*International Journal of*
*Emerging Research in Management &Technology*
*ISSN: 2278-9359 (Volume-6, Issue-6)*

**Research Article** | **June 2017**

# News Category Classification Using Distinctive Bag of Words and ANN Classifier

**Amritpal Singh**
Student, Information Technology, CDAC,
Mohali, Punjab, India

**Sunil Kumar Chhillar**
Pr. Engineer, CDAC,
Mohali, Punjab, India

***Abstract—***

***Category classification, for news, is a multi-label text classification problem. The goal is to assign one or more categories to a news article. A standard technique in multi-label text classification is to use a set of binary classifiers. For each category, a classifier is used to give a "yes" or "no" answer on if the category should be assigned to a text. Some of the standard algorithms for text classification that are used for binary classifiers include Naive Bayesian Classifiers, Support Vector Machines, artificial neural networks etc. In this distinctive bag of words have been used as feature set based on high frequency word tokens found in individual category of news. The algorithm presented in this work is based on a keyword extraction algorithm that is capable of dealing with English language in which different news categories i.e. Business, entertainment, politics, sports etc. has been considered. Intra-class news classification has been carried out in which Cricket and Football in sports category has been selected to verify the performance of the algorithm. Experimental results shows high classification rate in describing category of a news document.***

***Keywords— Artificial neural network, News category classification, Bag of words, word tokenization***

## I. INTRODUCTION

Data mining is a process of discovering patterns, associations, changes, anomalies and significant structures from large amount of data stored in database, data warehouse or other information repositories. As huge amount of data in electronic form is available, turning such data into useful information is the most imminent task. Knowledge for broad application including market analysis, business management and decision support, data mining has attracted a great deal of attention in information industry in recent times. It has been widely considered as synonym of knowledge discovery in database, although view point of some researchers is that data mining is an essential step of knowledge discovery. With the emergence of WWW, it has become essential to handle a very large amount of electronic data majority of which is in the form of text by various data mining techniques like ANN,SVM, Naïve Bayes classifiers, Meta classifiers etc. [2]

A knowledge discovery process consist of an iterative sequence of following steps[1]:

*A. Data cleaning*: which handles noisy, erroneous, missing, or irrelevant data.

*B. Data integration*: where multiple, heterogeneous data source may be integrated into one.

*C. Data selection:*where data relevant to analysis task are retrieved from database.

*D. Data transformation:* where data are transformed orconsolidated into form appropriate for mining by performing aggregate operations.

*E. Data mining:* which is essential process where intelligent methods are applied in order to extract data patterns.

*F. Pattern evaluation:* which is to identify the truly interesting patterns representing knowledge based on some interestingness measure?

*G. Knowledge presentation:* where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

## II. TEXT CLASSIFICATION

The commonly used ways to store information is in the form of text like e-mails, web pages, newspaper article, market research reports, complaint letter from customer and internally generated reports. Online news papers provide news under various categories like national, international, politics, finance, sports, entertainment etc[2]. News articles on topical issue are helpful for company manager and other decision maker. It is time consuming task to select the interesting one from large amount of news article . News categorization is an easy approach to retreive the information quickly.Text classification is also an important part of text mining [3] which is based on expert knowledge and classification of the

 

document under the given set of categories. Data mining classification start with training set of documents that are already labeled with class. Text classification has two flavours - single label and multi label [3]. A single label document belong to only one class and multi label document may belong to more than one class.

## III. EXISTING WORK

Liang-Chih Yu et al. (2013) proposed a contextual entropy model to expand a set of seed words by discovering similar emotion words and their corresponding intensities from online stock market news articles. This was accomplished by calculating the similarity between the seed words and candidate words from their contextual distributions using an entropy measure. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Their proposed method considered both co-occurrence strength and contextual distribution, thus acquiring more useful emotion words and fewer noisy words and outperforming Pointwise Mutual Information (PMI) which only considers the co-occurrence strength.

YanghuiRao et al. (2014) propose an efficient algorithm and three pruning strategies to automatically build a word-level emotional dictionary for social emotion detection. In the dictionary, each word is associated with the distribution on a series of human emotions. In addition, a method based on topic modeling is proposed to construct a topic-level dictionary, where each topic is correlated with social emotions. Compared with other lexicons, the dictionary generated using our approach is language-independent, fine-grained, and volume-unlimited. The generated dictionary has a wide range of applications, including predicting the emotional distribution of news articles, identifying social emotions on certain entities and news events.

