typically do not attribute to negative words in finance. 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 words in the negative word list are equally negative and all words in the positive list are equally positive, 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 () illustrates the overview of this process.

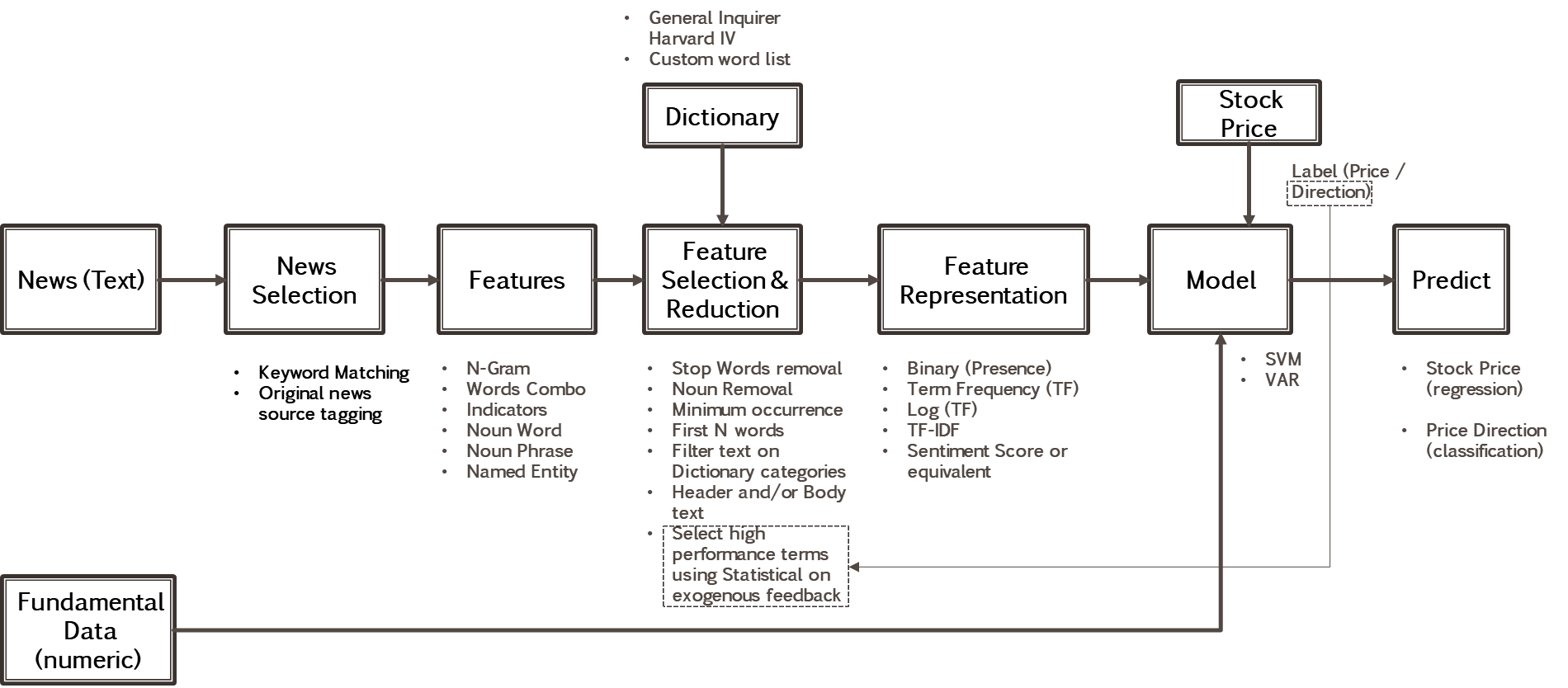


Figure .: Machine Learning Based Approach

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 whether financial news that has different degrees of relevance to the target stock can provide an advantage in financial news-based forecasting. Stocks are assigned to five news categories based on the Global Industry Classification Standard (GICS), which are stock specific (SS), sub-industry-specific (SIS), industry-specific (IS), group-industry-specific (GIS) and sector-specific (SES) news items, hence a total of five subsets for relevance analysis.

shows the proposed predictive system by (Shynkevich et al., 2016). News articles are downloaded from LexisNexis. LexisNexis contains additional information such as relevant companies and their relevance scores, to supplement its news articles. The news article that has a relevance score of more than 85% is added to the corresponding subset. Authors in (Shynkevich et al., 2016) classify news articles to be positive or negative based on the next trading day, whether the stock price increased or decreased 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, each news article is represented as a vector of 500 TF\*IDF values each, which corresponds to a feature. A sparse matrix size of [number of data points] \* 500 is then constructed. This procedure is to be applied separately to other subsets. Each data distance is assigned an ‘Up’ or ‘Down’ label for machine learning. The methodology is based on the first principle and no pre-defined dictionary library is used for preprocessing (Shynkevich et al., 2016).

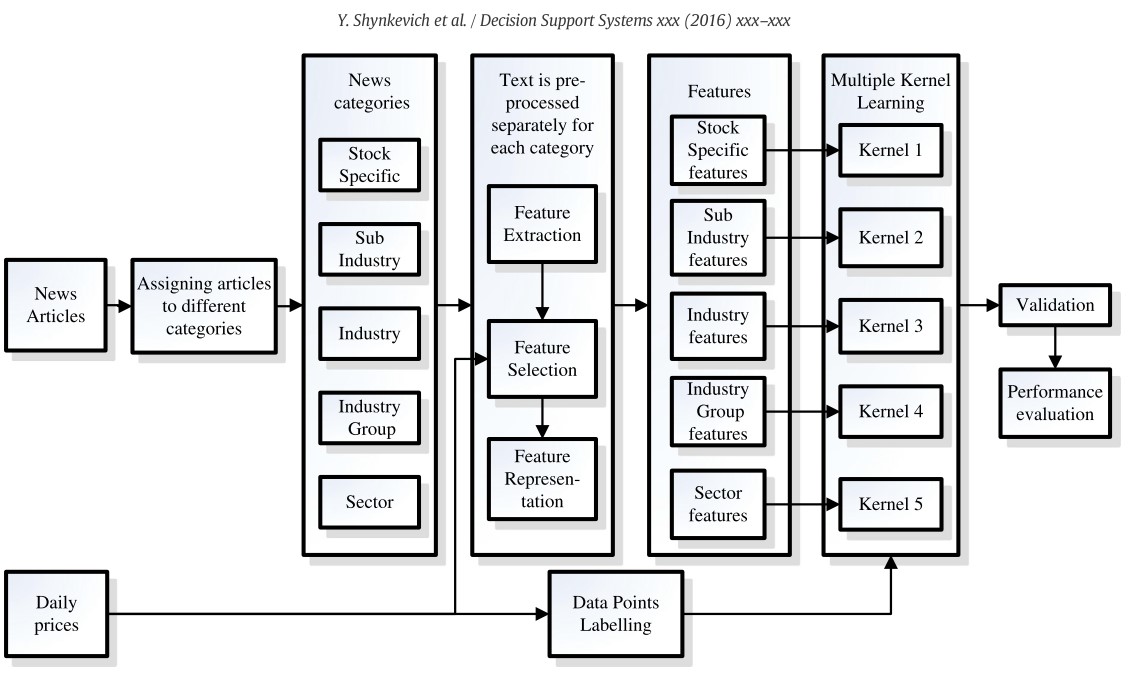


Figure .: Proposed Predictive System, adopted from (Shynkevich et al., 2016)

## News Sources

Multiple sources of textual data can be analyzed for stock price prediction, as presented below (Table 2.2). 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 in a prospect statement. Analyst recommendations, news, and discussion boards can provide a less biased view of a company (Schumaker & Chen, 2009). These textual data can be classified by categories which can be relevant to the stock market price prediction. These textual data can be classified by categories which can be relevant to the stock market price prediction.

Table .: Financial Event News Types In 8-K Reports (Schumaker & Chen, 2009)

|  |  |  |  |
| --- | --- | --- | --- |
| **Textual Source** | **Types** | **Examples** | **Description** |
| Company Generated Sources | SEC Reports | 8K | Reports on Significant Changes |
| 10K | Annual Reports |
| SC Reports | Financial Results | |
| General Announcements | |
| General Meetings | |
| Independently Generated Sources | Analyst Created | Recommendations | Buy/Hold /Sell assessments |
| Stock Alerts | Alerts for Share Prices |
| Bursa Market Place | Buy/Hold /Sell Assessments |
| News | Financial Times | Financial News Stories |
| Wall Street Journal | Financial News Stories |
| The Edge Market | Financial News Stories |
| Malaysia -StockBiz | Stock Market Blogs, News and Stock Forum |
| PRNewsWire | Breaking Financial News Article |
| Yahoo Finance | 45 Financial News Wire Sources |
| Discussion Boards | The Motley Fool | Forum to share stock-related Information |
| Binvestor | Stock Market Blogs, News and Stock Forum |
| Facebook | Stock Market Blogs, Forum |

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

|  |  |
| --- | --- |
|  | **Business Event Categories In 8-K Reports** |
| 1 | Material Definitive Agreements |
| 2 | Bankruptcies or Receiverships |
| 3 | Director is Elected |
| 4 | Director Departs |
| 5 | Asset movement: Acquisition or Sale |
| 6 | Result of Operations and Financial Condition |
| 7 | Material Direct Financial Obligations (bonds, debentures) |
| 8 | Exit or Disposal Plans |
| 9 | Material Impairments |
| 10 | Delisting or Transfer Exchange Notices |
| 11 | Changes in Control of the Company |
| 12 | Changes in Executive Management |

They build a corpus that aligns the description of financial events reported in 8-K documents with the corresponding stock price. In order to build the corpus, the difference in the company’s stock price before and after the publishing of the report is calculated for each 8-K document, which is then converted into an Up/Down/Stay label. Authors in (Lee et al., 2014) then generated numeric or event categorical features of four non-linguistic feature types, including the volatility index, recent stock price changes, earning surprise and event category (the reason for filing the form 8-K). 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, DOWN, STAY), in response to a financial event. The features and the labels are then used for model training (Lee et al., 2014).

summarizes the accuracy results. All the non-linguistic features achieve an accuracy of about 50% and with the help of unigrams with simple negation heuristics, the performance is improved 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))

|  |  |
| --- | --- |
| **System** | **Accuracy** |
| Baseline 1 | 49.4 |
| Baseline2 | 50.1 |
| Unigram model | 54.4 |
| NMF 50 | 54.7 |
| NMF 100 | 55.4 |
| NMF 200 | 55.3 |
| Ensemble | 55.5 |

below from the same paper shows the predictive power of linguistic and non-linguistic features as we move farther away from the event. 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)

|  |  |  |  |
| --- | --- | --- | --- |
| **Price Change Interval** | **Baseline2** | **Unigram** | **Difference** |
| BE -> l-Day | 50.11 | 54.4 | 4.29 |
| BE -> 2-Day | 49.72 | 53.07 | 3.35 |
| BE -> 3-Day | 50.56 | 52.45 | 1.9 |
| BE -> 4-Day | 50.95 | 52.62 | 1.67 |
| BE -> 5-Day | 51.06 | 52.37 | 1.31 |

Similar research conducted by (Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2013) also affirmed the findings. The authors quantify the news variable based on the events extracted from the text of news messages using the proprietary application. They concluded that the assigned expert-defined impacts of the extracted events correlate well with the returns generated by these events. Authors in (Mitra & Mitra, 2011) also find that event novelty is an important element of a news-based strategy.

