Evolutionary Trading Signal Prediction Model Optimization based on Chinese News and Technical Indicators in the Internet of Things

Chun-Hao Chen, Ping Shih, Gautam Srivastava, Shih-Ting Hung, and Jerry Chun-Wei Lin*

***Abstract*—The Internet of Things technologies are essential in deploying successful IoT-based services, especially in the financial services sector in recent years. Stock market prediction which could also be an IoT-based service is a very attractive topic that has inspired countless studies. Using financial news articles to forecast the effect of certain events, understand investors' emotions, and react accordingly has been proved viable in existing pieces of literature. In this study, we utilized Chinese financial news in an attempt to predict the stock price movement and to derive a trading strategy based on news factors and technical indicators. Firstly, the Stock Trend Prediction (STP) approach is proposed. It first extracts keywords from the given articles. Then, the 2-word combination is employed to generate more meaningful keywords. The feature extraction and selection are followed to obtain important attributes for building a trading signal prediction model. Also, to make the trading signal more reliable, the technical indicators are considered to confirm the trading signal. Because the hyperparameters for the STP and technical indicators will have influenced the final results, an enhanced approach, namely the genetic algorithm (GA)-based Stock Trend Prediction (GASTP) approach, is then proposed to find hyperparameters automatically for constructing a better prediction model. Experiments on real datasets were also made to show the effectiveness of the proposed algorithms. The results show that the GASTP performs better than the buy-and-hold strategy as well as the STP.**

***Index Terms*—Genetic algorithm, Chinese news mining, trading strategy, technical indicators, expected fluctuation analysis.**

I. Introduction

The Internet of Things (IoT) technologies [13] are essential in deploying successful IoT-based services, especially in the financial services sector, including insurance, banking, and investments [8], [23]. Financial market forecasting [15] has long been an interesting subject that inspired prolific researches, and stock price prediction [42], [43] which could also be an IoT-based service is most of all [3], [5], [7], [17], [22]. The ability to predict stock market trends, even if only slightly better than the market average, will garner incredible gain in the trading market. Hence, over the past few years, researchers, financial experts, trading specialists, and other enthusiasts have dedicated an enormous amount of time and resources in discovering and optimizing ways to accurately predict the bullish, bearish trends of the stock market. However, unsurprisingly, market trends are extremely difficult, if even possible, to predict.

In general, two major "philosophies" categorized the way traders assess equities and market indexes. Namely fundamental analysis and technical analysis. By and large, the two are distinguished by the input data taken into consideration. Fundamental analysis focuses on the performance and financial well-being of a target company, taking in company-released financial statements, reports, as well as financial or general news regarding the company, its industry, or even on a larger scale, statistical figures of countries and global development. On the other hand, technical analysis assumes that the stock prices and market trends have in themselves a certain cycle within a certain period which can be deduced via close observation of their historical prices and trading volumes. Albeit different the two approaches, often times, skillful traders adapt both philosophies to form their trading strategies.

A. Motivation

Abundant researches have been made for technical analysis approach in the past owing to the higher availability of historical trading data, whereas fundamental analysis presents tougher challenges due to the often unstructured and noisy quality of the data. The reasons for this paper to present the stock trend prediction approaches are stated as follows:

1) The market is a dynamic organism where people engage and participate, therefore human emotions, greed, overconfidence, fear of loss, or even deliberate manipulation of such, play an important part, and financial news is the primary conduit from which retail investors obtain market information. It will be unfair to disregard such significant components of the market.
2) Making investment decisions based solely on news reports may not be a very prudent idea. News reports by their nature are inherently hind sighted and financial news can be prone to manipulation [4]. Individuals who are not trading professionals or gigantic market tycoons who could have obtained inside information or

Chun-Hao Chen and Shih-Ting Hung are with Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan. Email: chchen@ntut.edu.tw, t109749005@ntut.org.tw

Ping Shih is with Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan. Email: ryanjshih@gmail.com

G. Srivastava is with the Department of Mathematics & Computer Science, Brandon University, Canada. Email: Srivastavag@brandonu.ca and also with the Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan.

J. Chun-Wei Lin is with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. Email: jerrylin@ieee.org (*Corresponding author)

singlehandedly influence market trends are subject to the information provided to them through newspapers, online stock forums, and most of all, financial news websites. Many existing studies have suggested that sentiment, keywords, and events mentioned in news articles have a reasonable effect on traders' decision-making [27], [33], [39] and have tried and succeeded in developing trading strategies that incorporate the influence of financial news [11], [18], [32], [37].

3) In existing pieces of literature, studies have been conducted majorly using one or many of the three sources as their experimental corpus namely financial news, corporate announcements, and social media contents. The three are quite different in their content as well as in their reaction speed and content authors. Therefore, it is very interesting and insightful to compare corpus from diverse sources. However, not many of them have taken into account the period in which the articles are published comparing to the peak-trough cycle of the stock price in which the target equity is situated. Identifying the situation period during which certain positive/ negative news message is the most efficacious can be very helpful for individual traders to avoid being manipulated and make profits using this newly obtained insight.

