NgramPOS: A Bigram-based Linguistic and Statistical Feature Process Model for Unstructured Text Classification

Sepideh Foroozan Yazdani[a]

[a] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
foroozan.sepideh@gmail.com

Zhiyuan Tan[b]

[b] School of Computing, Edinburgh Napier University
z.tan@napier.ac.uk

Mohsen Kakavand[c]

[c] School of Science and Technology, Sunway University, Malaysia
mohsenk@sunway.edu.my

Aida Mustapha[d]

[d] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
aidam@uthm.edu.my

Abstract. Research in financial domain has shown that sentiment aspects of stock news have a profound impact on volume trades, volatility, stock prices and firm earnings. With the ever growing social networking and online marketing sites, the reviews obtained from those, act as an important source for further analysis and improved decision making. These reviews are mostly unstructured by nature and thus, need processing like clustering or classification to provide different polarity categories such as positive and negative in order to extract a meaningful information for future uses. Accordingly, in this study we investigate the use of Natural Language processing (NLP) in a way to improve the sentiment classification performance to evaluate the information content of financial news as an instrument for using in investment decisions system. Since the proposed feature extraction approach is based on the occurrence frequency of words, low-frequency linguist features that could be critical in sentiment classification are typically ignored. In this research, therefore, we attempt to improve current sentiment analysis approaches for financial news classification in consideration of low-frequency, informative, linguistic expressions. Our proposed combination of low and high-frequency linguistic expressions contributes a novel set of features for text sentiment analysis and classification. The experimental results show that an optimal Ngram feature selection (combination of optimal unigram and bigram features) enhances sentiment classification accuracy than other types feature sets.

Keywords: Unstructured Text; Bigram Model; Machine Learning; Natural Language Processing; Sentiment Classification; Financial News Analysis.

1. Introduction

Financial news is considered a significant source to estimate stock price by analysts and investors. Since it conveys the latest news of financial markets and firms as well as the local and global financial policies, financial news has irresistible influence on stock markets and returns. In accordance with the Efficient Market Hypothesis (Fama, 1965), all available information of stocks reflects on their market prices, Tetlock (2007) showed the qualitative textual impact stock prices. Hence, information particularly including financial, market and economic news from traditional and new media plays an essential role when investors estimate stock prices. This is because the collection of massive vital information contained in the news. Besides, due to the rapid increase of financial news in new media over the past decades, it causes great difficulties for investors to fully investigate all available information before making financial decisions. Although attempts have been made utilizing text mining to convert unstructured information to a standardized format for classification, the studies in automated classification of textual financial news are still in their infancy [3]. Nevertheless, other challenges are found in text mining and classification of financial news, such as feature extraction, feature selection, and classification process.

In literature, only a few studies (Hagenau et al., 2013; Khadjeh Nassirtoussi et al., 2015) have employed sophisticated feature extraction approaches, such as noun phrase to extract low-frequency-based features. Most of the current studies (Koppel and Shtrimberg, 2006; Groth and Muntermann, 2011; Yu et al., 2013) on sentiment-based financial news analysis typically rely on

simple frequency-based textual presentations, such as Bag-of-Words (BoW) in which each piece of news is represented by the occurrence frequencies of distinct words. Some other research works (Généreux et al., 2011; Zhai et al., 2011), in contrast, have employed unigram text characterization, which shows similar characteristics to the BoW approach thanks to their commonality in consideration of occurrence frequency. However, in coping with a large volume of data, the process is encountered with a lot of the low frequency bigrams which can be considered as an informative and sentiment feature while the extraction of words is based on their high frequency which leads to ignore low frequency-based linguist features that can be worth to sentiment classification [9].

This research focuses closely on related studies that explore the impact of different feature types as input for SVM classification. this research is partly related to Généreux et al. (2011). The authors employed the different types of features as unigrams, stems, financial terms, health metaphors and agent-metaphors along with some feature selection methods (Information Gain (IG), (Chi-Square (CHI) and Document Frequency (DF)) and two feature weighting methods (binary and TF (Term Frequency) which assign different values to each feature. For example, binary feature weighting assigns zero and one to each feature as the presence or absence of that word in the content. The following sections will discuss feature weighting methods in detail. As previously mentioned unigrams and single words due to the possibility of having a higher frequency typically achieve higher accuracy (67.6%) than the other type of features. However, each unigram cannot convey sentiment orientation of each document. Although, the authors separately employed financial terms and health metaphors in order to help to determine the polarity of text but it led to the lowest accuracy, 59.2%, 52.4%, respectively. On the other hand, the highest accuracy was achieved by TF feature weighting method while TF can only determine the importance and value of a term in a document.

Another relevant study was performed by Hagenau et al. (2013) who also focused on different types of features including dictionary-based single words that retrieved from corpus, bigrams, 2-word combination, and noun phrase, and used CHI as feature selection. The results indicated that the worst accuracy is related to bigrams with frequency-based feature reduction which is expected while 2-word combination has the high accuracy by two feature selection methods (frequency-based and CHI) with an accuracy of 62.0%, 72.6% respectively. This shows the high impact of CHI on classification accuracy while CHI method is strongly sensitive to sample size and small frequencies. It is also, independent of the strength of the relationship between the features.

Another related work was conducted by Joshi et al. (2016). The authors used news articles in order to predict stock market. The authors performed sentiment analysis based on positive and negative single words in the news documents. As discussed, the single words alone are unable to convey sentiment and semantic of textual news. Observations showed that Random Forest and SVM perform well in all types of testing from 86% to 92%, while Naïve Bayes (NB) performance is around 86%.

Chen & Chen, 2017 integrated and analyzed news article and financial blogs to develop a public mood dynamic prediction model for stock markets based on behavioral finance. These authors used a text segmentation toolkit to break the text into words and remove low frequency vocabulary. In the stage of feature acquisition and extension they utilized two approaches to extract words where the first way is without considering the part of speech and the second is the extension of only nouns. Finally, the authors used PCA for dimensionality reduction of features and achieved the highest performance by applying feature weighing method (TF-IDF) in one of the scenarios with an accuracy of 65.81%.