## Conclusion

A general framework on feature processing and transformation is discussed, which generalized into dictionary based and machine learning-based approach. 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 and trading strategy. 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 for a particular day |
| Reuters Instrument Code (RIC) | It is the short code used by Reuters to represent 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) | A software intermediary that allows two applications to talk to each other. |
| DataFrame | A data structure representing a 2-dimensional table with variables as column and rows representing records. |
| Natural Language Processing (NLP) | Broadly defined as the automatic manipulation of natural language, like speech and text, by software. |
| Open Source Software | Software with source code that anyone can inspect, modify, and enhance |
| Wordnet | WordNet is a lexical database for the English language based on conceptual lookup. It organizes lexical information in terms of word meanings rather than word form. It is a project began by George A.Miller (Miller, 1995). |
| Senti-Wordnet | A lexical resource for opinion mining. It extends WordNet synset with concept of sentiment score. Each synset is assigned three sentiment scores, namely positive, negative and objective scores. (Baccianella, Esuli, & Sebastiani, 2010) |

## Data Sources and Acquisition

This section describe in methods to acquire data from sources as defined in Section 1.6 Scope of Research.

### News Acquisition (Text)

This research focused only on business and finance related news to minimize noises. News data is 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. The algorithm took Exchange Ticker and Keywords (Table 3.3) as the search parameters and loops through all stocks in scope, described in detail below:

1. Identify the URL of the news provider search engine
2. For each stock in scope:
   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
   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 computer file

Table .: News Search Keywords

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RIC** | **EXCHANGE TICKER** | **SECTOR** | **COMMONNAME** | **News Search Keywords** |
| MBBM.KL | MAYBANK | FS | Malayan Banking Bhd | MAYBANK, Malayan Banking |
| PUBM.KL | PBBANK | FS | Public Bank Bhd | PBBANK, Public Bank |
| CIMB.KL | CIMB | FS | CIMB Group Holdings Bhd | CIMB Bank, CIMB |
| HLBB.KL | HLBANK | FS | Hong Leong Bank Bhd | Hong Leong Bank, HLBANK |
| RHBC.KL | RHBBANK | FS | RHB Bank Bhd | RHB Bank, RHBBANK |
| HLCB.KL | HLFG | FS | Hong Leong Financial Group Bhd | HLFG, Hong Leong Financial Group |
| AMMB.KL | AMBANK | FS | AMMB Holdings Bhd | Ammb Holdings, AMMB, Ambank |
| BIMB.KL | BIMB | FS | BIMB Holdings Bhd | BIMB, Bank Islam |
| LOND.KL | LPI | FS | LPI Capital Bhd | Lonpac, LPI |
| MBSS.KL | MBSB | FS | Malaysia Building Society Bhd | Malaysia Building Society, MBSB |
| MXSC.KL | MAXIS | TELCO | Maxis Bhd | Maxis |
| DSOM.KL | DIGI | TELCO | DiGi.Com Bhd | Digi, DIGI.COM |
| AXIA.KL | AXIATA | TELCO | Axiata Group Bhd | Axiata |
| TLMM.KL | TM | TELCO | Telekom Malaysia Bhd | Telekom, TM |
| ASTR.KL | ASTRO | TELCO | Astro Malaysia Holdings Bhd | Astro |
| TCOM.KL | TIMECOM | TELCO | TIME dotCom Bhd | TIMECOM, Time dotCom |
| GRNP.KL | GPACKET | TELCO | Green Packet Bhd | Green Packet |
| OCKG.KL | OCK | TELCO | OCK Group Bhd | OCK |
| MDCH.KL | MEDIAC | TELCO | Media Chinese International Ltd | Media Chinese |
| STAR.KL | STAR | TELCO | Star Media Group Bhd | The Star |

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.

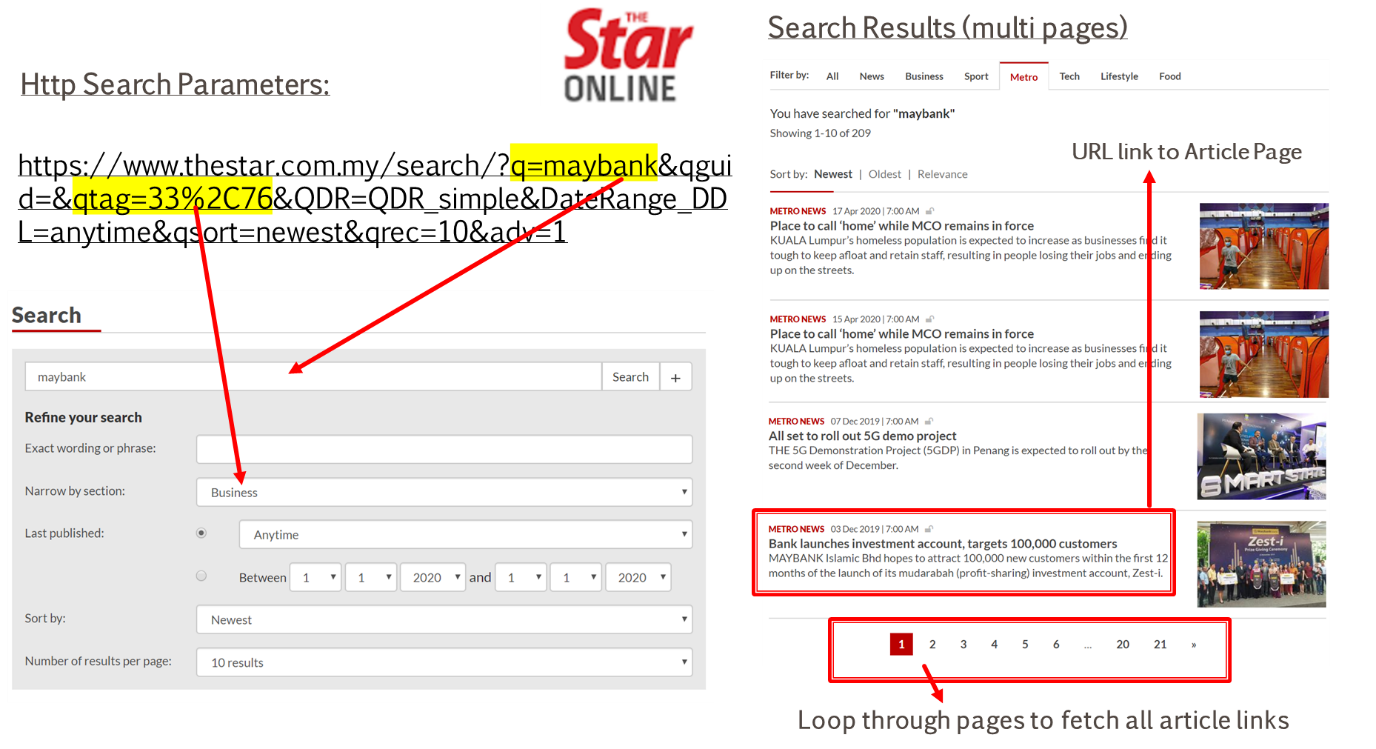


Figure .: Web Page For News Searching

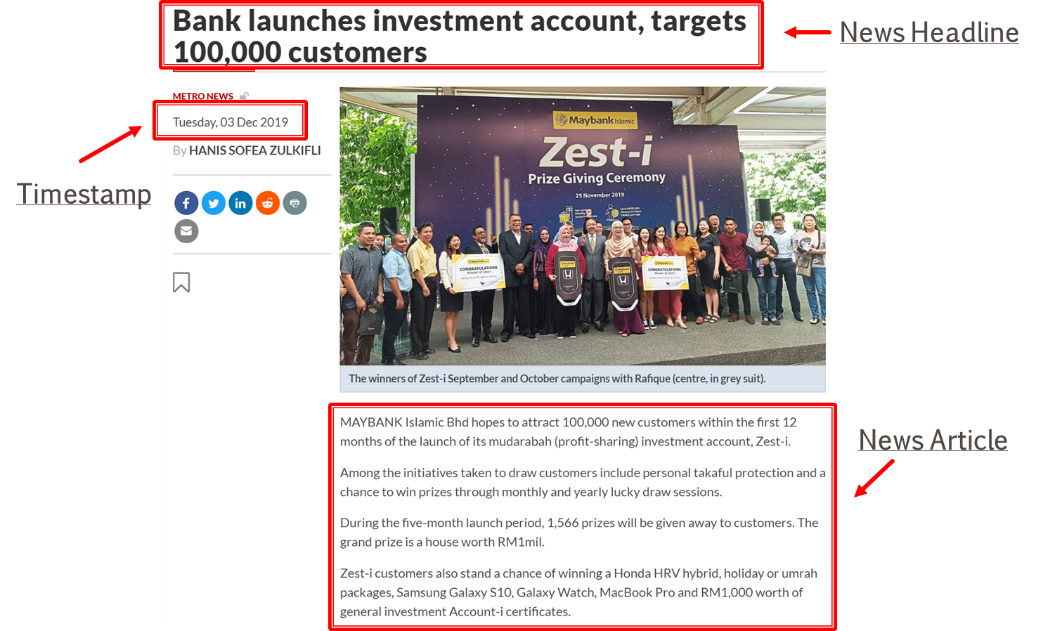


Figure .: Web Article Web Page

### Stock Price Acquisition (Numerical)

Stock Price related data is relatively simpler as to its structured data source. 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 in KLSE context is very limited (close to none for most stocks).

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

This project uses Eikon API (Figure 3.3) to only obtain three historical variables, namely OPENPRICE, CLOSEPRICE and OUTSTANDINGSHARES with the algorithm below:

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

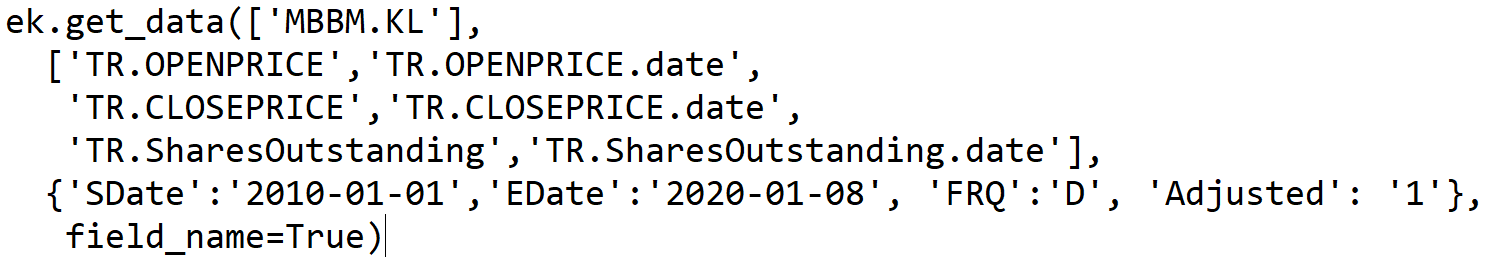


Figure .: Eikon API Call For Pricing Data In Python

“Adjusted” daily price changes and outstanding shares are captured for 10 years duration within scope. Number adjustment is important to reflect true value after the effect of corporate action such as stock splits, consolidation, or right issues. For example, when a stock split from one to two due to bonus issue, the share value will be reduced by half (Figure 3.4). If the outstanding shares are also not adjusted, market capitalization would be wrongly reported as halved overnight.