B. Contributions

In this study, we attempt to predict stock movement using Chinese financial news, which we have obtained from one of the popular Taiwanese financial news websites, and along with technical indicators to determine the optimal time for trading. The contributions are listed as follows:

1) Firstly, we proposed the stock trend prediction (STP) algorithm based on test mining for financial news and discovered that the settings of hyperparameters played an important role in the prediction accuracy and simulated trading returns.
2) Hence we then applied the genetic algorithm (GA), namely the GA-based stock trend prediction (GASTP) approach, for finding appropriate hyperparameters for constructing the prediction model.
3) Finally, experiments on the real datasets were made to show the effectiveness of the proposed approaches. By using the proposed approaches, they can also be applied to various engineering applications, e.g., trading robot [24], pair trading [14], exchange rate prediction [38].

The rest of the paper is organized as follows. Section II presents the literature review; Section III describes the design and implementation of the proposed approaches; Section IV contains the result and analysis of the experiment; Conclusion and recommendation for future work are given in Section V.

II. Literature Review

According to the efficient-market hypothesis of Fama [9], the market is "informationally efficient". It suggests that once the information is made available, the market will efficiently reflect its underlying value and until any new information is released, the price fluctuates completely randomly. On this premise, one cannot consistently achieve returns greater than the market average. However, Fama later revised this theory to include three states of efficiency: strongly efficient, semi-strongly efficient, and weakly efficient [10]. When the market is strongly efficient, the information is completely available to all of the investors, a market prediction is unachievable. Nevertheless, this strongly efficient market is highly theoretical and rarely the case in real life. Additionally, the market efficiency is not always "static". Lim *et al.* examined the market efficiency of eight Asian countries during the 1997 financial crisis [21]. They found that the financial crisis adversely affected the market efficiency of most Asian markets, but the market efficiency recovered during the post-crisis period. In a recent study, Wu *et al.* conducted an empirical analysis on the predictability of the Taiwan Stock Exchange using news reports from 2008 to 2014 and found that news variables are useful for predicting the stock returns [41]. Boudoukh *et al.* classified trading days into "news days", "no news days" based on whether or not any related news is published on that day and, depending on the identification of any particular event, further classified news into "event identified" and "no event identified" [1]. They found that when news articles are identified as an event, it has an effect on the price, and consequently, price variances are higher when news occurs. They also discovered that price reversal happens on "no news days" while price continuation happens on news days. Schumaker *et al.* classified news articles into subjective news and objective news and found out that subjective news articles have better predictive power than objective ones [39].

Utilizing textual data to predict the market is not a counter-intuitive idea. Most investors take news articles, as well as online forums, blog posts, etc. into account when making investment decisions. Sources of corpus include mandatory corporate reports, traditional newspapers, social media contents, etc. Yu *et al.* examined the effectiveness of conventional and social media on firm equity values [44]. Their study included different sources of the corpus from various outlets like newspapers, business magazines, television broadcasts, Twitter tweets, blogs, and forum posts, and compare their usefulness and found out that social media has a stronger relationship with firm stock performance than conventional media. Yang *et al.* made use of the faster information speed of Twitter tweets and combined it with the reliability of financial news to develop a trading strategy based on the derived sentiment feedback strength [45]. Li compared different sources of financial news in Taiwan and found out that different sources of financial news had significantly different effects on investors' decision-making [20]. Other studies mainly apply one source of the financial corpus. The sources and the period length of a financial corpus in past studies are shown in Table I.

There are three main disciplines in this field of study, namely feature extraction, feature selection, and machine learning algorithm. Most of the existing pieces of literature fall into one or more of these categories. In the following, literature about feature extraction is described. Textual data by its nature is populated with noises and syntactical auxiliaries

TABLE I: Sources and length of financial corpus.

Reference	Text Type	Period Length
Schumaker *et al.*, 2012 [39]	Financial news	1 month
Boudoukh *et al.*, 2013 [1]	Financial news	10 years
Hagenau *et al.*, 2013 [18]	Corporate announcement	14 years
Yu *et al.*, 2013 [44]	Multiple sources	3 months
Kim *et al.*, 2014 [19]	Financial news	1 year
Li *et al.*, 2014 [27]	Financial news	1 year
Li *et al.*, 2014 [26]	Financial news, social media	3 months
Nassirtoussi *et al.* [29]	Financial news	4 years
Nuij *et al.*, 2014 [32]	Financial news	5 months
Nguyen *et al.*, 2014 [33]	Social media	1 year
Lee, 2016 [20]	Financial news	2 years
Yang *et al.*, 2017 [45]	Financial news, social media	3 years

that do not possess much predictive power. The purpose of feature extraction is to extract the terms or sentences and turn them into useful "features" before putting them through any machine learning algorithms.