Chan & Chong, 2017 proposed a sentiment analysis engine (SAE) which used linguistic analyses based on grammars where this engine extends sentiment analysis as word token and phrase for each sentence. The experimental result in this research, in terms of accuracy, of different models in rule-based unsupervised sentiment classification using two different word lists in the range of 49.8% to 82.1% have been reported.

As discussed earlier, the main focus of the existing works is to identify the polarity of news contents. However, the existing approaches have employed variety of feature extraction and selection methods to classify news that have leads to different outcomes, this indicates feature extraction and selection methods play a significant role in sentiment classification. The performance of the related works will be improved if these weaknesses can be addressed.

- Identify Linguistic and statistical relevant features
- High Dimensionality of feature space.

This research proposes a model to sentiment classification of financial news based on the combination of statistical-based and linguistic-based approaches. This model will be able to generate an optimal feature space based on high and low frequency features which is presented as NgramPOS model. The model that contains stronger sentimental low-frequency bigrams influences the decision making behavior of investors and stockholders. The contributions of this paper are two-fold:

- Demonstrate the Strength of Incorporation of Low Frequency-based Prominent (bigram_POS) Statistical and Linguistic Features to Determine the Polarity of News as Positive and Negative.
- Using Uni_POS and Bi_POS phrases as sentiment features for supervised machine learning.

Hence, the purpose of this article is to enhance the performance of high frequency-based features (Uni-POS) by incorporating more sentiment-rich information in the form of Bi-POS phrases (low frequency-based features) with the Uni-POS feature where these features are initially extracted using fixed patterns based on part of speech (POS) and principal component analysis (PCA).

The rest of this paper is organized as follows. Section 2 describes the research methodology, which focuses on how to build a bigram-based linguistic and statistical feature process model to financial news sentiment classification. Section 3 describes the experimental design and results for this paper. A conclusion and further research directions are given in Section 4 and 5.

2. Framework of Financial News Classification

This framework based on the shortcomings discussed in the previous section is designed in different phases including financial news preprocessing, feature processing, and financial sentiment classification. The first phase, includes all activities related to the preparation of news text and the second phase is concerned with all processes related to feature extraction and selection and the last phase are included the classification methods. The components of this framework and the requirements with respect to each individual component are discussed as follows.

2.1. Financial News Preprocessing

This phrase is composed of three components: extract and collect financial news, financial news labeling, and news preparing and cleaning, each of which is described as follows.

2.1.1. Financial News Collection

Various sources of textual financial news are available, which include media articles, public disclosures and internet posting. In this work, financial news was gathered from *Google Finance* that is part of Internet posting. Internet posts are a useful source of textual sentiment because many people spend a notable amount of time every day reading and writing internet posting, however, it is potentially noisier than other sources because it contains more view from individual traders.

2.1.2. Financial News Labeling

After storing financial news items extracted from *National Association of Securities Dealers Automated Quotations* (NASDAQ) and *New York Stock Exchange* (NYSE) stock markets as text files, news labeling was performed on the news data using *R* packages, namely *tm.plugin.tags* and *tm.plugin.sentiment*. The label of each stored textual news file was reviewed and validated manually [14] against an existing financial news classification (i.e. news documents categorized by *Lydia* sentiment analysis) and then labeling errors were fixed.

2.1.3. News Cleaning

As mentioned in subsection 3.1.1, the financial news collected from *Google Finance* contains a variety of news articles in the form of HTML. The main content body of each news is enclosed in between XML tags and includes four elements: news headline, news body, stock ticker, and publication date and time [15]. **This step deals with a variety of words and signs in financial news**, **which** is **irrelevant and without any sentiment** such as XML tags, email address, currency signs and so forth**. Therefore, this makes the cleaning text process for these financial news pieces required and different from other types of text. Hence before feature preprocessing phase, the text cleaning is applied.**

2.2. Feature Processing

According to what is discussed in the introduction, the structure of the most of the related works to financial news sentiment relies on fairly simple approaches of feature selection and extraction such as *bag of word* (BoW) or unigrams. Hence, it is interesting to know if the combination of statistical and linguistic methods can actually enhance the performance of financial news classification and this is the main idea behind designing the proposed models in this research.

2.2.1. Part-of-Speech tagging and Tokenization

Generally, algorithms make use of a Natural Language Processing (NLP) technique called Part-of-Speech (PoS) tagging. PoS tagging is a task of labeling, where each word in a sentence is tagged with an appropriate PoS category. Common PoS categories in English grammar are: noun, verb, adjective, adverb, preposition, pronoun, conjunction, and interjection which other categories can be arose from them as different forms. Since PoS identifies sentiment expressions and semantic relationships between these expressions, we use it to distinguish candidate features that indicate sentiment orientation.

2.2.2. *N*-gram Model

An *n*-gram model is a sequence of n consecutive symbols that can be characters, words, bytes, or any other continuous symbols. For instance, a 1-gram (or unigram) is one-symbol, a 2-symbol (bigram) is a two-symbol sequence of symbols, and 3-gram (trigram) is a three-symbol sequence of symbols so on and so forth (Généreux et al., 2011; Zhai et al., 2011). It uses the previous $n-1$ symbol to predict the next $p(t_n|t_{n-1})$. Research (Pederson, 2001; Dave et al., 2003) in sentiment analysis has shown further that *n*-grams are effective features for word sense disambiguation. Hence, this study uses word-based *n*-gram model to extract unigram, bigram and trigram features where it will be able to reveal correlations between the words and the importance of individual phrases (Mejova and Srinivasan, 2011). Following the PoS tagging and *n*-gram approaches, processing is then performed as Tokenize the financial news documents into individual tokens, remove punctuations, and uniform the whole text.

2.2.3. Unsupervised Feature Weighting Methods

Term weighting methods are commonly grouped into two categories, namely supervised and unsupervised methods (Lan, Tan, Su, & Lu, 2009; Pham Xuan & Le Quang, 2014). Generally, the unsupervised or traditional term weighting methods are originated from information retrieval field. In contrast to the unsupervised feature weighting methods, the supervised weighting methods use the prior information of training document in predefined categories. In this study, we use unsupervised weighting methods are derived from information retrieval. *Binary*, *Term Frequency* (TF), and *Term Frequency-Inverse Document Frequency* (TF-IDF) [21] belong to unsupervised term weighting method.