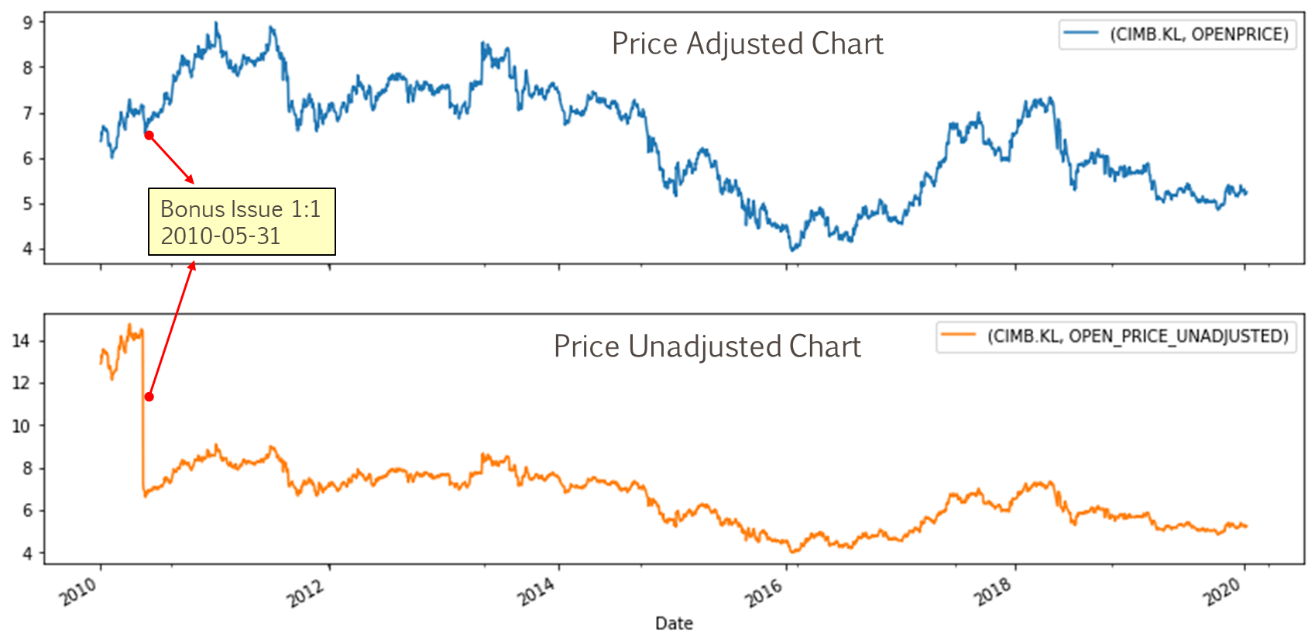


Figure .: CIMB Price Chart

## System And Data Management Design

This project used Python as the only programming language for end to end data pipeline, on a single computer. Except for Eikon which require license, all other components and library used are from Open Source. System diagram depicted in Figure 3.5 below shows the end to end data flow from Internet sources to file based local data management system, data preprocessing, feature engineer, modeling and finally produced the result. Two Python object library has been created as reusable components to acquire price and text data from internet.

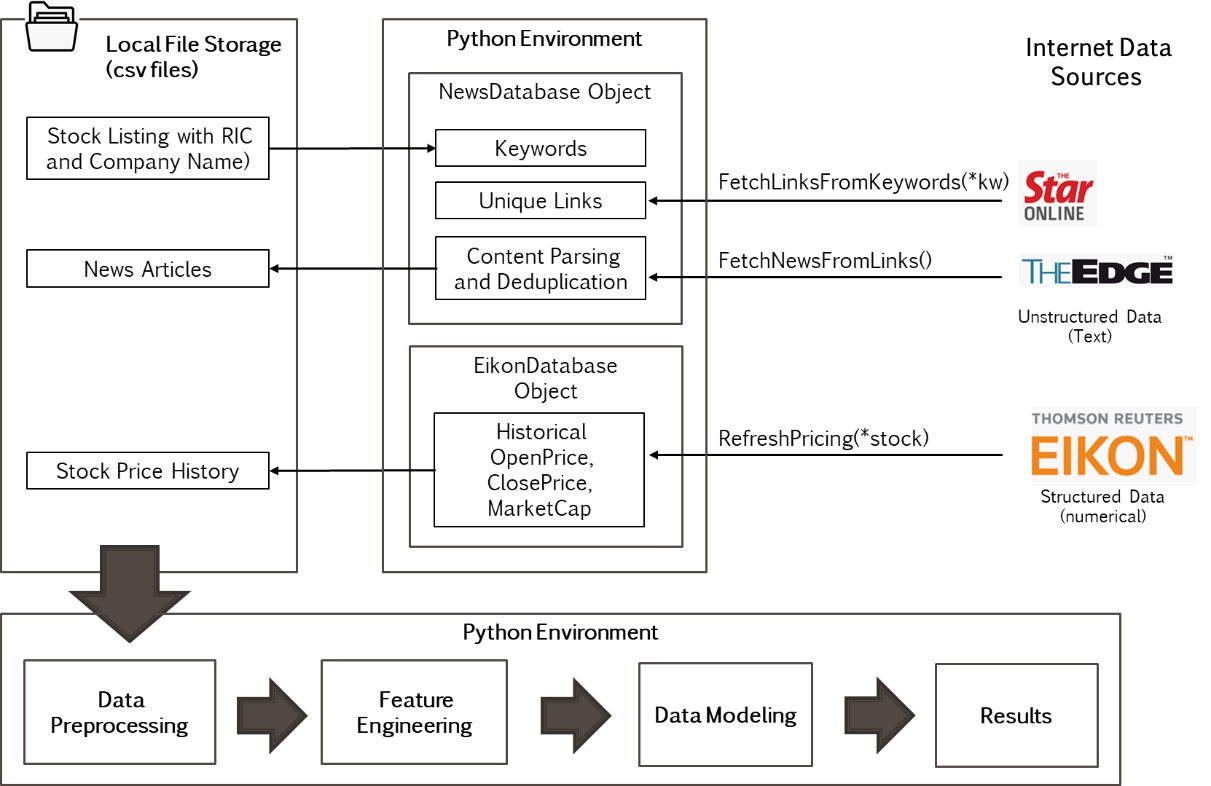


Figure .: Overall System Design

### 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.

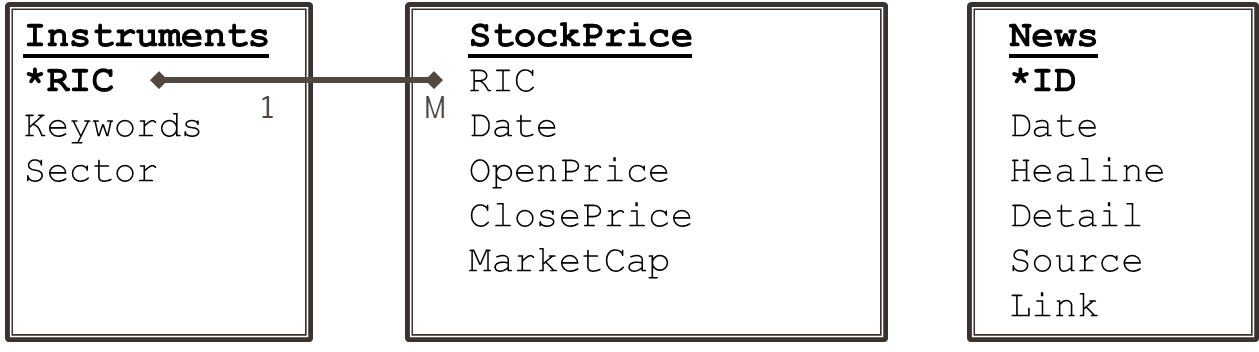


Figure .: Database Schema

* Instruments (Table 3.3)

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

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 weightage for sector index price.

* News

Key news data are the published date, news headline and its detail body content. Further stage of the data processing will derive news sentiment scores, a numeric variable to quantity polarity. Later part of this document will explain the method to relate news sentiment to stocks/sector index pricing.

## Data Processing

### Processing 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 below data point count (Figure 3.4), the quantity of data is all very good (around 2,400). The only exception is ASTR.KL and OCKG.KL due to their IPO date on year 2012. Data quality is revalidate by visually inspecting all price charts as in Figure 3.4, to ensure no obvious gaps and jumps on pricing.

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 needs treatment:

1. Publish timestamp relate – 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 may 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 dropped in this analysis to avoid skewing of result.

### Feature Engineering

#### Stock Price Return

One of the goals of this project is to evaluate near team returns for same day 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 uses fixed Open Price at day 0 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.

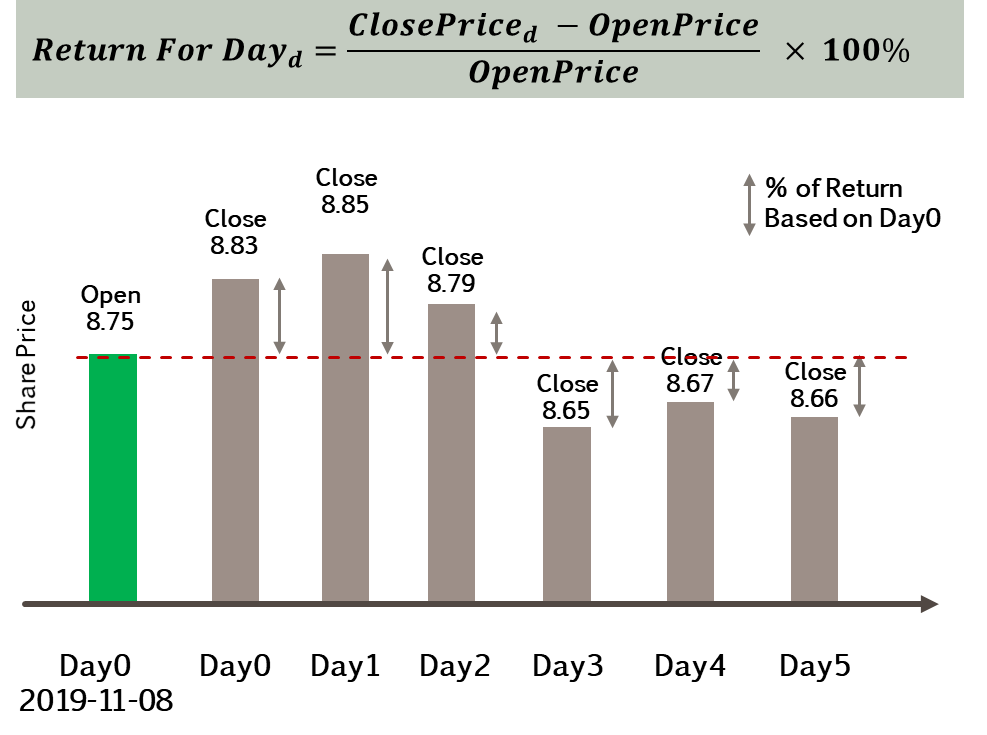


Figure .: MBBM.KL Stock Price Return Illustration

For computation this time series pricing data, intermediate dataframe is created by shifting up a day for five consecutive days in the future (upper part of Table 3.5). Final step to get returns simply perform 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). It also shows the effect of having blank values at the last few rows as an effect of shifting up the CLOSE PRICE. To avoid this, this research had obtained time series data till 2020-01-08 to ensure not missing gaps of returns value till 2019-12-31.

Table .: MBBM.KL Stock Price Return Example

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CURRENT** | | **FUTURE** | | | | |
| **MBBM.KL** | **OPEN PRICE** | **CLOSE PRICE** | **CLOSE PRICE** | | | | |
|  | **DAY 0** | **DAY 0** | **DAY 1** | **DAY 2** | **DAY 3** | **DAY 4** | **DAY 5** |
| 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 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **MBBM.KL** |  | **CURRENT RETURN** | **FUTURE RETURN** | | | | |
|  |  | **DAY 0** | **DAY 1** | **DAY 2** | **DAY 3** | **DAY 4** | **DAY 5** |
| 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 Price 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 method is a type of market index with individual components that are weighted according to their total market capitalization. (Chen, 2019).

Like 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 (**Error! Reference source not found.**).