Limeng Cui et al. (2014) proposed a hierarchy method based on LDA and SVM. They first introduced the significance of news classification. Then, the concepts of topic model and SVM were introduced. They also did algorithm parameter adjusting.

Zhu Li-Juan et al. (2015) first of all, introduces the characteristics of Vietnamese news, on which the trigger words were selected; next, determine the event type based on the title, keywords and trigger words; finally, achieved the classification of Vietnamese news events through the event template and combining with the maximum entropy model.

WeitaoWeng et al. (2016) explore a combination of softmax regression model and the structured data of stocks to develop a structure-based classification model for stock news, which weights the probability output by the softmax regression model and the structured data to get the final probability values, and finally classified the news by the classification selection algorithm.

Yu-Chen Wei et al. (2017) constructed an Aggregate news sentiment index (ANS)I with the incorporation of public news relating to each of the individually listed firms. They proposed appropriate hypotheses to facilitate the investigation of the relationships between the ANSI, market returns, trading value, turnover ratio and Taiwan volatility index. The levels of changes in the weekly and monthly ANSI were also examined, and the ANSI regions were subsequently classified to confirm the findings.

## IV. PROPOSED WORK

In this work, Bag of words features have been used in which following steps are followed for feature extraction as shown in figure 4.1. The other features described above are not beneficial as vocab is the only solid source to extract the features.

- Tokenizing strings and giving a whole number symbol for every doable token by exploitation whitespaces as token separators.
- Remodeling the individual tokens for the aim of knowledge cleansing. During this step, all unique words are evaluated.
- Filtering the tokens so as to get rid of words with low data price. All tokens that occur below the threshold within the training set are filtered.
- Investigating what percentage of time the tokens occur in every document.
- Features and samples are outlined as follows:
- Every individual token incidence frequency (normalized or not) is treated as a feature.
- The vector of all the token frequencies for a given document is taken into account a variable sample.
- Concatenation of features in a matrix for different classes
- Training and testing the feature set with ANN classifier

Description of documents depends upon word utilized in it with very little or no care of wherever the word is employed within the document. This follows that a corpus of documents may be delineated through a matrix kind, whereby matrix $X$ with a row for every document and column for every token (i.e. word) happening within the corpus. The worth of cell $x_{ij}$ represents the amount of times the word $j$ seems within the example. In our analysis, this feature set is denoted as "Bow" (bag-of words). Since this approach is widely utilized in text classification, this feature set is deemed to be a baseline.

 