2.2.4. Dimensionality Reduction Using Principle Components Analysis

One of the most popular feature extraction methods is Principle Component Analysis (PCA) which has been broadly applied to a variety of data from social science, biology, finance [22] and so on. In brief, PCA seeks to map data points, such as financial news documents, from a high dimensional space to a low dimensional space while keeping the most significant linear structure intact.

Given a dataset consists of n rows (e.g. financial news documents) and m columns (e.g. their features), a $n \times m$ matrix $\boldsymbol{X}$ is derived. Let k be the dimensionality of the new space to which we seek to map the data, and $k \ll m$. The covariance between two features is defined by equation (1).

$$\sigma_{jp} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x_j})(x_{ip} - \bar{x_p}) \quad (1)$$

In this study, two criteria are applied to determine the *k* most significant components in the data space through PCA. They are proportion of variance, and Kaiser's rule (Alpaydin, 2010).

- **Proportion of Variance**

The proportion of variance criterion search for the first *k* eigenvectors of the covariance matrix with the largest eigenvalues. After λ_i (eigenvalues) are sorted in descending, the cumulative proportion is described by the k principal components as follows (Alpaydin, 2010) :

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \leq threshold, \quad (k \ll m) \tag{2}$$

where λ_i are eigenvalues and m is a dimensionality of different feature spaces in this model. $k \ll m$ happens when the dimensions are extremely correlated. Therefore, we will have a small number of eigenvectors, and so a large reduction in dimensionality may be achieved. Normally, the threshold is placed in range 80%-90%, we consider 90% as a reasonable criterion here.

- **Kaiser's Rule**

Kaiser's rule (Alpaydin, 2010) is another criterion to eliminate the eigenvectors with low eigenvalues. Since $\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} \sigma_i^2$ this means the average eigenvalues is equal to the average input variance. Kaiser's rule suggests retaining the eigenvectors with eigenvalues greater than the average eigenvalues. As such, only those with a variance greater than the overeager input variance are kept.

$$\lambda_i \in top\ k\ eigenvalues\ \{\lambda_1, \lambda_2, \dots, \lambda_k\}\ if\ \lambda_i \geq \sum_{i=1}^{m} \frac{\lambda_i}{m} \tag{3}$$

2.3. Financial News Classification

The studies by Sebastiani (2001) and Khadjeh Nassirtoussi et al. (2014) confirm that SVM has been extensively and successfully applied as a textual classification and sentiment learning approach. The specific characteristics of textual data classification, such as High-dimensional space and low-discriminative lexical feature, Irrelevant features, and Data feature sparsity, encourage the adaptation of SVM classifier (Joachims, 1998; Hajek & Henriques, 2017) in this work.

2.3.1. Support Vector Machine

The SVM is a generalization of a simple and intuitive classifier which was developed by Vladimir Vapnik [27] . SVM uses a linear boundary called a hyperplane to separate news data into groups of the same elements (positive news, negative news), typically represented by the class labels (i.e. *pos* and *neg*). The news data is said to be linearly separable if the data can be separated by a straight line in a 2-dimensional space or a flat surface in a higher-dimensional space. The primary goal of SVM algorithm is to identify a Maximum Margin Hyperplane (MMH), which creates the greatest separation between classes. The instance (news) from each class that is the closest to the MMH or on the margin is called a support vector (Lantz, 2013). It is noteworthy that SVM hyperplanes are determined by the support vectors (a small subset of the train set). Nonlinear SVM applies kernel methods for mapping the data (news) that has a nonlinear decision boundary in input space to a higher dimensional space to obtain a MMH. The used methodology in nonlinear SVM is performed by mapping each vector x (each news document) from n-dimensional input space to a new space $\Phi(x)$ (feature space) with m-dimensional to construct a hyperplane to separate data in feature space. The methodology defined by Schölkopf & Smola (2005) is shown in Equation (5):

$$f(x) = \boldsymbol{w}.\boldsymbol{\Phi}(x) + b \tag{4}$$

f(x) demonstrates a linear classifier as the inner product of two vectors, so that the vector w is known as the weight vector, and b is called the bias. (Joachims, 1998; Ooi et al., 2010) showed that the RBF kernel usually outperforms other kernels in both accuracy and convergence time if the data are normalized and proper values of the kernel parameters are chosen. Hence, this research uses a linear kernel first as baseline, and then used RBF kernel to determine best classifier to sentiment classification.

2.3.2. Model Tuning in SVM

Grid-search optimization method is more computational cost-effective than the other sophisticated optimization methods (e.g. the Newton's or Quasi-Newton's method) in searching for the best values of the parameters. This is because fewer parameters, which are actually two parameters in the case of grid-search, need to be optimized, that is why that the grid-search approach can be easily parallelized because C and γ parameters are independent, unlike other methods that have iterative processes (Hsu et al., 2010). In line with the prior research presented in (Hsu et al., 2010; James et al., 2013), we treat a normal sequence of $(\mathbf{C}, \boldsymbol{\gamma})$ as $\gamma = 10^{\wedge}(-6:-1)$, $\gamma = 10^{\wedge}(-3:2)$ for RBF SVM and a range of C as $C = 10^{\wedge}(-3:2)$ for Linear SVM. Due to the sparse distribution of features in dataset that was used to classify sentiments of financial text news and using all of the news documents in the corpus as both training and testing, we utilizes *k-fold cross-validation* method to assess (Kohavi, 1995).

3. Experimentation and Results

NgramPOS model is constructed using the linguistic-based model relying on Ngram model that is consisted of five main stages. Fig. 1 shows the conducted experiment to create the NgramPOS model.