Table .: Sector Index Price Calculation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **10 Sector Stocks (A to J)** | | | | **Sector Index**  **(OPEN or CLOSE PRICE)** |
| A | B | .. | J |
| Date | Open |  |  | .. |  |  |
| Close |  |  | .. |  |  |

Table .: Stock MarketCap Weighing

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **10 Sector Stocks (A to J)** | | | | **Total Market Capital** |
| A | B | .. | J |
| Date | Market Capital |  |  | .. |  |  |
| Weight |  |  | .. |  |  |

As an example, market capital for all Financial Services stocks within scope are calculated as below 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. The Index price for that day is 7.0256.

Table .: Market Capitalization Weighted Financial Services Index Price

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **2010-01-04** | | | |
| **Stocks** | **MarketCap** | **Weightage** | **Open Price** | **Weighted OpenPrice** |
| **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 calculated using the same weighting method above, procedures in Section 3.4.2.1 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 fundamental to understand the 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 use under computer program control. For each English word, it produces a set of synonyms (called Synset) based on its built-in dictionary and the part of speech tag of the word such as nouns, verbs, adjectives, and adverbs. Wordnet superficially resembles thesaurus in the aspect of providing synonyms, however, below are the few key distinctions (Priceton, 2020):

* WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated
* WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity

In terms of coverage, WordNet’s goal differs from good standard dictionary in that WordNet is based on the notion of word sense that lexicographers have traditionally used in writing dictionaries (Raj, 2013). Wordnet connects its large lexical databased of English words by interlinking words of similar conceptual-semantic from the same POS. There are various of semantic relation and their support for POS is described in Table 3.9 below.

Table .: Structure of WordNet (Keselj, 2009)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 version 3.0, which contains 147,306 words 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, it is clear that the same word can have different meanings depending on its Part of Speech. In WordNet, a basic construct of such is called a SyncSet, which is a tuple of (Word, POS, Word Sense). Word Sense is an 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 next section.

Table .: Various Meaning of The Same Word

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Various Synsets of ‘Bank’** | | |
| **SynSet** | **(‘Bank’, Noun, 1)** | **(‘Bank’, Noun, 2)** | **(‘Bank’, Verb, 1)** |
| **Definition** | Sloping land (especially the slope beside a body of water) | A financial institution that accepts deposits and channels the money into lending activities | Tip laterally |
| **Example** | They pulled the canoe up on the bank | He cashed a check at the bank | the pilot had to bank the aircraft |
| **Hypernym** | (‘Slope’, Noun, 1) | (‘Financial Institution, Noun, 1) | (‘Tip’, Verb, 1) |

1. Senti-Wordnet

SentiWordNet (SWN) is a lexical resource for opinion mining. It is a sentiment lexicon associating sentiment information to each WordNet Synset. In high level, SWN employs two learning algorithm (SVM, Rochhio) and four (4) training sets to yield eight (8) classifiers. As the result from the work in (Baccianella et al., 2010), SWN assigns to each synset of WordNet three sentiment is represented by three categories of scores: positivity, negativity and objectivity. The relationship among the three sentiment categories is illustrated in Figure 3.8 below. The sum of the all three sentiment scores equals to one.

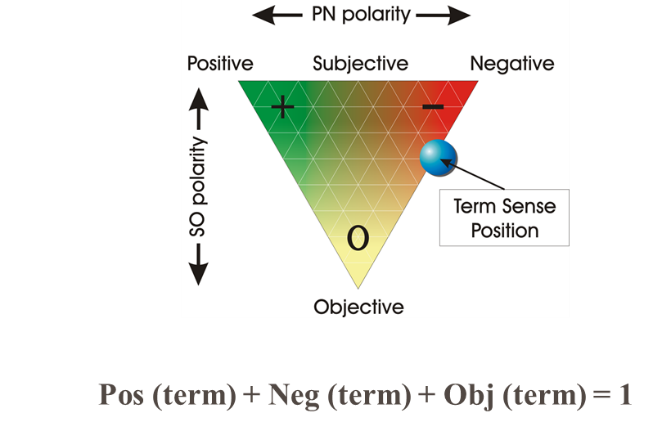


Figure .: Graphical Representation of Senti WordNet Opinion Related Properties (Esuli & Sebastiani, 2006)

Because SWN rides on Synsets of WordNet, they uses the same lexical resource which essentially provides a synset-based sentiment representation. That also means diﬀerent senses of the same term may have diﬀerent sentiment scores. As shown in Figure Table 3.11 below, different senses of word “Happy” results different sentiment associations. Ideally, SentiWordNet needs to be coupled with a Word Sense Disambiguation (WSD) algorithm to identify the most promising meaning. However, WSD is beyond the focus of this research, and this project always use first sense as the most common scenario.

Table .: SWN 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

In order to evaluate the sentiment of a news article, this research scores the news header and detail separately and later combined for an average score to represent the final sentiment of the article. There are two steps to evaluate the text sentiment scores as illustrated in Figure 3.9 below. Firstly count the frequency of positive and negative terms (determined using Senti-SynSet Pos, Neg scoring, Equation 3.1) , secondly use the Equation 3.2 to score the underlying text. This technique was also found in work (Hor et al., 2018; Li et al., 2014; Nassirtoussi et al., 2015). The value one (1) indicates extreme positive sentiment, whereas 0 represents extreme negative sentiment. Value of zero 0.5 represents neutral. See examples in Table 3.12 above.

In Equation 3.1, if more positive terms detected than negative ones, the term is labeled as Positive. On the reverse, it is Negative labeled. When there are equal number of positive and negative terms (equally weighted), the equation evaluates to 0.5 (neutral). Positive or negative terms may not present in short text (especially headers), hence Equation 3.2 adjusted the sentiment score to 0.5 (neutral) for this scenario. Figure 3.9 below gives a simplified view of the above-mentioned workflow.

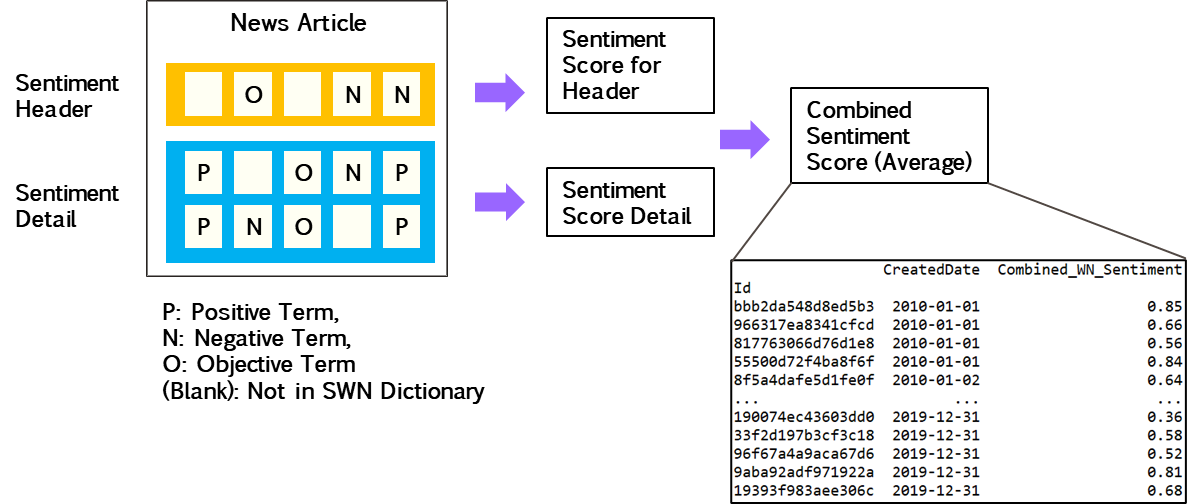


Figure .: Sentiment Scoring and Label For News Article

Equation .: Determine Sentiment Category For A Term

Equation .: Determine Sentiment Score For A News Headline or Detail

Few sample news articles and their sentiment score calculation are illustrated below.

Table .: News Sentiment Score (Positive Example 1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Article** | | **WN Positive Words Count** | **WN Negative Words Count** | **Sentiment Score** | |
| **ID** | af47b304b16e80a9 |  |  |  | |
| **Date (ID)** | 2019-11-28 |  |  |  | |
| **Headline** | Maybank gets BNM nod to raise paid-up by 539m new shares | 1  (new) | 0 | 1/1 = 1 | |
| **Headline** | KUALA LUMPUR (March 20): Malayan Banking Bhd had received Bank Negara Malaysia’s approval to increase its issued and paid-up share capital by up to 539.3 million new shares. Maybank said on Tuesday the new shares arose from the implementation of the dividend reinvestment plan. The recurrent and optional dividend reinvestment plan allows shareholders to reinvest the electable portion of their final cash dividend for the six-month financial period ended Dec 31, 2011 of 36 sen per share less 25% taxation into new shares. | 5  (had, new, new, financial, new) | 1  (less) | 5/6 = 0.8333 | |
| **Source** | TheEdge |  |  |  | |
|  | **Combined Sentiment Score (Average)** | | | | (1 + 0.8333)/2  = 0.92 | |

Table .: News Sentiment Score (Positive Example 2)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Article** | | **WN Positive Words Count** | **WN Negative Words Count** | **Sentiment Score** | |
| **ID** | af47b304b16e80a9 |  |  |  | |
| **Date (ID)** | 2010-08-07 |  |  |  | |
| **Headline** | Haisan gets letter on term-loan facility' | 1  (new) | 0 | 1/1 = 1 | |
| **Headline** | 'PETALING JAYA: HAISAN RESOURCES BHD told Bursa Malaysia yesterday its wholly owned subsidiary IGLO (Saigon) Inc has received a letter from MALAYAN BANKING BHD ’s Ho Chi Minh City branch demanding 44.34 billion dong (RM7.43mil), being the principal and interest due related to a term-loan facility. It said the company had appointed a scheme adviser, UHY Diong Advisory (KL) Sdn Bhd, to formulate a conclusive debt restructuring proposal. HAISAN : [Stock Watch][News] | 6  (wholly, has, Chi, being, had, scheme) | 2  (demanding, formulate) | 6/8 = 0.75 | |
| **Source** | TheStar |  |  |  | |
|  | **Combined Sentiment Score (Average)** | | | | (1 + 0.75)/2  = 0.88 | |

Table .: News Sentiment Score (Negative Example 1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Article** | | **WN Positive Words Count** | **WN Negative Words Count** | **Sentiment Score** | |
| **ID** | af47b304b16e80a9 |  |  |  | |
| **Date (ID)** | 2010-08-07 |  |  |  | |
| **Headline** | Microlink sees shareholding changes | 0 | 0 | 0.5 | |
| **Headline** | KUALA LUMPUR: ACE Market-listed MICROLINK SOLUTIONS BHD saw Teak Capital Sdn Bhd ceasing to be a substantial shareholder in the former after it sold about 11.44 million shares in a direct business transaction to Yong Kar Seng. The indirect shareholders of Microlink Solutions by virtue of indirect stakes in Teak Capital are Normah Raja Nong Chik, which is Yong’s spouse, MALAYAN BANKING BHD, Khazanah Nasional Bhd, Bank Perusahaan Kecil & Sederhana Malaysia Bhd and Central Paradigm Sdn Bhd. | 6  (wholly, has, Chi, being, had, scheme) | 2  (demanding, formulate) | 6/8 = 0.75 | |
| **Source** | TheStar |  |  |  | |
|  | **Combined Sentiment Score (Average)** | | | | (1 + 0.75)/2  = 0.88 | |

1. Text Processing Implementation

Every news article stored in News database (refer Section 3.3 above) is processed by computer program row by row to determine its sentiment score. The program workflow executes various calculation as defined in 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 cleans up null value with empty character replacement 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.