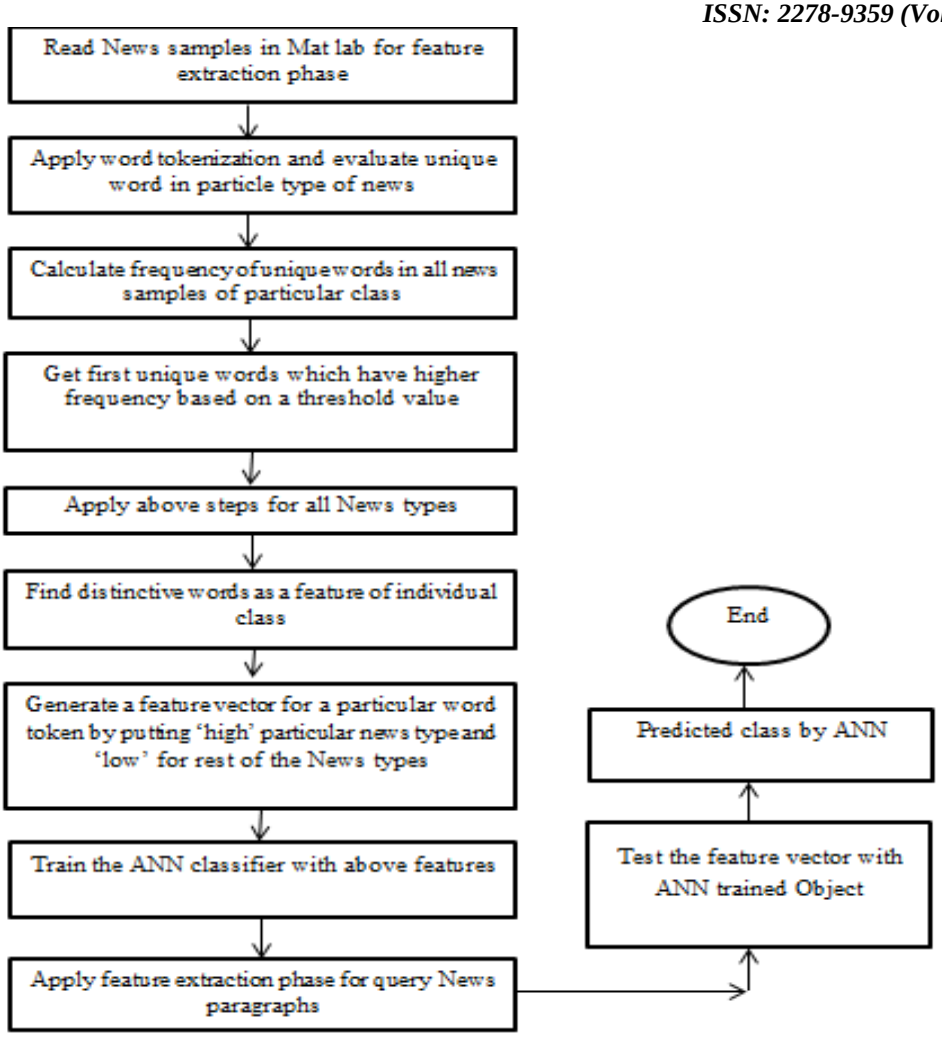


Fig. 1. Flowchart of the present work

## V. RESULTS AND DISCUSSIONS

As shown in the following table the news samples of different classes are differentiated using code implementation in the MATLAB. In Table 1 chosen news classes are classified on the basis of the distinctive word tokens having high frequency in the documents.

Table 1 Content Found In News Samples

| Properties Content %age | Business | Entertainment | Politics | Cricket | Football | Tech |
|---|---|---|---|---|---|---|
| business | 80.8630 | 2.7379 | 5.2288 | 3.0155 | 2.8218 | 5.3326 |
| Entertainment | 2.8932 | 82.5798 | 4.3093 | 2.9461 | 2.6205 | 4.6509 |
| Politics | 4.9699 | 3.38786 | 80.5787 | 3.2603 | 3.3399 | 4.46305 |
| Cricket | 0.3589 | 0.3149 | 0.5517 | 97.328 | 0.8671 | 0.57858 |
| Football | 1.9455 | 1.9394 | 2.4323 | 3.1787 | 88.5406 | 1.96327 |
| Tech | 4.1286 | 3.8203 | 3.9657 | 2.653 | 2.1129 | 83.3191 |

Table 2 Percentage of Samples Classified Accurately To Class

| **Class** | **%age** |
|---|---|
| **Business** | 100 |
| **Entertainment** | 100 |
| **Politics** | 100 |
| **Cricket** | 100 |
| **Football** | 100 |
| **Technology** | 100 |

 

The Fig. 2 is showing the mining News with combined polarity percentage for all News samples.

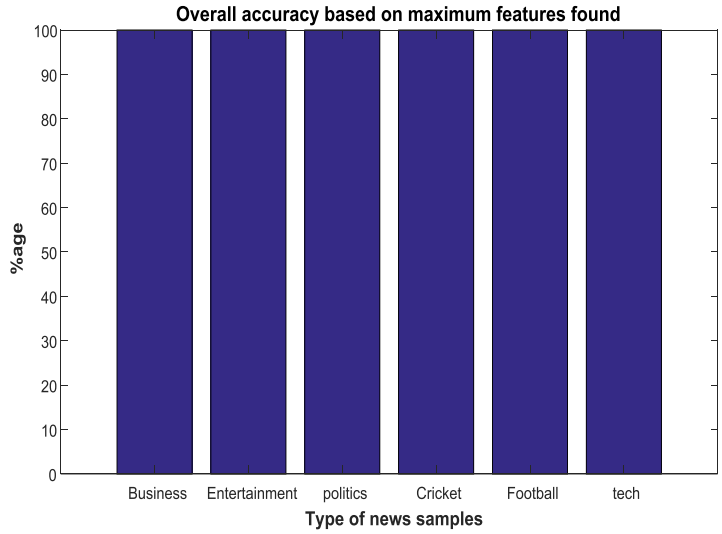


Fig. 2 Graphical representation of accuracy of classification

Pictorial representation of the outcomes of sample counted as per output of the code on our dataset are shown. Fig. 3 shows the bar chart of the mining samples for Business class.