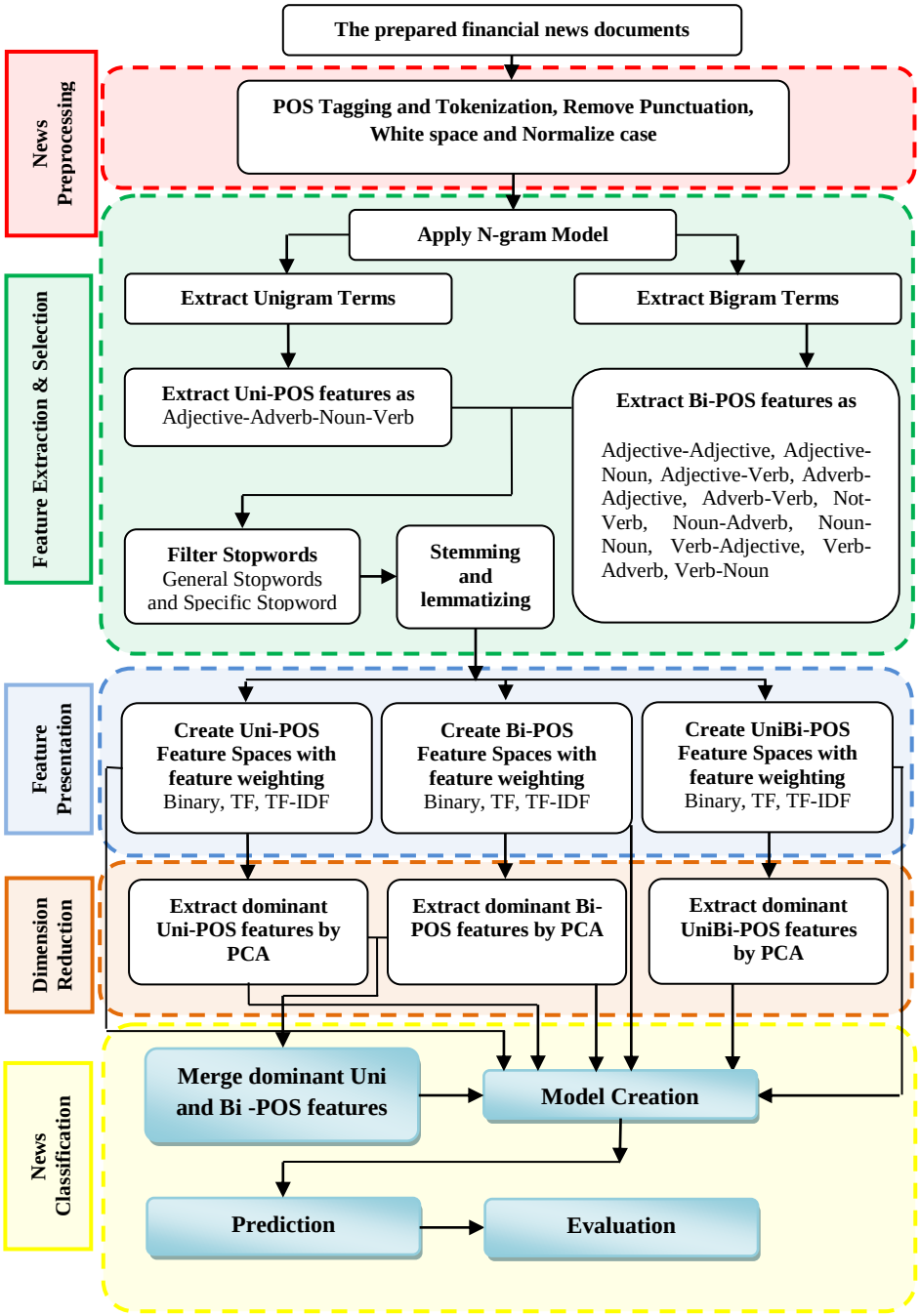

Fig. 1. The Process Flow in NgramPOS-based Methodology

3.1. Financial News Preprocessing in NgramPOS-based Methodology

The financial news documents after several processes are prepared for preprocessing.

- **POS Tagging and Ngram**

In this step, at first a few basic *Natural Language Processing* (NLP) procedures are applied on the cleaned text news that utilizes the standard *Penn Treebank POS Tags*[1] and Stanford[2] [34]. Fig. 2 shows a sentence of financial news after applying POS tagger.

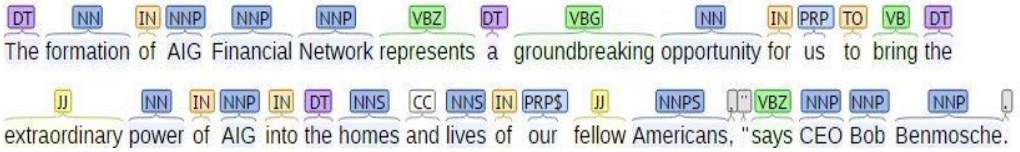

Fig. 2. The POS Tagging Results for a Sample Sentence

To have a clear understanding of the preprocessing step, for instance, let $\boldsymbol{\mathcal{T}}$is assumed with all possible words in the context of news documents, and an input cleaned text news file will be considered as $d_i = (t_1, t_2, \ldots, t_{m-1}, t_m)$, where $t_j \in \boldsymbol{\mathcal{T}}$ and $d_i \in \boldsymbol{\mathcal{D}}$, where $\boldsymbol{\mathcal{D}} = \{d_1, d_2, \ldots, d_n\}$, so that m, n are the size of the vocabulary that are extracted from the text news and the total number of financial news documents, respectively.

Table 2 shows the information about total number of financial news documents and depicts distribution of N-gram models (unigrams, bigrams and trigrams) collected from clean financial news documents, where "the number of terms" column indicates the number of unigrams, bigrams, and trigrams with their frequency (their iterations) in the whole corpus and the last column shows the richness of the n-gram terms in the entire of news documents which is obtained by dividing the number of the number of n-gram terms (unigrams, bigrams, trigrams) on the list of their vocabulary ($Richness = \frac{the\ number\ of\ n-gram\ words}{the\ list\ of\ n-gram\ words\ vocabulary}$), where the vocabulary is list of n-gram terms without their frequency.

Table 2. Data Set Inforamtion and Distribution of N-gram Models in the Financial News Documents

Data	Ngram models	The number of terms	Richness
News documents: 1917	Unigram	28792	903063/28792=31.37
Positive news (POS): 1160	Bigram	273172	901146/273172=3.30
Negative news (NEG): 757	Trigram	539110	899229/539110=1.67

As can be seen, the unigrams have the highest frequency which leads to high richness while the bigrams and the trigrams have much lower richness than the unigrams. Hence, we use unigrams and bigrams in order to feature extraction.