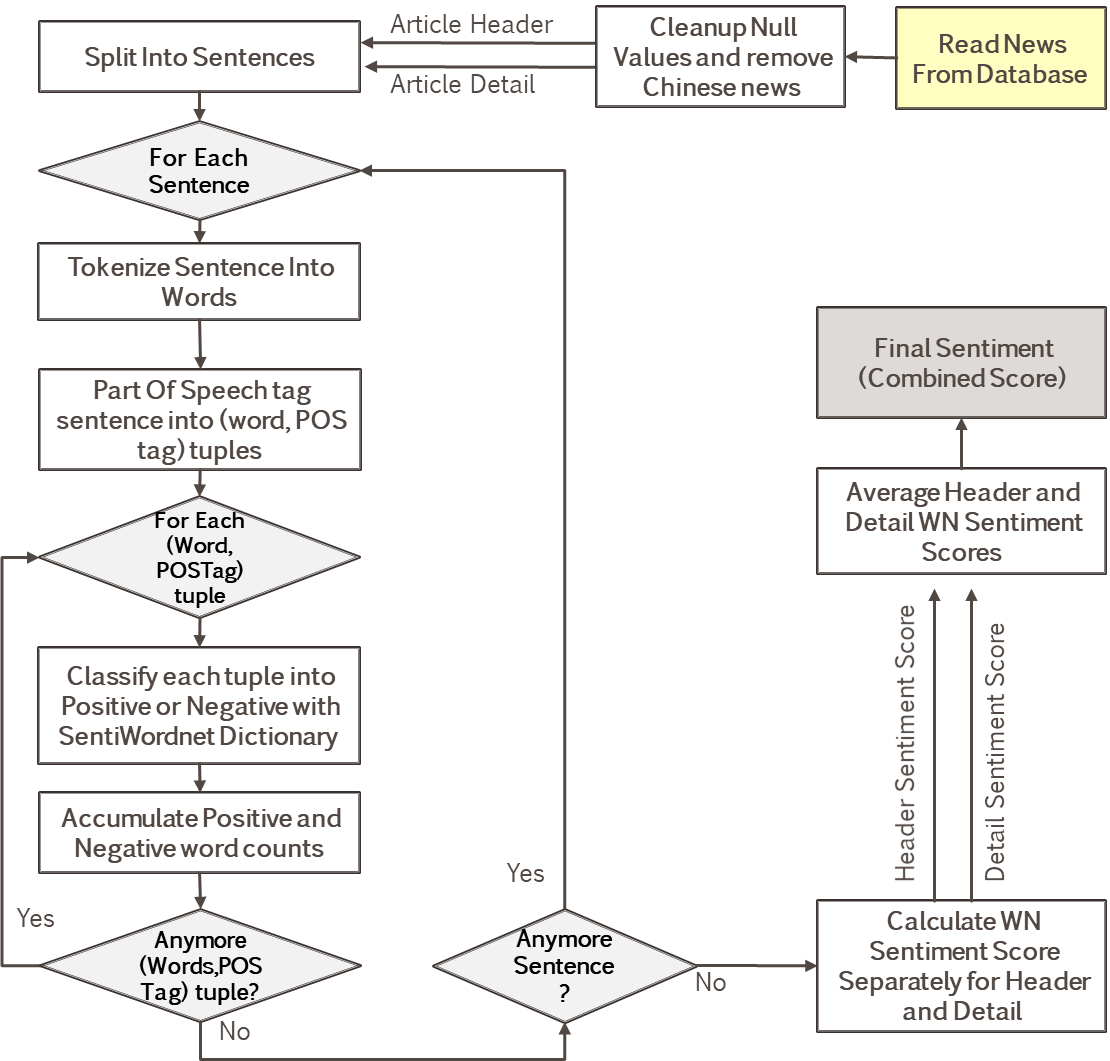


Figure .: Sentiment Score Processing Flowchart

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

As explained earlier, SWN requires that terms being POS-tagged. NLTK included pretrained POS tagger and Sentence Tokenizer. NLTK tokenizer split text into sentences so that words within sentences can be POS-tagged. The 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 3.15 below is necessary as part of the tagging process. Simple mapping rules that match first character of Penn tag is sufficient to this step.

Table .: Penn Treebank POS Tag Mapping to WordNet

|  |  |  |
| --- | --- | --- |
| **WordNet Tag** | **Penn Tag** | **Mapping Rules** |
| Noun (N) | Noun (NN)  Noun, plural (NNS)  Proper noun (NP)  Proper noun, plural (NPS) | Start with ‘N’ |
| Adverb (ADV) | Adverb (RB)  Adverb, comparative (RBR)  Adverb, superlative (RBS)  Particle (RP) | Start with ‘R’ |
| Verb (VERB) | Verb, base form (VB)  Verb, past tense (VBD)  Verb, present participle (VBG)  Verb, past participle (VBN)  Verb, non-3rd person singular present (VBP)  Verb, 3rd person singular present (VBZ) | Start with ‘V’ |
| Adjective (ADJ) | Adjective (JJ)  Adjective, comparative (JJR)  Adjective, superlative (JJS) | Starts with ‘J’ |

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

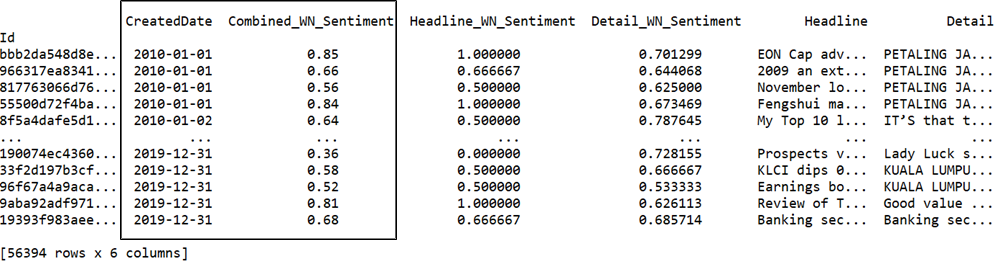


Figure .: Example Sentiment Scores DataFrame

### 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. It is the same technique applied for searching relevant content during web scrapping phase (refer to Section 3.2.1).

The relevancy mapping is represented in a DataFrame with RIC in columns and news id as row index. The cells with Boolean True indicates relevancy of news id to stock. The It is possible that news articles contain keywords from multiple stocks hence multiple True value in a single row is observed in such scenario. **Error! Reference source not found.** below shows a snapshot of news related to Financial Services sector stocks. The last news id (bottom row) is related to both PUBM.KL and HLBB.KL.

Sector news relevancy is represented with two new columns, namely ‘FSI\_INDEX’ and ‘TELCO\_INDEX’. Their value will be True if news is relevant to any of its constituents. In computer programs, this is simply derived from logical operator OR applied column wise across all underlying stocks. The final DataFrame representation has eleven (11) columns for each sector, ten (10) component stocks and one sector Index. This is convenient for data filtering and aggregation phase in the subsequent section.

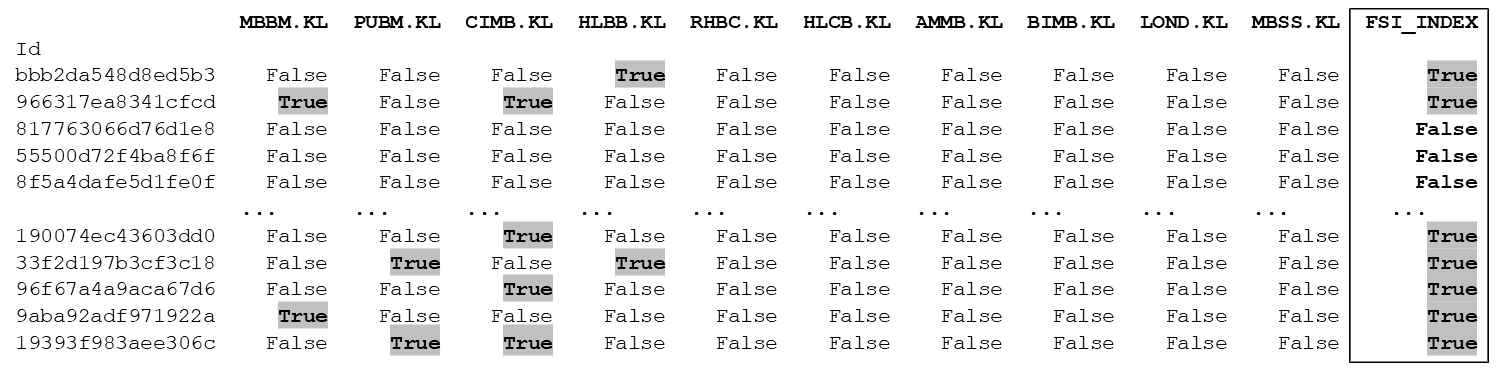


Figure .: Example News To Stock Entity Mapping DataFrame

### Data Aggregation

This stage simply joins data from News Entity DataFrame (output from Section 3.4.2.3) and News Sentiment DataFrame (output from Section 3.4.3) using news Id as the common key. The result of this join is a neat DataFrame that maps all news to stock entities (including Sector Index) together with its individual news sentiment. Shown in Figure 3.13 below is the snippet of this intermediate output filtered on Financial Services stocks.

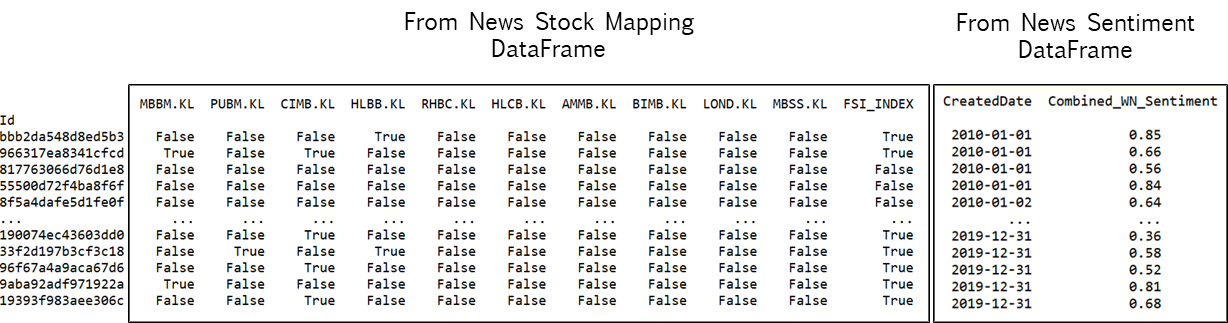


Figure .: News Sentiment To Entity DataFrame Example

Stock price analysis in this research focus on daily returns and modeled with news sentiment as predictor. Hence, news sentiment score data must be aligned to daily to match stock price data. This alignment only take place when a stock has multiple news within the same day. This 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 new unavailability on that day. At this point, news sentiment features are fully ready for modeling.

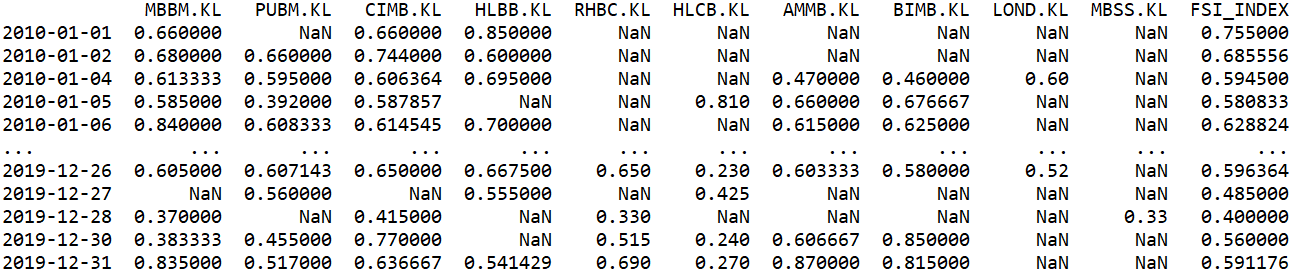


Figure .: News Sentiment To Entity DataFrame Example

### Returns Labelling

This research differs from majority work found in literature review by its categorization of price movement. Three categories of price direction, namely UP, DOWN and STAY is used as label in this research. The idea of “STAY” label is to mimic human decision as a reaction to external influence such as news sentiment. One obvious factor which may influence investor decision is the trading cost when buying or selling shares.