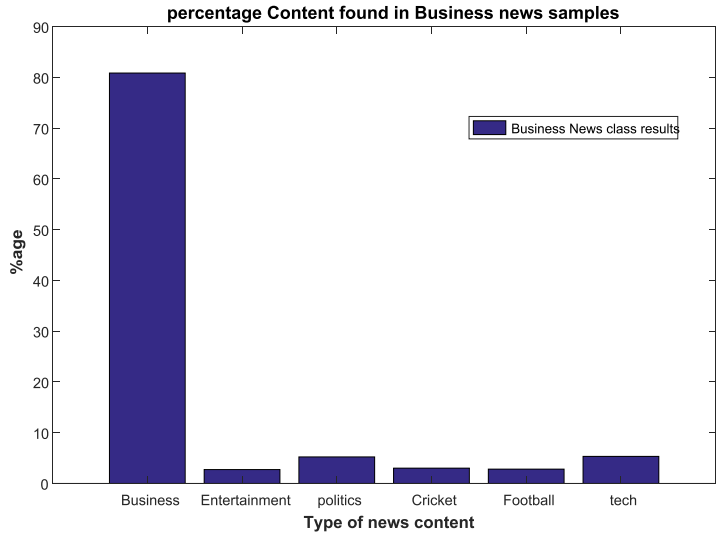


Fig. 3 Bar chart of the mining samples for Business

Following figure shows the bar chart of the mining News for Entertainment class

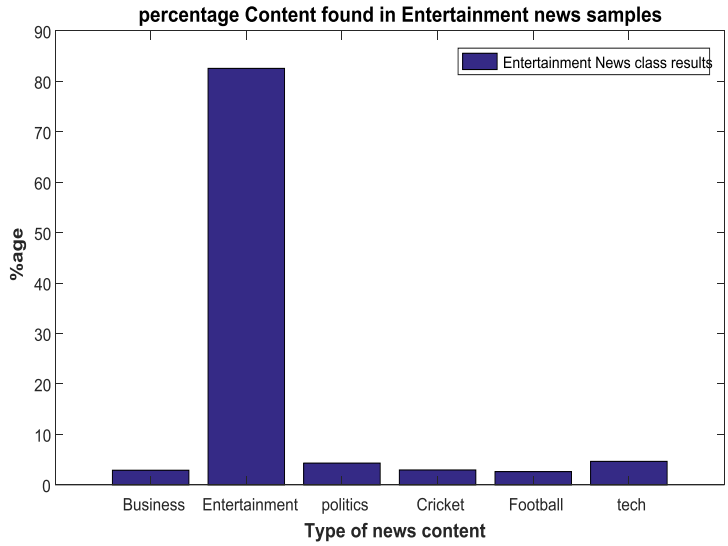


Fig. 4 Bar chart of the mining samples for Entertainment

 