3.2. Uni-POS and Bi-POS Feature Extraction and Selection

After the tokenization has been performed and a POS for each word was assigned, the feature process will start to choose the most relevant and riches terms as feature. Most of the researches on sentiment analysis (Benamara et al., 2007; [36] [40][46]Hatzivassiloglou et al., 2009) have focused on recognition of sentiment expressions, opinion words.

[1] . https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[2]. http://nlp.stanford.edu:8080/corenlp/process

Hence, POS is applied to identify terms that are used for sentiment analysis in order to recognize polarity each news for sentiment classification, such as adjective, adverb, noun and verb (unigram pattern), and collocations like adjective+noun (bigram pattern).

Besides, adjectives, other content words such as adverbs, nouns, and verbs are also used to express sentiment and opinion. For instance, a sentiment expression that uses an adjective as "negative" denotes the sentiment towards its modifier noun such as in "negative value", and the whole noun phrase "negative value" itself becomes a sentiment expression with the same polarity as the sentiment adjective, in this case negative for "negative". Likewise for an adverb, if a sentiment expression uses an adverb as "dramatically", denotes the sentiment towards it modifier verb such as in "increase dramatically". Hence, this section extracts different feature sets based on specific rules; this means that only unigrams are considered as feature that the POS tags assigned to them to be in one of the forms of adjective, adverb, noun, or verb. This rule is named as Unigram Pattern, and the unigrams extracted by using this pattern are called Uni-POS features. After extracting unigrams, the rules are intended to extract Bi-POS features from bigrams that are adopted from study of Turney (2002) who considered phrases that included adjective or adverb. Moreover other patterns have included that contain verbs or nouns to express sentiment. This model uses McDonald's stop word list is provided by Bill McDonald to financial domain; it is made up of several groups of generic, names, dates and numbers, geographic, and currencies with 13775 stop words. As can be seen in Fig. 3, four uni-POS and eleven bi-POS feature sets are created in the process of feature extraction and selection.

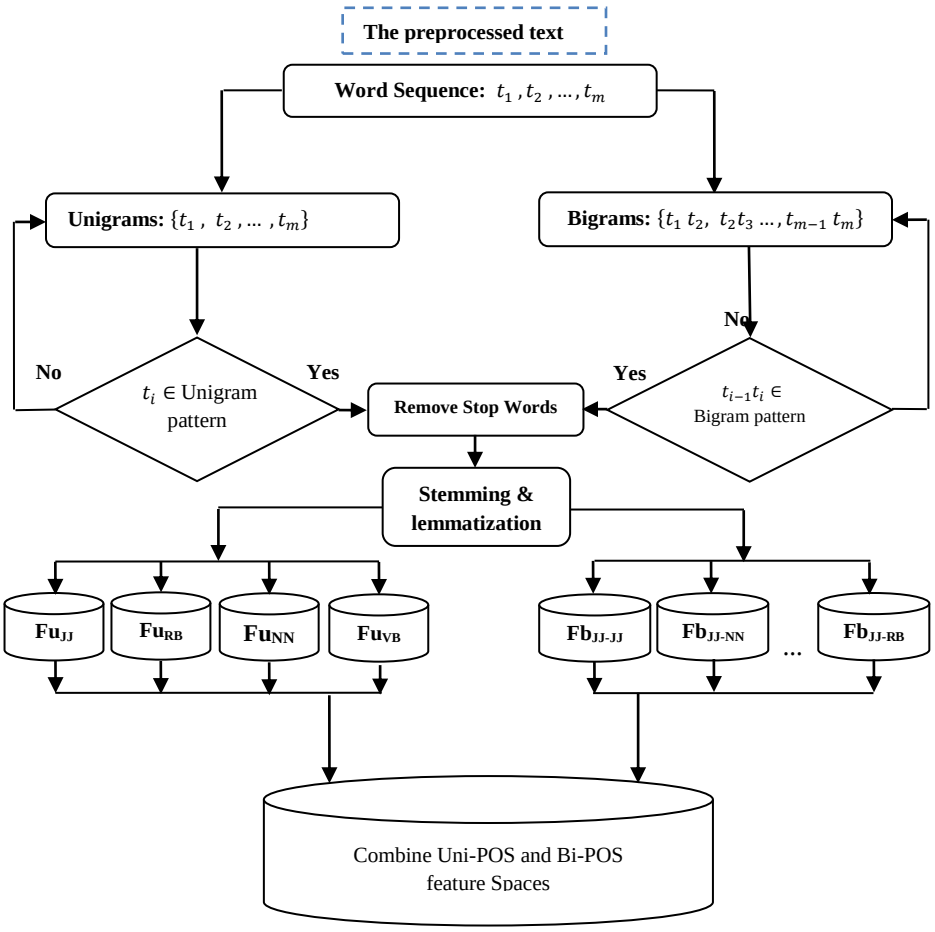

Fig. 3. The Process Uni-POS and Bi-POS Feature Extraction and Selection

let $\mathcal{D}$ be the set of financial news documents and $\mathcal{F}_{\text{Uni}-\text{POS}}$ be the union of Uni-POS feature sets where each of them are defined as follows.

$$\boldsymbol{\mathcal{D}} = \{d_1 , d_2 , \dots , d_n\} \quad (7)$$

$$\boldsymbol{\mathcal{F}u}_k = \{fu_1 , fu_2 , \dots , fu_{|\mathcal{F}u_k|}\} \; where \; k \in \{JJ, RB, NN, VB\} \quad (8)$$

$$\boldsymbol{\mathcal{F}}_{Uni-POS} = \boldsymbol{\mathcal{F}u}_{JJ} \cup \boldsymbol{\mathcal{F}u}_{RB} \cup \boldsymbol{\mathcal{F}u}_{NN} \cup \boldsymbol{\mathcal{F}u}_{VB} \quad (9)$$

$$\boldsymbol{\mathcal{F}}_{Uni-POS} = \{fu_1, fu_2, \dots , fu_{|\mathcal{F}_{Uni-POS}|}\} \quad (10)$$

$$|\boldsymbol{\mathcal{F}}_{Uni-POS}| < |\boldsymbol{\mathcal{F}u}_{JJ}| + |\boldsymbol{\mathcal{F}u}_{RB}| + |\boldsymbol{\mathcal{F}u}_{NN}| + |\boldsymbol{\mathcal{F}u}_{VB}| \quad (11)$$