Three (3) types of trading cost involved for every transaction in KLSE according to (Bursa, 2020). These costs can be summarized below:

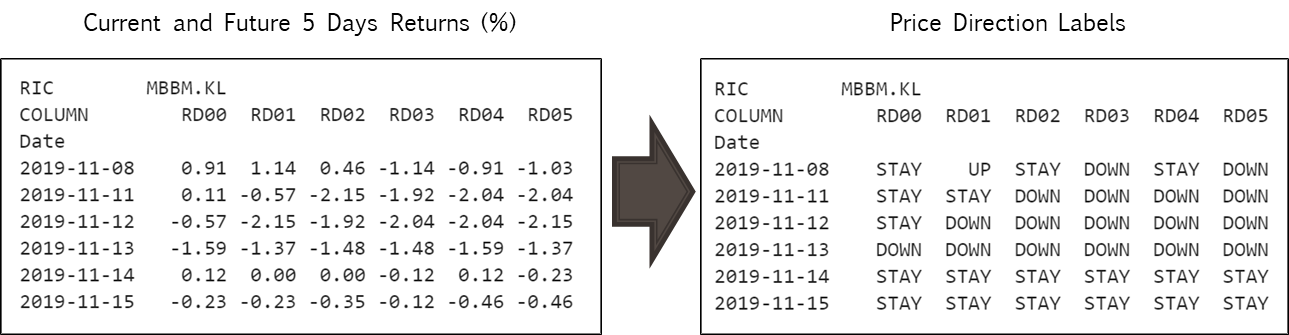
* Broker’s Commission - charges depending on brokers, generally 0.1% for cash and 0.3% for contra account. applicable to buyer and seller. Reference on various broker charges and criteria is available in this website (Investor.com, 2020).
* Stamp Duty - charged at RM1 for every Rm1000 transacted.
* Clearing Fees - charged at 0.03% and capped at RM1000 per transaction applicable to seller and buyer.

This research considered transaction cost of 0.5%, covering the cost for both way trades (buy and sell). This is based on thinking process of an investor who buy only if he believes the stock price will gain beyond this threshold (UP). Reversely, sell the stock if he thinks 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 (0.5%) is also justifiable based on analysis in Section 4.2.1, where most the average returns fell within this boundary. This also means returns the research is investigating on abnormal returns.

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 model building.

Equation .: Share Price Direction Labeling

Figure .: Share Price Labels Example



# Data Interpetation

## News

### Overall News

Total amount of 147,930 news were 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%), which is expected as TheEdge is a major business news provider in Malaysia.

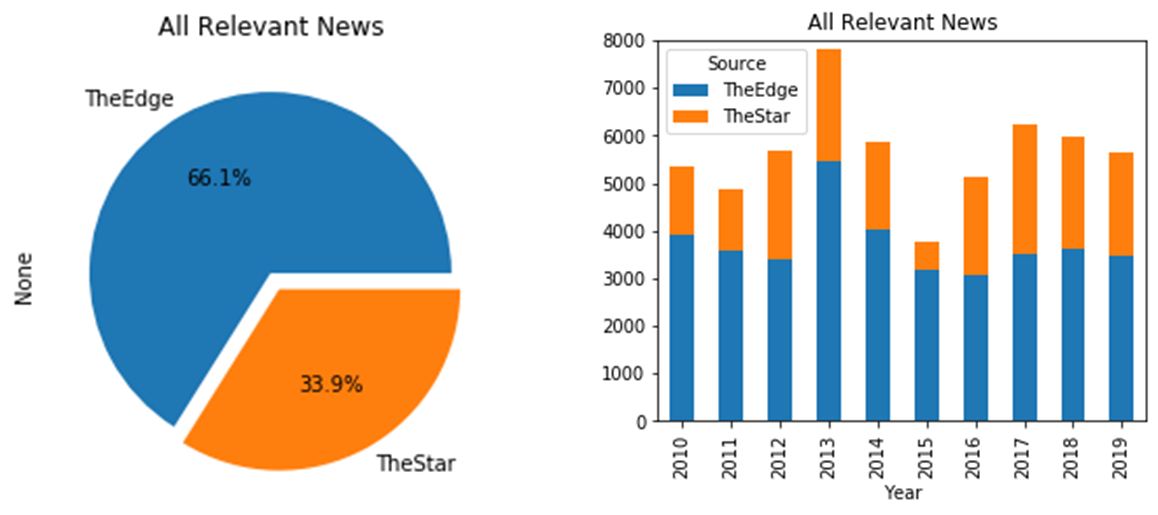


Figure .: Overall News By Sources

It is also found that generally the large capital stocks generate more interest to the news publisher. With exception of CIMB.KL and MBBM.KL. 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 most interest from publisher (22,715 news). Correlation between News Count and company Market Capital is illustrated in scatter plot below (Figure 4.2).

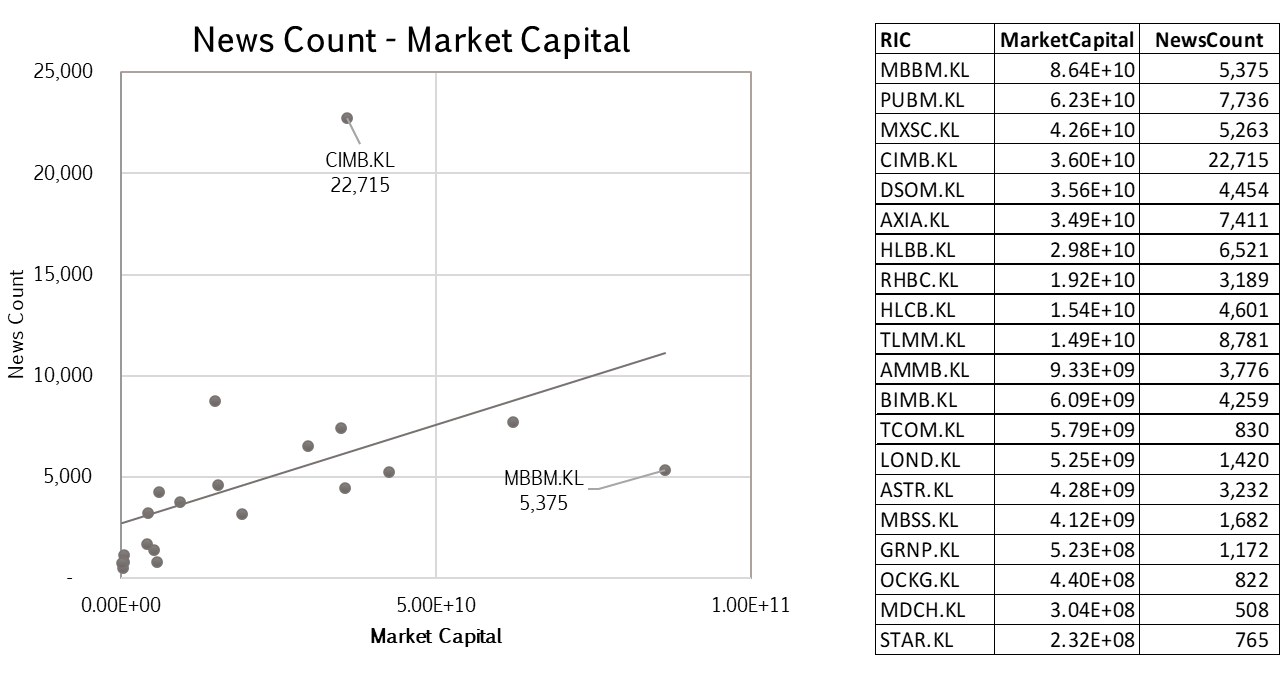


Figure .: News and Market Capital Scatter Plot

### 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, which also means 64% of entire news coverage in this database.

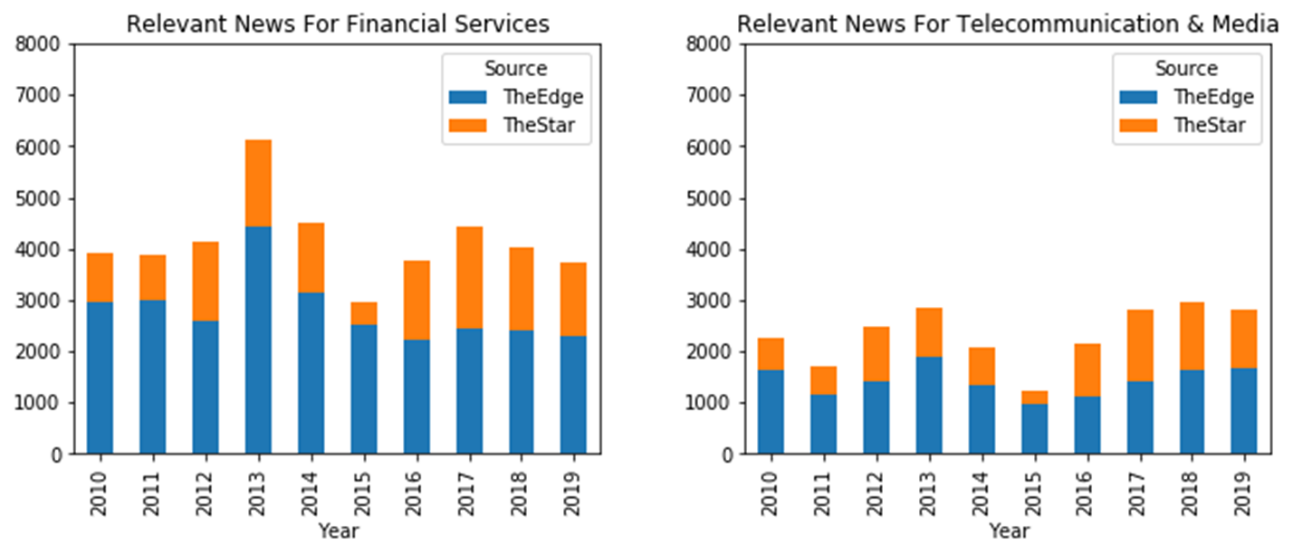


Figure .: Sector News

## Share Price

This section analyzes share price returns (Section 3.4.2.1) and its direction labels as produced in Section 3.4.5.

### Share Price Returns

Share price 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 of price returns of FS stocks trend from 0.02 in RD00 to 0.13 (RD05). This 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, Telco stocks are more volatile in short terms, as seen in the heteroscedastic plots of individual stocks performance in Figure 4.5. However, when combined all Telco stocks, its return has very similar trend to FS, which trends up from 0.06 (RD00) to 0.15 (RD05). 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.

For comparison purpose, below charts are arranged side by side with the same scale. On the left is Financial Services and on the right is Telecommunication & Media.