Following figure shows the bar chart of the mining News for Politics class

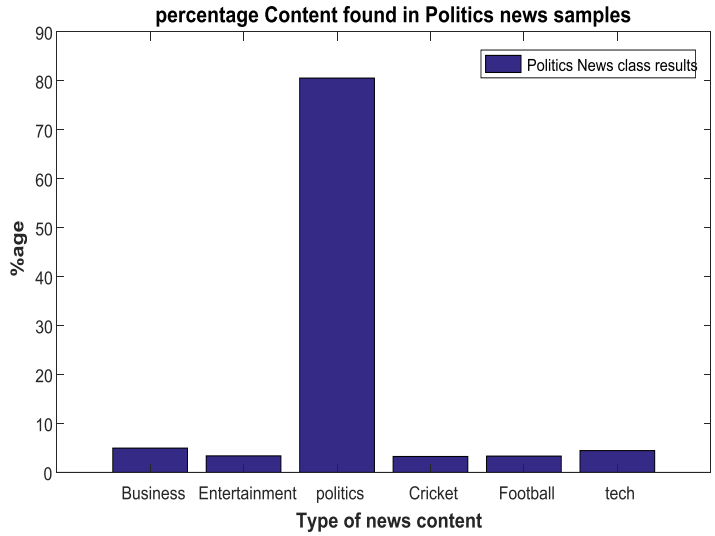


Fig. 5 Bar chart of the mining samples for Politics

Following figure represents the bar chart of the mining News for Cricket class

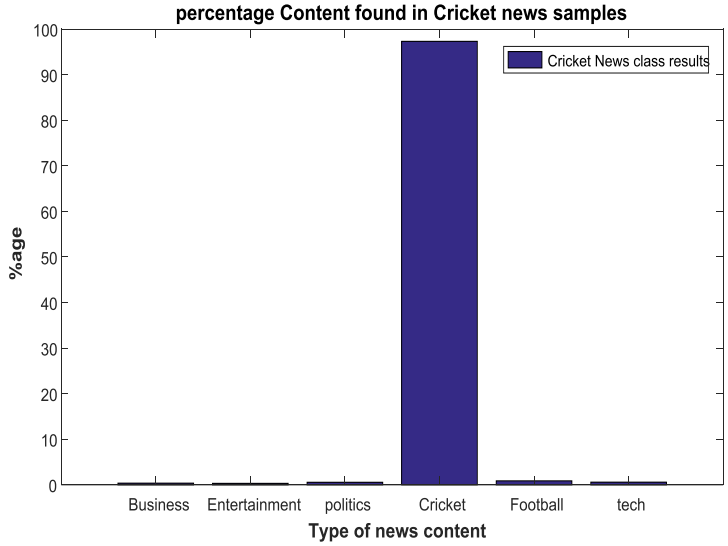


Fig. 6 Bar chart of the mining samples for Cricket

Following figure shows the bar chart of the mining News for Football class

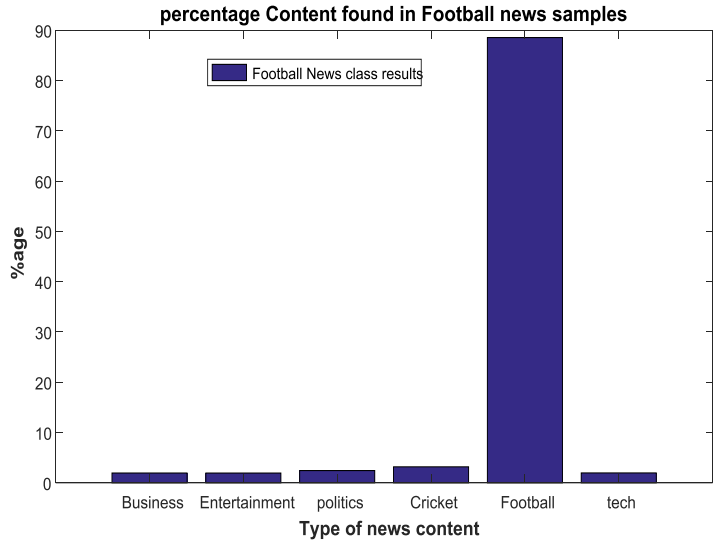


Fig. 7 Bar chart of the mining samples for Football

 