Likewise for Bi-POS feature sets, let $\mathcal{F}_{\mathrm{Bi-POS}}$ be the union of Bi-POS feature sets; since none of them alone cover all of news documents, the bigrams are combined as a unique Bi-POS feature set. Before combine these feature spaces we remove features with frequency less than 2 (freq<2), thus the final Bi-POS feature set would be as follows:

$$\boldsymbol{\mathcal{F}b}_k = \{fb_1 , fb_2 , \dots , fb_{|\mathcal{F}b_k|}\} \; where \; k \in \{JJ - JJ, JJ - NN, \dots , JJ - RB\} \quad (12)$$

$$\boldsymbol{\mathcal{F}}_{Bi-POS} = \boldsymbol{\mathcal{F}b}_{JJ-JJ} \cup \boldsymbol{\mathcal{F}b}_{JJ-NN} \cup \dots \cup \boldsymbol{\mathcal{F}b}_{JJ-RB} \quad (13)$$

$$\boldsymbol{\mathcal{F}}_{Bi-POS} = \{fb_1, fb_2, \dots , fb_{|\mathcal{F}_{Bi-POS}|}\} \quad (14)$$

$$|\boldsymbol{\mathcal{F}}_{Bi-POS}| \leq |\boldsymbol{\mathcal{F}b}_{JJ-JJ}| + |\boldsymbol{\mathcal{F}b}_{JJ-NN}| + \dots + |\boldsymbol{\mathcal{F}b}_{JJ-RB}| \quad (15)$$

3.3. Financial News Documents NgramPOS-based Representation

The third phase of our experiment is document presentation. This research uses Vector Space Model (VSM) to transform the content each news document to a vector in the desired feature space so that the news could be recognized and classified by a classifier. According to studies conducted by Foroozan Yazdani et al. (2016), the different feature spaces and feature weighting methods have different effects in the performance of sentiment classification, hence in this stage, VSM would be created based on the three feature sets Uni-POS, Bi-POS, and UniBi-POS. Just like the previous subsection; $\boldsymbol{\mathcal{D}} = \{d_1 , d_2 , \dots , d_n\}$ is considered as a set of news documents, in this case, three document-feature matrix is created, where rows and columns are indicated by the documents and the features. To measure the importance of each feature in a news document or the whole corpus, it must be associated with a value which is called weight. The three unsupervised feature weighting methods *binary*, *Term Frequency* (TF), and *Term Frequency Inverse Document Frequency* (TF-IDF) are assigned to any feature.

With regard to the nature of contextual data (financial news), since PCA is computationally inexpensive and can handle sparse and skewed data, it can be the best choice as a method to dimensionality reduction of feature space (Han and Kamber, 2006).

3.4. Financial News Classification Using RBF SVM for Uni-POS, Bi-POS, and UniBi-POS Feature Spaces after Applying PCA

In this step, PCA are applied along with two criterions proportions of variance and Kaiser's rule in order to reduce dimensions of feature spaces. Therefore, the experiment is conducted in two ways with two criterions: firstly, k prominent features are extracted from UniBi-POS as a new feature space, and then in secondly, k optimal features are extracted individually, from Uni-POS and Bi-POS and then merged as another new feature space, which are defined as follows:

$$\boldsymbol{\mathcal{F}}_{UniBi-POS} = \{fu_1, fu_2, \dots , fu_{|\boldsymbol{\mathcal{F}}_{Uni-POS}|}, fb_1, fb_2 \; \dots , fb_{|\boldsymbol{\mathcal{F}}_{Bi-POS}|}\}$$

$$|\boldsymbol{\mathcal{F}}_{UniBi-POS}| = |\boldsymbol{\mathcal{F}}_{Uni-POS}| + |\boldsymbol{\mathcal{F}}_{Bi-POS}|$$

$$\boldsymbol{\mathcal{F}}^*_{UniBi-POS-PCA} = \{f^*ub_1, f^*ub_2, \dots , f^*ub_{|\boldsymbol{\mathcal{F}}_{UniBi-POS-PCA}|}\}$$

$$|\mathcal{F}^*_{UniBi-POS-PCA}| \ll |\mathcal{F}_{UniBi-POS}| \quad (16)$$

As mentioned earlier, $\mathcal{F}_{\text{UniBi-POS}}$ feature space is union of Uni-POS and Bi-POS features while $\mathcal{F}^*_{\text{UniBi-POS-PCA}}$ feature space includes new features extracted from $\mathcal{F}_{\text{UniBi-POS}}$ in which it is not clear what percentage of the $\mathcal{F}_{\text{Uni-POS}}$ and $\mathcal{F}_{\text{Bi-POS}}$ has been extracted, since PCA is unsupervised method and chooses optimal features based on variance measurement. The second approach will also be determined as follows:

$$\mathcal{F}_{Uni-POS} = \{fu_1, fu_2, \dots, fu_{|\mathcal{F}_{Uni-POS}|}\}$$

$$\mathcal{F}_{Bi-POS} = \{fb_1, fb_2, \dots, fb_{|\mathcal{F}_{Bi-POS}|}\}$$

$$\mathcal{F}^*_{Uni-POS-PCA} = \{f^*u_1, f^*u_2, \dots, f^*u_{|\mathcal{F}_{Uni-POS-PCA}|}\} \quad (17)$$

$$\mathcal{F}^*_{Bi-POS-PCA} = \{f^*b_1, f^*b_2, \dots, f^*b_{|\mathcal{F}_{Bi-POS-PCA}|}\} \quad (18)$$

$$\mathcal{F}^*_{Uni-Bi-POS-PCA} = \mathcal{F}^*_{Uni-POS-PCA} \cup \mathcal{F}^*_{Bi-POS-PCA} \quad (19)$$

$$\mathcal{F}^*_{Uni-Bi-POS-PCA} = \{f^*u_1, f^*u_2, \dots, f^*u_{|\mathcal{F}_{Uni-POS-PCA}|}, f^*b_1, f^*b_2, \dots, f^*b_{|\mathcal{F}_{Bi-POS-PCA}|}\}$$