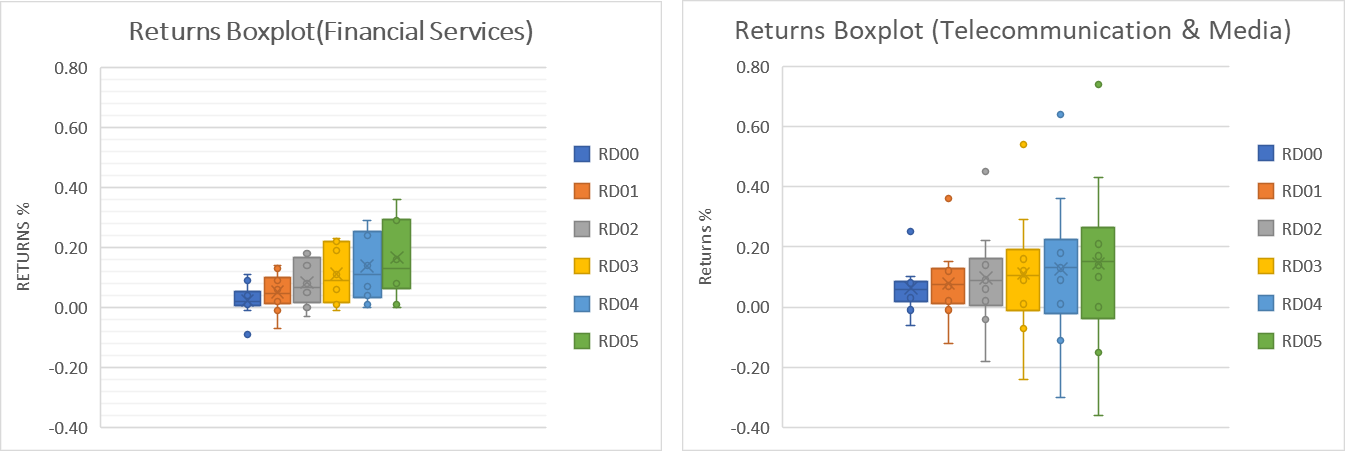


Figure .: Sector Price Returns Overall Chart

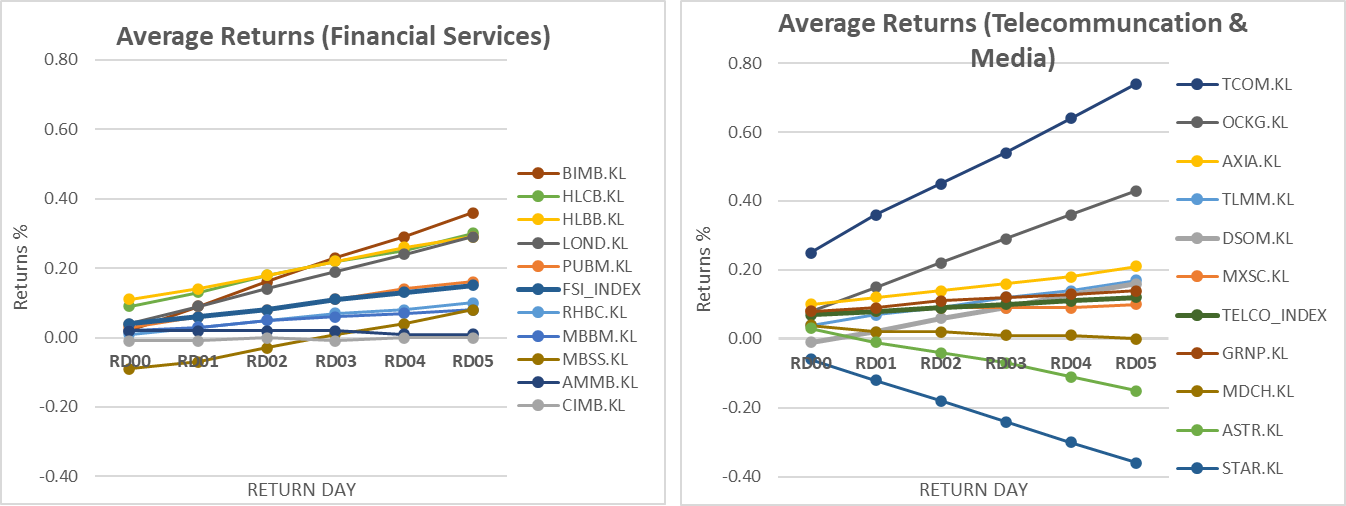


Figure .: Shares Price Returns Chart By Stock

### Returns Label

Share price direction labels are being analyzed in the next chart below (Figure 4.6). Distribution of labels (UP/STAY/DOWN) is visualize over the current to near term (RD00 to RD05). Recalled that the threshold separating upper and lower boundaries STAY label is at 0.5%.

Both FS and TELCO sectors shows similar pattern where majority of 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 boxplots and line charts above suggested up trends for price returns, the magnitude of hike is too subtle to push up the number of UP labels. This is shown in Figure 4.6 where the shift is well balanced over time in both UP/DOWN direction, though slightly upwards for TELCO.

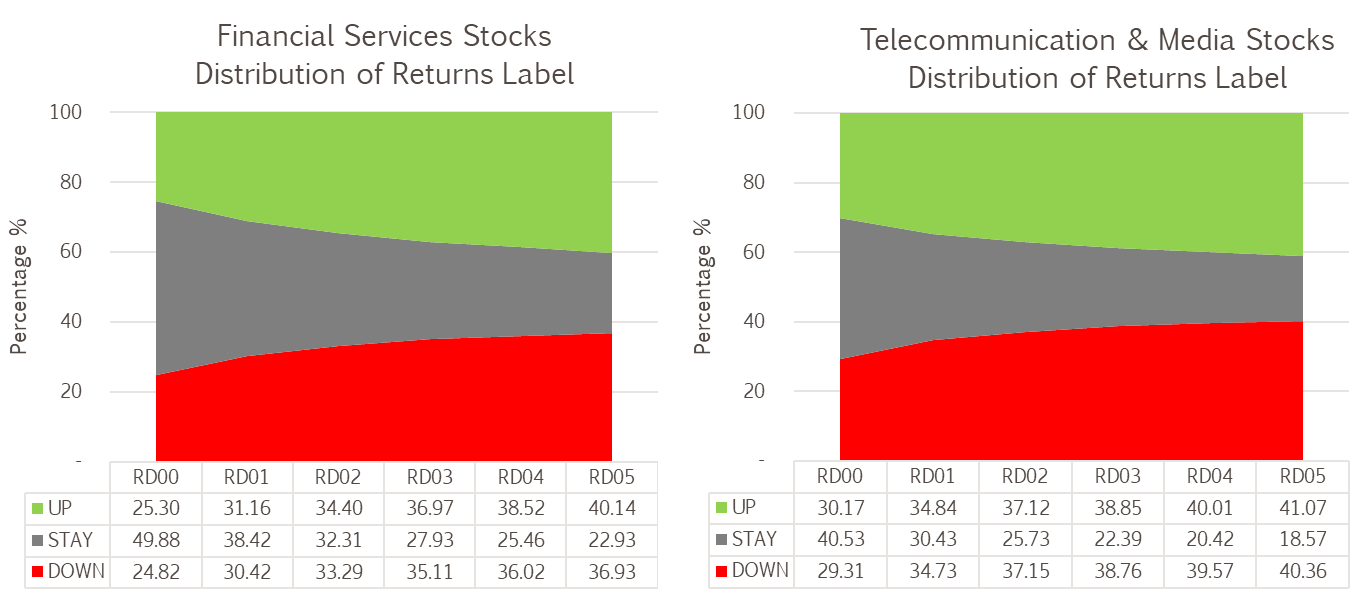


Figure .: Returns Label Distribution For Stocks

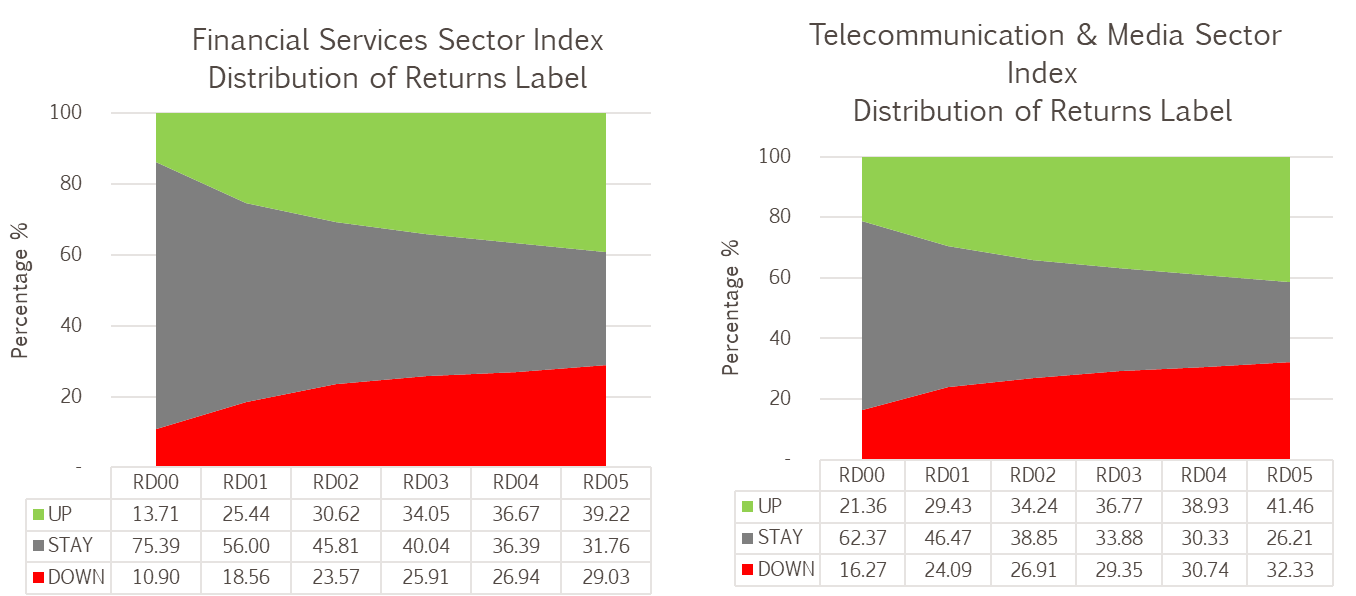


Figure .: Returns Label Distribution For Sector Index

# Data Modeling

### The Model

Modeling of sentiment scores to predict current and near-term returns involved multiple steps as depicted in Figure 5.1 below. Sentiment scores generated 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 boundary mapped to STAY. A grid search is performed on the centre and threshold to discover the parameters that optimizes different objectives.

This result from the mapping is used for prediction. The model’s prediction is evaluated against actual labels for current and next five (5) days – RD00 to RD05. The evaluation used three (3) categories confusion matrix to measure performance of accuracy, precision, recall and F1 means (see Section 5.1.2 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 Section 5.1.3 later. The process discussed here is repeated for various group of stocks and sector index within the scope of this research.

