Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm

Shuhaida Mohamed Shuhidan[1], Saidatul Rahah Hamidi [2], Soheil Kazemian[3], Shamila Mohamed Shuhidan[4] and Maizatul Akmar Ismail[5]

[1,3] Accounting Research Institute (ARI), [1,2] Faculty of Computer and Mathematical Sciences, [4] Faculty of Information Management,
Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia.
[5] Department of Information Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
[1]shuhaida@tmsk.uitm.edu.my, [2]rahah@tmsk.uitm.edu.my,
[3]soheilkazemian@salam.uitm.edu.my, [4]shamila@salam.uitm.edu.my,
[5]maizatul@um.edu.my

Abstract. The study covers the implementation of machine learning algorithm approaches in sentiment analysis of Malaysia financial news headlines. This study can be used for stakeholders who want to know about the financial news and seek knowledge or data in the financial world. The data are gained from Malaysia online financial news, which are from Business section of New Straits Times. Our study applies Opinion Lexicon-based algorithm and Naïve Bayes algorithm as the method to perform sentiment analysis. This study consists of several phases in pre-processing such as extract data, stop word removal, and stemming to clean the dataset and make it as data preparation before performing the sentiment analysis with the selected machine learning algorithms. In the stop word removal, tm package in R is used to clean the dataset while for stemming process, Snowball stemmer is used to set the data to its root word. Sample outcomes of analysis are explained for both algorithms. The conclusion describes the summation of the study and future works.

Keywords: Data analytics, Sentiment Analysis, Financial headlines, opinion-lexicon based algorithm.

1 Introduction

Sentiment is a view or attitude toward a situation or event that can also be called an opinion. Sentiment analysis, known as opinion mining, is the field of the study regarding people's views, sentiments, evaluations, attitudes, and emotion from text [1]. Sentiment analysis is the task to recognize the text's polarity; resulted in either positive, negative, or neutral. The text may be in the form of electronic data such as reviews, messages, or comments.

Life choices are mostly conditioned on how others see and evaluate the world. For this reason, when individuals need to make a decision, they often seek out the opinion of others. This is not only for individuals but also for organization. Our work aims to extract and analyse the data of Malaysia financial news headlines.

A. M. Lokman et al. (Eds.): KEER 2018, AISC 739, pp. 64–72, 2018.
https://doi.org/10.1007/978-981-10-8612-0_8

Financial headline often shares the price movements which show what is trending in the economy. Negative headlines may effect share price and it also can hamper the firm's capability to raise finance on the stock market.

2 Research Work

2.1 Text Mining

Text mining can be defined as looking for patterns in a text that contains high quality information which refers to innovation, importance, and interestingness of the way to write the text [2]. Basically, text mining is a complex task where computer algorithms are used to process text and to derive meaning or pattern from the text. The use of text mining has been broadly deliberate in the area of financial markets. Nevertheless, it is hard for a computer to recognise and it needs some intelligent methods to make the information easily understood. There is a need of a method to structure input text, add linguistic features, and remove other linguistic features and store the text into the database.

2.2 Sentiment Analysis

Sentiment is a view or attitude towards an event or a situation which is also called an opinion. Sentiment analysis on online review has become a hot topic in the field of data mining consistent with the sheer volume of rich web resources such as digital newspaper, Facebook, Twitter, and e-forum. Sentiment analysis, also identified as opinion mining, is a *Natural Language Processing* task directed at the automatic identification and analysis of people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes in text [1, 3]. It allows users to recognize the emotional state of the author in writing, and the proposed expressive emotions that influence what the author desires to offer to the reader.

In life, principles and views of reality and the choice that they make are mostly based on how peoples see and evaluate the world. This is not only for individuals, in fact, it also apply to an organizations. In this modern technological era, the world that people lives in, they can use data technologies to know the opinions of others toward specific issues. Therefore, they can make better judgment and decision based on the sentiment.

In financial economics, it was used differently, as sentiment analysis refers to the derivation of market confidence indicators from proxies (i.e. stocks prices or trading volumes) [4]. Sentiment analysis as a whole helps a business observe public moods to achieve their goals either in terms of political movement, financial market, measure of customer satisfaction, and many other factors that affect their general emotions [5].