$$|\mathcal{F}^*_{Uni-Bi-POS-PCA}| = |\mathcal{F}^*_{Uni-POS-PCA}| + |\mathcal{F}^*_{Bi-POS-PCA}|$$

$$|\mathcal{F}^*_{Uni-Bi-POS-PCA}| \ll |\mathcal{F}_{UniBi-POS}| \quad (20)$$

As can be seen in equations 17-20, $\mathcal{F}^*_{\text{Uni-POS-PCA}}$ and $\mathcal{F}^*_{\text{Bi-POS-PCA}}$ feature spaces include the optimal features extracted using PCA from $\mathcal{F}_{\text{Uni-POS}}$ and $\mathcal{F}_{\text{Bi-POS}}$, respectively. Unlike $\mathcal{F}^*_{\text{UniBi-POS-PCA}}$, $\mathcal{F}^*_{\text{Uni-Bi-POS-PCA}}$ feature space comprises of a certain percentage of each feature spaces $\mathcal{F}^*_{\text{Uni-POS-PCA}}$ and $\mathcal{F}^*_{\text{Bi-POS-PCA}}$ based on two criterions (proportion of variance and Kaiser's rule).

Table 3. The Maximum Accuracy Obtained for Different Feature Spaces $\mathcal{F}^*_{UniBi-POS-PCA}$, $\mathcal{F}^*_{Uni-POS-PCA}$, and $\mathcal{F}^*_{Bi-POS-PCA}$ using RBF SVM

Feature spaces	Number of features	Memory usage (MB)	RBF SVM parameters		Train set		Test set	
			C	gamma	Accuracy	Training time	Accuracy	Testing time
UniBi_B_PCAKi	1605	23.5	10	0.01	96.72±0.24	5.13±0.13	64.00±2.24	0.57±0.01
Uni-TF-PCACUM90	866	12.7	10	0.001	92.06±0.81	2.75±0.84	**66.93±2.91**	0.28±0.01
Bi-TF-PCACUM90	874	12.8	1	0.001	83.74±0.34	2.88±0.10	64.06±1.89	0.32±0.01

Table 4. Sentiment Classification Accuracy for Merge Uni-POS-PCA-based and Bi-POS-PCA-based Feature Spaces using RBF SVM

Feature spaces	Number of features	Memory usage (MB)	RBF SVM parameters		Train set		Test set	
			C	gamma	Accuracy	Training time	Accuracy	Testing time
Uni-POS-B-PCACUM90 **Bi-POS-B-PCACUM90**	1810 919,891	26.5	10	0.001	97.50±0.20	5.58±0.06	**67.19±2.73**	0.62±0.01
Uni-POS-B-PCAKi **Bi-POS-B-PCAKi**	2539 1303,1236	37.2	10	0.001	98.65±0.16	8.26±0.03	64.06±2.10	0.92±0.01
Uni-POS-TF-IDF-PCACUM90 **Bi-POS-TF-IDF-PCACUM90**	2180 1103,1077	31.9	10	0.0001	95.45±0.33	6.65±0.05	66.51±2.07	0.74±0.01
Uni-POS-TF-IDF-PCAKi **Bi-POS-TF-IDF-PCAKi**	2958 1502,1456	43.3	10	0.0001	97.69±0.28	9.24±0.91	63.54±2.84	1.06±0.01
Uni-POS-TF-PCACUM90 **Bi-POS-TF-PCACUM90**	1740 866,874	25.5	10	0.001	96.71±0.22	5.34±0.03	65.94±3.61	0.6±0.00
Uni-POS-TF-PCAKi **Uni-POS-TF-PCAKi**	2478 1258,1220	36.3	10	0.0001	95.34±0.17	7.68±0.05	65.05±2.14	0.85±0.01

Let consider "PCACUM90" and "PCAKi" to refer to portion of variance 90% (Equation 3) and Kaiser's rule (Equation 2) as a threshold, respectively. For instance, Uni-POS-TF-PCACUM90 feature space includes 90

percentages of the transformed unigrams with TF feature weighting method by PCA method and UniBi-POS-B-PCAKi feature space refers to the transformed UniBi features along with binary feature weighting method by PCA with considering Kaiser's rule as criterion and likewise for other combinations in Tables 3 and 4.

- **The Impact of the dimensionality reduction on financial news sentiment classification in low frequency-based feature spaces model (NgramPOS model)**

Table 3 shows the maximum accuracy obtained for different feature spaces $\mathcal{F}^*_{UniBi-POS-PCA}$, $\mathcal{F}^*_{Uni-POS-PCA}$, and $\mathcal{F}^*_{Bi-POS-PCA}$. The accuracy results indicate a high correlation between Uni-POS and Bi-POS features separately that has led to the improvement of sentiment classification performance for $\mathcal{F}^*_{Uni-Bi-POS-PCA}$ feature space. The impact Uni-POS-PCA features is more because they have higher variance, where PCA indicate the importance of features base their variance.

The sentiment classification accuracy result (67.19±2.73) for the feature space of $\mathcal{F}^*_{Uni-Bi-POS-PCA}$ (merge Uni-POS-B-PCACUM90 and Bi-POS-B-PCACUM90) in Table 4 in comparison with the classification accuracy Uni-TF-PCACUM90 (66.93±2.91) and UniBi-B-PCAKi (64.00±2.24) in Table 3, implies to the fact that despite the low frequency of Bi-POS features, these features can be used to increase the accuracy of document classification as an effective feature. Since $\mathcal{F}^*_{Uni-Bi-B-POS-PCA}$ includes the combination of Uni-POS-B and Bi-POS-B after applying the PCA method separately, while UniBi-B-PCAKi feature space created by applying PCA over the combination of unigrams and bigrams features.

Therefore, the merge the two transformed feature spaces ($\mathcal{F}^*_{\mathrm{Uni-POS-PCA}}$, $\mathcal{F}^*_{\mathrm{Bi-POS-PCA}}$) produced by PCA model, constructs a new feature space that is used for RBF SVM and provides a promising result since it keeps optimal features from Uni-POS-B and Bi-POS-B features spaces.