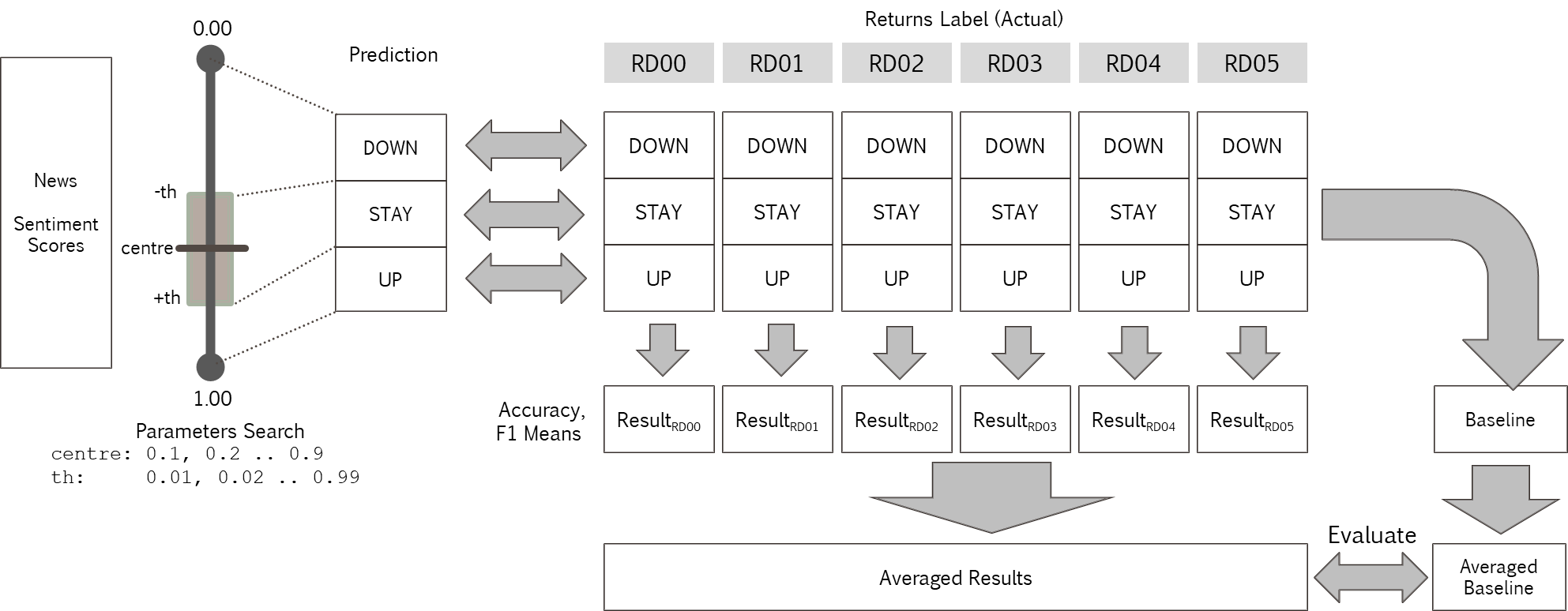


Figure .: Prediction Model

### Performance Metrics

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

* True Positive (TP) – Correctly predicted
* False Positive (FP) – Incorrectly predicted
* False Negative (FN) – Was not predicted

When the confusion matrix is read in a columnar manner calculation (illustrated in upper part of Figure 5.2), TP and FP values are used to calculate precision. 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 evaluates the entire confusion table. It tells overall how many observations were correctly predicted. Accuracy may not be the best or only performance metric when there are overly too many TP in the uninteresting category to investor, such as STAY. A high accuracy due to STAY does not motivate investor to trade. A more useful evaluation will be 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.

In short, Accuracy and F1 means are two key performance metrics evaluated in this research.

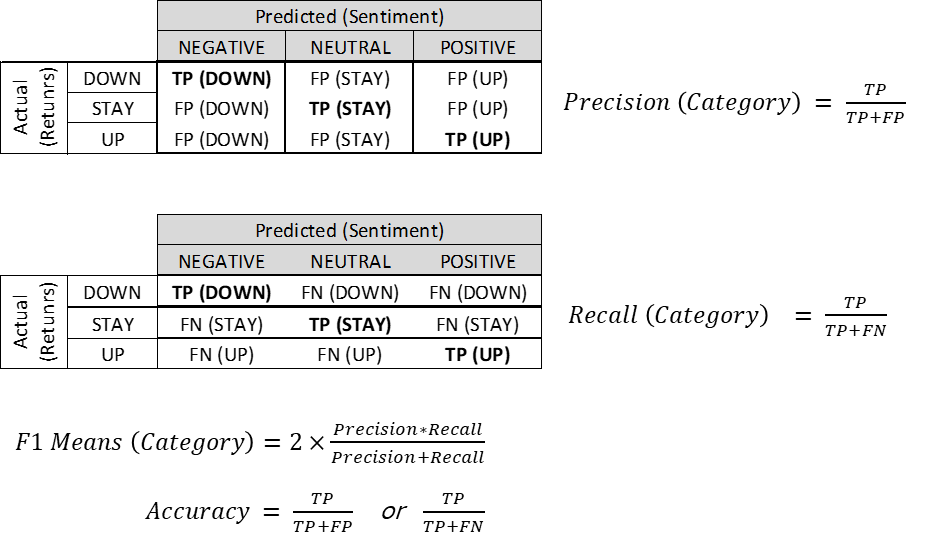


Figure .: Three Categories Confusion Matrix

### Performance Baseline

Baseline is derived from existing labels by naive human guessing. Two metrics are used as the overall measurement for model fitness, namely accuracy and F1 means. As there are multiple sets of labels (RD00 to RD05), the average metrics are used as baselines.

Naïve human guessing simply predicts label 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. Accuracies derived from this approach across RD00 till RD05 are averaged to get the final accuracy baseline.

Average of probability on each category signifies the baseline F1 means. It is from the same idea that if one placed ‘all-in’ to a specific label, precision and recall for that label will be equivalent to the probability of label. Hence, value of F1 means will be the same as the probability. Each category will have a baseline F1 means.

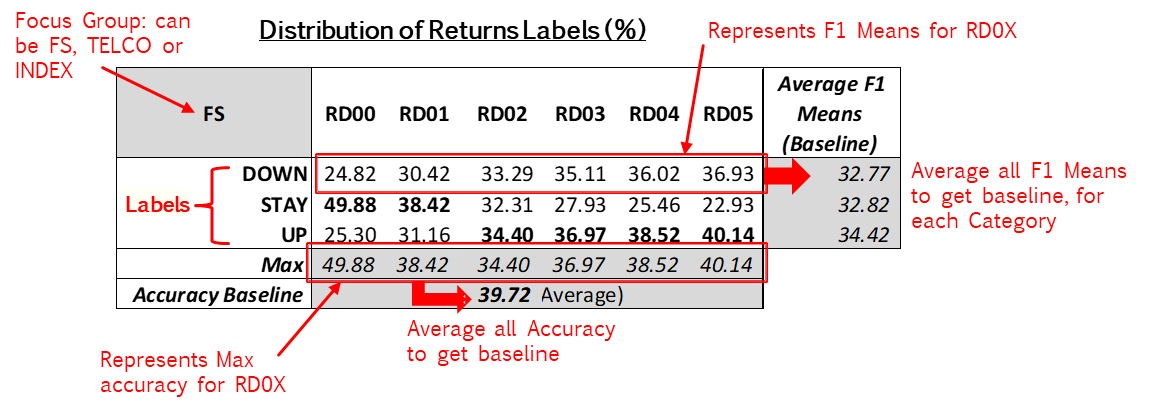


Figure .: Baselining Method

Using the baselining method discussed above, four (4) groups of baselines have been calculated (Table 5.1 below). This is 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 process discussed previously. Baseline values from are summarized below with side by side comparison with model performance, together with the parameters that optimized the metrics.

On accuracy, the model performs better than baseline only in TELCO group at 47.20%. Accuracy prediction for all groups are below 50%.

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 allows more predictions for UP labels where probability of this label increased from RD00 to RD05.

On the other hand, the optimum center parameter is lower for F1 means prediction on DOWN category (between 0.80 and 0.90). For the same reason as UP labels, probability for DOWN label increased from RD00 to RD05 too. F1 means prediction for DOWN performs 5% to 33% better than baseline.

As for STAY, only FS INDEX group did not outperform baseline, with F1 means 8% lower than baseline. The other groups outperform baseline within range of 16% to 36%. As STAY is the zone of ‘neutral’ opinion, its sentiment centers is quite balanced with 0.5 and 0.6.

The result is summarized with table and chart in Table 6.1 below. The optimum parameters in the table is interpreted with “C” for center value and “T” for threshold.

Table .: Model Performance Per Group

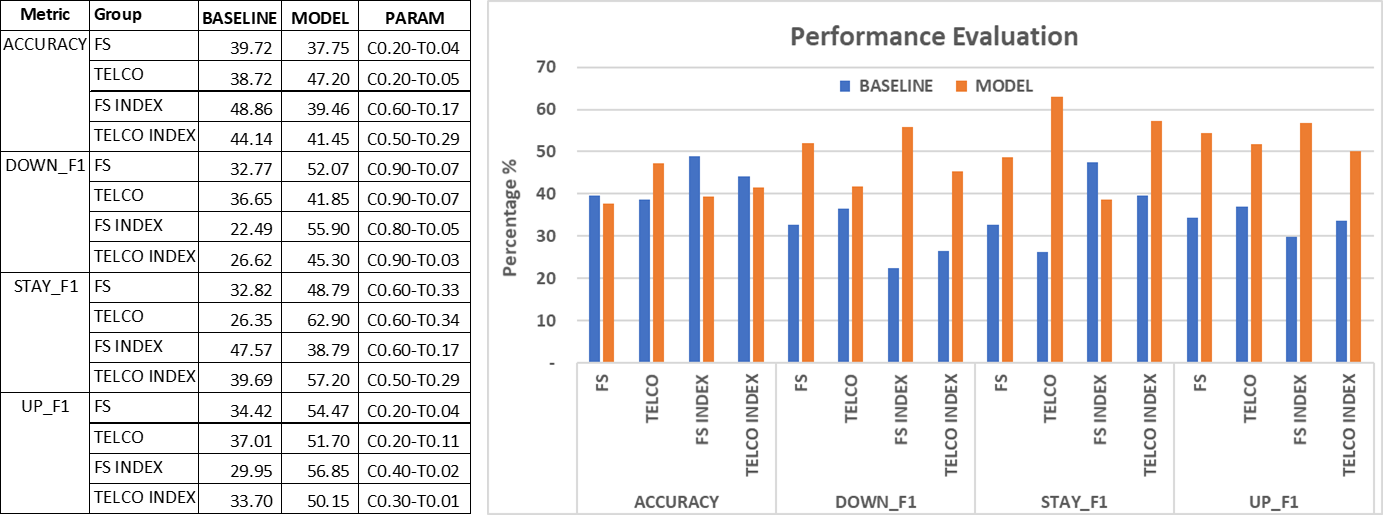


Table 6.2: Model Performance Per Stock

Table 6.2 below broaden the evaluation by individual stock basis. The average model accuracy is less than 1% below baseline while F1 means outperform baseline by 15% to 19%.



# CONCLUSION

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

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 ability of the model producing within 38% to 62% in F1 means across all categories indicates that news sentiment indeed related to short term stock price direction. In short, results obtained in this research opposed to randomness and proved the relation between news sentiment to stock price direction.

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 baseline ranged from -8% to 33.41% in F1 means. The variability is higher compared to stocks group. Sector news for Financial Services performs -8% weaker than baseline on STAY category. As for accuracy, both sector Index performed weaker than baseline. In summary, the impact of sector news to index direction is not as strong as stocks groups.

# Future Works

# 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.

Bursa, M. (2020). Transaction Costs. Retrieved from <https://www.bursamalaysia.com/trade/trading_resources/equities/transaction_costs>

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.

Chen, J. (2019). Capitalization Weighted Index. Retrieved from <https://www.investopedia.com/terms/c/capitalizationweightedindex.asp>

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.

Fast, G. (1999). Lingua: EN: Syllable Perl module for estimating syllable count in words. Available at CPAN (1999). In.

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.

Investor.com. (2020). Broker Comparison. Retrieved from <https://klse.i3investor.com/jsp/hti/brokers.jsp>

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

Keselj, V. (2009). Speech and Language Processing Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6, $115.00. In: MIT Press.

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.

Priceton, U. (2020). What Is WordNet. Retrieved from <https://wordnet.princeton.edu/>

Raj, G. (2013). WordNet A Lexical Knowledgebase. Retrieved from <https://www.slideshare.net/govindraj15/wordnet-27805299/>

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.

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., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Tan, L. I., Phang, W. S., Chin, K. O., Alfred, R., & Anthony, P. (2014). Impact of financial news headline and content to market sentiment.

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.

Appendix A: News Database Code

NewsDatabase