Due to social media, the development of sentiment analysis has become widely used and implementations of applications have increased, as there is a huge

growth of information available on the Internet [3]. People become eager to share their opinion in the web in regards to their daily life or global issues, which provides a transparent view of the world-wide-web. Any information acquired from Internet or online activities lead to huge data obtained for businesses to analyse and transform it into viable information. As such, it is prevalent for the business and service industry, which needs customer express and opinion on their product or services [5]. Social media real-time evaluations of business performance also allow investors to predict their business value [6]. However, currently, many research are rooted in examining the relationship of financial news articles to the stock market behaviour [7].

There is a study that shows that any aspect of news affects the market in many ways, either impacting on stocks, volatility, or future earnings of a firm [4].

2.3 Opinion Lexicon-Based Algorithm

Sentiment analysis can be performed using opinion lexicon-based approach [8]. This opinion lexicon algorithm approach depends on opinion words which are words that express the polarity of the sentiments. Positive opinion words are used to express some desired states while negative opinion words are used to express some undesired states. Examples of positive words: good, amazing, beautiful, and success. Examples of negative opinion words are ugly, worse, bad, and lose. In addition, there are also opinion phrases and idioms such as cost someone an arm and a leg. Collectively, they are called the opinion lexicon.

There are three main approaches that have been investigated; 1. Manual approach which is very time-consuming and thus it is not usually used alone but combined with automated approaches as the final check because automated methods make mistakes; 2. Dictionary based approach is based on bootstrapping using a small set of seed opinion words and an online dictionary such as thesaurus or *WordNet*. This method first collects a small set of opinion words manually with known orientations, and then grows the set by searching in the *WordNet* or thesaurus for their synonyms or antonyms. The iterative process stops when no more new words are found. This method has a major shortcoming. The approach is unable to find opinion words with domain and context specific orientations; and 3. Corpus based approach rely on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus [1].

The method is not effective as the dictionary based approach because it is hard to prepare a huge corpus to cover all English words. However, this approach has a major advantage that the dictionary based approach does not have. It can help find domain context specific opinion words and their orientations using a domain corpus. This algorithm requires a scoring function to score every sentence by counting the *positive* and *negative* occurrences in a statement. Fig. 1 shows the lexicon based sentiment analysis method. The lexicon based method uses a lexicon, a set of positive and negative words, combined with a scoring function to determine the sentiment analysis.

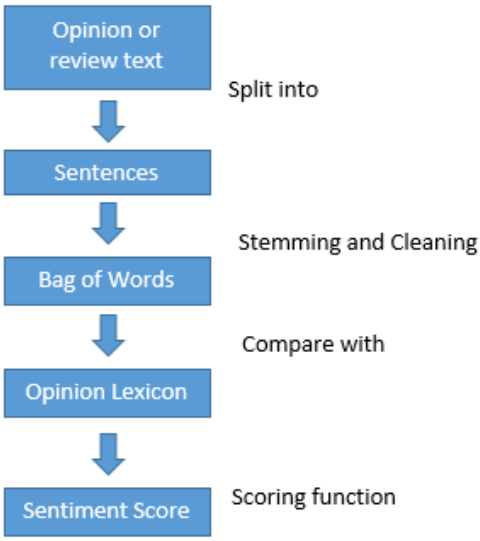

Fig. 1. Lexicon-based sentiment analysis approach

2.4 Naïve Bayes Algorithm

The Naïve Bayes algorithm is a set of supervised learning algorithm that widely uses algorithms for document classification. This algorithm is based on the Bayes theorem. Naïve Bayes learners and classifiers can be extremely fast compared to more sophisticated methods [9]. This algorithm is based on the bag-of-words model which the word of the text document appears in a positive-words-list or a negative-words-list. According to [10], there are two commonly used models for text categorization, which are:

Multinomial model
The multinomial classifier is suitable for classification with discrete features. The multinomial distribution normally requires integer feature counts. This classifier works with occurrence counts

Multivariate Bernoulli model
This classifier is also suitable for discrete data. This classifier is designed for binary/Boolean features.