4. Conclusion

This study has proposed an effective feature selection model for sentiment classification of financial news which is able to enhance the performance of feature processing in Ngram-based models. NgramPOS model employs a combination of statistical and linguistic approaches to extract sentiment information as features in order to classify financial news. This low frequency-based model extracts the combination of sentiment-rich words and phrases as unigrams and bigrams (Uni-POS, Bi-POS) using the defined POS-based fixed patterns along with the binary weighting method while applying Principle Component Analysis (PCA) method to reduce dimension of the extracted feature space. The experiment results indicated that the NgramPOS model can extract the important Bi-POS features with low frequency to classify financial text with promising accuracy rate. The focus of this research is on enhancing the performance sentiment classification of financial news by extracting sentiment and opinion words. Hence, this study can be extended by applying on automatic trading systems which utilize financial news as a parameter to predict stock market.

Acknowledgment

This work is supported in partial by Universiti Tun Hussein Onn Malaysia.

REFERENCES

[1] E. Fama, "Random Walks in Stock Market Prices," *Financ. Anal. J.*, vol. 21, no. 5, pp. 55–59, 1965.

[2] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *J. Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[3] F. Li, "Textual Analysis of Corporate Disclosures: A Survey of the Literature," 2011.

[4] M. Koppel and I. Shtrimberg, "Good News or Bad News ? Let the Market Decide," *Comput. Attitude Affect Text Theory Appl. Inf. Retr.*, vol. 20, pp. 297–301, 2006.

[5] S. S. Groth and J. Muntermann, "An intraday market risk management approach based on textual analysis," *Decis. Support Syst.*, vol. 50, no. 4, pp. 680–691, 2011.

[6] Y. Yu, W. Duan, and Q. Cao, "The impact of social and conventional media on fi rm equity value : A sentiment analysis approach," *Decis. Support Syst.*, vol. 55, no. 4, pp. 919–926, 2013.

[7] M. Généreux, T. Poibeau, and M. Koppel, "Sentiment Analysis Using Automatically Labelled Financial News Items," in *Affective Computing and Sentiment Analysis*, vol. 45, no. 2, the series Text, Speech and Language Technology, Springer, 2011, pp. 101–114.

[8] J. J. Zhai, N. Cohen, and A. Atreya, "CS224N Final Project : Sentiment analysis of news articles for financial signal prediction," pp. 1–8, 2011.

[9] V. Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?," *Comput. Math. with Appl.*, vol. 65, no. 10, pp. 1427–1437, 2013.

[10] M. Hagenau, M. Liebmann, and D. Neumann, "Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features," *Decis. Support Syst.*, vol. 55, pp. 685–697, 2013.

[11] K. Joshi, B. H. N., and Jyothi Rao, "STock Trend Prediction Using News Sentiment Analysis," *CoRR*, vol. abs/1607.0, 2016.

[12] M. Y. Chen and T. H. Chen, "Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena," *Futur. Gener. Comput. Syst.*, no. 2011, 2017.

[13] S. W. K. Chan and M. W. C. Chong, "Sentiment analysis in financial texts," *Decis. Support Syst.*, vol. 94, no. 2017, pp. 53–64, 2017.

[14] A. Mayne, "Sentiment Analysis for Financial News," University of Sydney, 2010.

[15] S. Foroozan Yazdani, M. A. A. Murad, N. M. Sharef, Y. P. Singh, and A. R. A. Latiff, "Sentiment Classification of Financial News Using Statistical Features," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 3, p. 34, 2016.

[16] T. Pederson, "A Desision Tree of Bigrams is an Accurate Predictor of Word Sence," in *proceeding of the second NAACL*, 2001, pp. 79–86.

[17] K. Dave, I. Way, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery : Opinion Extraction and Semantic Classification of Product Reviews," pp. 519–528, 2003.

[18] Y. Mejova and P. Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 546–549.

[19] M. L. M. Lan, C. L. T. C. L. Tan, J. S. J. Su, and Y. L. Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 721–735, 2009.

[20] N. Pham Xuan and H. Le Quang, "A New Improved Term Weighting Scheme for Text Categorization," *Adv. Intell. Syst. Comput.*, vol. 271, pp. 261–270, 2014.

[21] F. Sebastiani, "Machine Learning in Automated Text Categorization," 2001.

[22] A.-I. Petrişor, I. Ianoş, D. Iurea, and M.-N. Văidianu, "Applications of Principal Component Analysis Integrated with GIS," *Procedia Environ. Sci.*, vol. 14, pp. 247–256, 2012.

[23] E. Alpaydin, *Introduction to machine learning*, Second. 2010.

[24] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.

[25] T. Joachims, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features," *Proc. 10th Eur. Conf. Mach. Learn. ECML '98*, pp. 137–142, 1998.

[26] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods," *Knowledge-Based Syst.*, vol. 0, pp. 1–14, 2017.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[28] B. Lantz, *Mahine Learning with R*. Birmingham B3 2PB, UK.: Packt Publishing Ltd, 2013.

[29] B. Schölkopf and A. Smola, "Support Vector Machines and Kernel Algorithms.," 2005, pp. 1–22.

[30] H. S. Ooi, G. Schneider, T. Lim, Y. Chan, B. Eisenhaber, and F. Eisenhaber, "Data Mining Techniques for the Life Sciences," vol. 609, pp. 129–144, 2010.

[31] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," *Bioinformatics*, vol. 1, no. 1, pp. 1–16, 2010.

[32] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[33] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, vol. 14, no. 12, pp. 1137–1143, 1995.

[34] A. Taylor, M. Marcus, and B. Santorini, "the Penn Treebank: an Overview," *Treebanks*, pp. 5–22, 2003.

[35] F. Benamara, C. Cesarano, and D. Reforgiato, "Sentiment Analysis : Adjectives and Adverbs are better than Adjectives Alone," *Proc Int Conf Weblogs Soc. Media*, pp. 1–4, 2007.

[36] P. D. Turney, "Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews," *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 417–424, 2002.

[37] V. Hatzivassiloglou, K. R. McKeown, B. Pang, L. Lee, S. Vaithyanathan, L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Predicting the Semantic Orientation of Adjectives," *ACM Trans. Inf. Syst.*, vol. 21, no. 4, pp. 315–346, 2009.

[38] J. Han and M. Kamber, *Data Mining (Concepts and Techniques)*. Elsevier (Morgan Kaufmann), 2006.