Following figure shows the bar chart of the mining News for Technology class

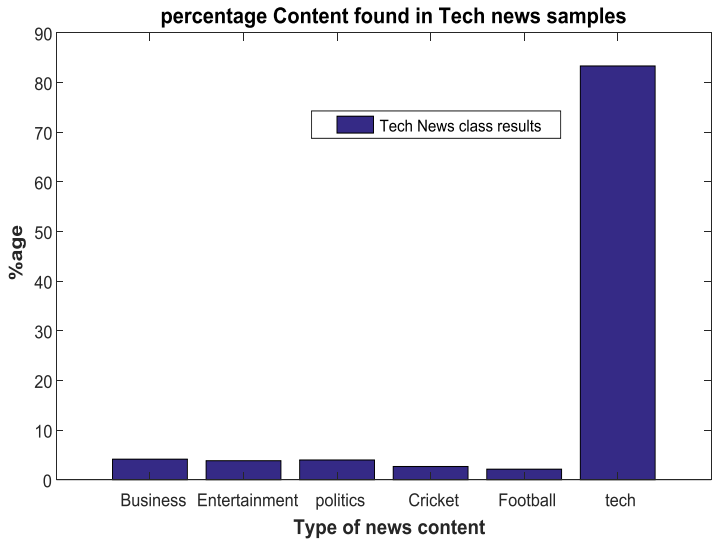


Fig. 8 Bar chart of the mining samples for Technology

As per the above showing bar charts and outcome results of the algorithm, the results are obtained in the form of percentage of news content in each news query sample. A news sample is classified as a particular type of news based on maximum of features. For the evaluation of results, the artificial neural network (ANN) is applied by collecting features based on bag of words. This neural network is helpful in creation of ranking level in the news samples. As per the ranking level is highest if all the word tokens features giving positive outcomes for a particular news class. It has been found that proposed methods gives 100% results in accuracy when tested on variety of news samples.

## VI. CONCLUSION

The past decade has seen the rapid development of different techniques to retrieve additional information from news. Various software and methods of news analytics are used to quantify textual information. Text sentiment extracts text's attitude by getting unique words and has proved extremely useful in a variety of contexts. There are many methods proposed to classify news in different classes but inter class classification is not touched much. As for example sports news has variety of sub –categories i.e. Cricket, football, hockey, athletics etc. and it is hard to separate them as most of the commentary words used are same. Therefore unique properties need to be extracted to classify such type of news. In this method, distinctive bag of words is taken as feature which keeps the individual property of a subclass. First of all, all unique words are evaluated in each class and then we sort the according to its frequency. Then a threshold is set to select first high frequency words in the news document. Then distinctive words have been chosen which are only found in one class. It eradicates the word features which can be found in two or more classes and remained only unique bag of words for each sub class. Then neural network objet has been trained by these distinctive word features. Then all query samples are tested with trained neural network object. Experimental results show 100 % results in classification. We have also analyzed the type of content its percentage found in each news class as some distinctive features found in one or another classes.

## REFERENCES

[1] Bing Liu "S*entiment Analysis and Opinion Mining*" Human Language Technologies, ISBN: 9781608458851, pp: 1-167, 2012

[2] Chee-Hong Chan Aixin Sun Ee-Peng Lim, "Automated Online News Classification with Personalization," in *Proc. 4th international conference on Asian Digital Libraries (ICADL2001),* pp: 320-329, Bangalore, December 2001

[3] Choiru Za'in, Mahardhika Pratama, Edwin Lughofer, Sreenatha G. Anavatti, "Evolving type-2 web news mining" Published in: *Applied Soft Computing*, Volume 54, pp: 200–220, May 2017

[4] Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, Hsuan-Shou Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news" Published in: *Knowledge-Based Systems*, Volume 41, pp: 89–97, March 2013

[5] Yang huiRao, Jingsheng Lei, Liu Wenyin, Qing Li, Mingliang Chen, "Building emotional dictionary for sentiment analysis of online news" Published in: *World Wide Web*, Issue 4, Volume 17, pp :723–742, July 2014

[6] L. Cui, F. Meng, Y. Shi, M. Li and A. Liu, "A Hierarchy Method Based on LDA and SVM for News Classification," in *Proc. 2014 IEEE International Conference on Data Mining Workshop, ICDM Workshops 2014,* pp: 60-64, Shenzhen, China, 14 December 2014

 

[7] E. Kiliç, M. R. Tavus and Z. Karhan, "Classification of breaking news taken from the online news sites," in *Proc. 2015 23nd Signal Processing and Communications Applications Conference (SIU)*, ISBN: 978-1-4673-7387-6, pp: 363-366, Malatya, Turkey, 16-19 May 2015

[8] Z. Li-juan, Z. Feng, P. Qing-qing, Y. Xin and Y. Zheng-tao, "A classification method of Vietnamese news events based on maximum entropy model," in *Proc. 2015 34th Chinese Control Conference (CCC)*, ISBN: 978-1-4673-7443-9, pp: 3981-3986, Hangzhou, China, 28-30 July 2015

[9] Yu-Chen Wei, Yang-Cheng Lu, Jen-Nan Chen, Yen-Ju Hsu, "Informativeness of the market news sentiment in the Taiwan stock market" Published in: *The North American Journal of Economics and Finance*, Volume 39, pp: 158–181, January 2017

 