3 Sentiment Analysis Approach

3.1 Data Collection

In order to sentiment the newspaper headlines (refer Fig. 2), the headlines must first be extracted from the online newspaper websites (link: https://www.nst.com.my/business). Headlines were extracted from New Straits Times web pages using R Programming and a tool called *SelectorGadget*. There were 536 financial headlines were scraped from News Strait Times newspaper online in the span of 12 months (January 2017 to December 2017).

Fig. 2. Data Collection Process

The data extracted will be stored into a *Comma-Separated Values* (CSV) spreadsheet file. Main reason for doing this is because it is faster to handle, smaller in size, easy to generate, and simple to implement. Fig. 3 shows a sample of the data extracted.

	A	B	C	D	E	F	G	H
1	Date	Headlines						
2	September 29, 2016	DRB-HICOM looking at 5 proposals for Proton partnership						
3	September 29, 2016	Better oil prices lift ringgit higher against US dollar						
4	September 28, 2016	Canada approves $8 bln LNG plant, but Petronas to review project						
5	September 28, 2016	Bursa Msia signs licensing agreement with KL Tin Market						
6	September 28, 2016	BCM Alliance aims to raise RM16.01m from IPO						
7	September 28, 2016	Malaysia's ranking drops to 25: WEF						
8	September 28, 2016	Need to address non-tariff barriers within Asean						
9	September 28, 2016	Satellite Pipeline inks technical deal with AWS Schäfer						
10	September 28, 2016	Citizens, govt and businesses play role in city's transformation						

Fig. 3. Example of extracted data from New Straits Times

3.2 Pre-processing Data

Next, the dataset will undergo the pre-processing process. Fig. 4 shows the pre-processing steps including stop words removal and stemming of the dataset.

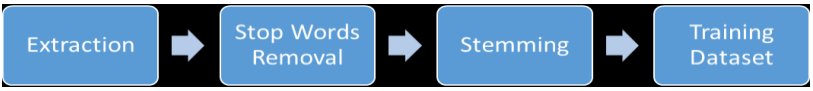

Fig. 4. Pre-processing step

Firstly, the data will be processed for stop words removal (refer Table 1).The next step is to perform word stem. Through this stemming process, the word will be stemmed to its root word. This project applied Snowball Stemmer of R packages, to stem the words.

Table 1. Sample data pre-processing (stop word removal and stemming)

Original Data	After Performing Stop Word Removal	After Performing Stemmer
DRB-HICOM looking at 5 proposals for Proton partnership	drbhicom looking proposals proton partnership	drbhicom look propose proton partner
Bursa Msia signs licensing agreement with KL Tin Market	bursa msia signs licensing agreement kl tin market	bursa msia sign license agree kl tin market
'No changes to GST in near term'	no changes gst near term	no change gst near term

3.3 Machine Learning Algorithm

The dataset will perform the sentiment analysis with the selected machine learning algorithms, which is Opinion Lexicon-based algorithm and Naïve Bayes algorithm.

Opinion Lexicon-Based Algorithm

The output obtained after the pre-processing phase can also be called the bag of words representation of the financial news headlines after applying the text mining. This algorithm compared each word in the sentence to the list and counted positive words with positive scores, and negative words with negative scores. In order to perform the sentiment, this algorithm needs the opinion lexicon and a scoring function, to assign the scores to financial news headlines. Hu and Liu published an “opinion lexicon” which contains a list of English positive and negative opinion words which around 6800 words separately. One of the advantages of using R is the ability to load and scan data. The positive and negative words are store in the project file and imported using scan function. In order to count the scoring sentiment, the ‘Plyr’ package is used to split data sets into smaller subsets and apply methods to parted subsets and

compare the words to the dictionaries of positive and negative words. Then combine back the results after the word was matched to the dictionaries.

	score	text
1	1	Better oil prices lift ringgit higher against US dollar
2	0	Bursa Msia signs licensing agreement with KL Tin Market
3	0	Malaysia on track to be high-income economy
4	0	KWAP crosses out Airbnb potential
5	-1	AirAsia wants govt to scrap planned airport tax increase

Fig. 5. Sample result of score function

Fig. 4 shows the result of sentiment analysis on financial news headlines after performing the score function. The result of scoring function:

- If the score is '2' means it highly positive
- If the score is '0' means it's a neutral sentence
- If the score is '-1' means it has no impact or negative sentence

Naïve Bayes Algorithm

The second approach applied to the pre-processed data is using the Naïve Bayes algorithm. Before performing the Naïve Bayes algorithm, the data must be classified into positive data and negative data. The data is converted into a set of frequency table and then a table was created by finding the probabilities. Next, the classified of positive and negative data will be load up. Then, the bags of words are loaded with its scores and categorize the words as very negative to very positive. In order to sentiment the sentences, calculate the number of words in each category within a sentence. The sentences are split up using *str_split* function. Next, build vector with matches between sentence and each category. The numbers of word in each category is summed up and later add a new row of score to the score table. Then, the tables of positive and negative sentences with scores and combine the positive and negative table are created. Finally, the data with the highest probability is the outcome of the prediction The sentiments are declared positive or negative. Fig. 6 shows the sample of result sentiment analysis of classified positive data and Fig. 7 shows the sample of result sentiment analysis of classified negative data.

	sentence	vNeg	neg	pos	vPos	sentiment
1	DRB-HICOM looking at 5 proposals for Proton partner...	0	0	0	0	positive
2	Better oil prices lift ringgit higher against US dollar	0	0	2	0	positive
3	Bursa Msia signs licensing agreement with KL Tin Mar...	0	0	1	0	positive
4	'No changes to GST in near term' .	0	1	0	0	positive
5	Malaysia on track to be high-income economy . Cana...	0	0	1	0	positive
6	Malaysia's innovation landscape, according to GE's in...	0	0	2	0	positive

Fig. 6. Sample result sentiment analysis of classified positive data

	sentence	vNeg	neg	pos	vPos	sentiment
1	Malaysia's ranking drops to 25: WEF . KWAP crosses o...	0	0	0	0	negative
2	Ringgit opens lower on weak oil prices .	0	1	0	0	negative
3	Less proactive new US administration to negatively i...	0	4	1	0	negative
4	Ringgit extends downtrend against greenback .	0	0	1	0	negative
5	Ringgit expected to trend lower against USD next w...	0	0	0	0	negative

Fig. 7. Sample result sentiment analysis of classified negative data

4 Conclusion

The study explains in detail the stages to conduct sentiment analysis. The datasets were scraped from the News Strait Times, Business section, for the duration of 12 months. The datasets undergo the pre-processing stages and then were applied to two machine learning algorithm; Opinion Lexicon-based algorithm and Naïve Bayes algorithm.

For future works, the researchers aim to study the datasets in depth and compare with other machine learning algorithm. Also, the researchers aim to develop a dashboard to display the results of the sentiment analysis to users.

5 Acknowledgements

The authors gratefully acknowledge the financial grant Accounting Research Institute (ARI), Universiti Teknologi MARA, Malaysia and Faculty of Computer and Mathematical Sciences for all supports and resources. Authors also would like to thank the research assistant, Mohd Fitri b Ali for the kind assistance. The authors also gratefully acknowledge the financial grant 600-IRMI/ Dana KCM 5/3/Lestari (156/2017) given by IRMI, Universiti Teknologi MARA, Malaysia and Faculty of Computer and Mathematical Sciences for all supports and resources.

References

1. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1): 1-167 (2012).
2. Puteh , M., Isa, N., Puteh, S., Redzuan, N.A..: Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. Proceedings of the World Congress on Engineering, vol.3 (2013).
3. Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., Herrera-Viedma, E.: Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38 (2015).
4. Devitt, A., Ahmad, K.: Sentiment polarity identification in financial news: A cohesion-based approach (2007).
5. Ravi, K., Ravi, V. A.: Survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46 (2015).
6. Yu, Y., Duan, W., Cao, Q.: The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems, 55(4), 919-926 (2013).

7. Tan, L. I., Phang, W. S., Chin, K. O., Anthony, P.: Rule-Based Sentiment Analysis for Financial News. In 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1601-1606, IEEE (2015).
8. Younis, E. M.: "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study." International Journal of Computer Applications 112(5) (2015).
9. Zhang, H.: The optimality of naive Bayes. AA 1, no. 2 vol 3. (2004).
10. Tan, S., Xueqi, C., Yuefen, W., and Hongbo, X.: Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. Advances in Information Retrieval, 337-349 (2009).