Schumaker *et al.* examined several more advanced feature extraction techniques namely noun-phrase, named entity, and proper noun [37], [39]. Noun phrase feature extraction uses part-of-speech tags to identify the noun phrases in the corpus and choose them as features. Named entity refers to certain categories of entities as the candidate to be used as features. They applied the so-called MUC-7 framework of named entities that comprise date, time, location, money, percentage, organization, and person. Later they considered the results generated by the two above-mentioned feature extraction techniques and further derived a new method called proper noun, which is a subset of a noun phrase and a superset of the named entity. The results showed that proper noun possesses better predictive power. Hagenau *et al.* used Bag-of-Words, noun phrase, and 2-Word Combination, N-gram and found out that the performance of 2-Word Combination outperformed other methods of feature extraction because it succeeded in capturing word combinations that when used together connotes different are more precise meaning than when used alone [18].

Nassirtoussi *et al.* invented a heuristic-hypernyms feature-selection algorithm to improve the feature extraction performance on news headlines for FOREX price prediction [31]. Nguyen *et al.* compared human-selected sentiment, SVM-classified sentiment, and LDA-based sentiment along with their proposed method: JST-based sentiment extraction and aspect-based sentiment extraction [33]. The result shows that while aspect-based sentiment extraction performs best, the human-selected sentiment is at a very close second. Li *et al.* extracted proper nouns and "emotion words" from financial news and social media [26]. They examined finance-specific emotion words along with general emotion words and found out that finance-specific sentiment dictionary well-performed general sentiment dictionary and when used together with a proper noun, finance-specific dictionary performs better. Nuij *et al.* used proprietary software to extract events from news articles and evaluated the effectiveness of different events [32]. Kim *et al.* used a self-built sentiment dictionary and evaluates the sentiment score of each financial news [19].

In conclusion, feature extraction generally falls into three categories, namely (1) Syntactical features, including Bag-of-Words, N-gram, noun phrase, proper noun, word combination; (2) Sentiment features, including human-selected sentiment tags, dictionary-based sentiment (general usage, finance-specific), and other derived measurements of sentiment score; (3) Specific subjects, including events and key-word mentions. Following the feature extraction phase is feature selection or dimensionality reduction. Some of the features are ready to be put into the learning process after extracted from the original corpus, especially features like events, key-word mentions, and sentiments. But, some of the features are still too large in number for machine learning algorithms, especially those generated by syntactical means of feature extraction like Bag-of-Words, noun phrase, word combination, etc., and therefore necessitates the process of selecting the features that are more significantly co-related with the prediction target.

Schumaker *et al.* used minimum document frequency to select only the features that appeared in at least a certain number of documents from the entire corpus to eliminates features that are too rare [39]. Hagenau *et al.* compared multiple methods of feature selection including a benchmark Bag-of-Word, dictionary-based features, and two exogenous-feedback-based feature selections namely, Chi-Square and Bi-Normal-Separation [18]. Their findings suggest that exogenously given feature selection outperformed other methods of feature selection. Nassirtoussi *et al.* proposed a synchronous targeted feature reduction (STFR) to reduce the number of features and predict FOREX price trends using financial news headlines [31]. Vu *et al.* used pre-defined company-related keywords, named entity recognition based on linear conditional random fields (CRF) [40].

In conclusion, there are no major "schools of thought" in feature selection, partly because this phase is highly dependent on the targeted outcome and is very domain-specific. Many researchers proposed their way of feature selection intending to generate better results. Once the textual features are ready, they will be put through one or more machine learning algorithms to learn their patterns and predict future outcomes. By far, the most popular machine learning algorithm when it comes to stock trend prediction is Support Vector Machine (SVM) [17], [18], [27], [31], [33], [37], [39]. For example, Hao *et al.* proposed the SVM with fuzzy hyperplane approach to find the hidden topic model and emotional information from news articles for stock trend prediction [17]. Another popular algorithm used in this domain is Naïve Bayes, e.g., Yu *et al.* use the Naïve Bayes approach to predict various sources of textual data [44]. In recent years, many algorithms based on deep learning have also been proposed for stock trend prediction [3], [5], [22]. Considering both stock market information and individual stock information, Chen *et al.* proposed the graph convolutional feature-based convolutional neural network (GC–CNN) model to predict stock trend [5]. Long *et al.* proposed an integrated framework for the prediction of stock price trends using the deep learning and knowledge graph [22]. To explore the public mood and emotion from articles, Chen *et al.* designed a Long Short-Term Memory (LSTM) network for stock market trend prediction [3]. Besides, considering the sentiment of stock social media, Derakhshan *et al.* proposed a part-of-speech graphical

model for extracting user's opinion for stock price movement prediction [7].

Most of the existing pieces of literature are done on English corpus. However, the pre-processing phase of text mining differs greatly among different languages. For example, one of the popular feature extraction methods, noun phrases, cannot be easily applied to Chinese text due to the syntactical nature of the Chinese language.

III. Our Proposed Approaches

This section depicts the two proposed approaches. The Stock Trend Prediction (STP) approach based on traditional Chinese news and technical indicators is described in Section III-A. The GA-based Stock Trend Prediction (GASTP) approach is described in Section III-B.

A. Proposed Stock Trend Prediction Approach

The flowchart of the designed STP is shown in Fig. 1.

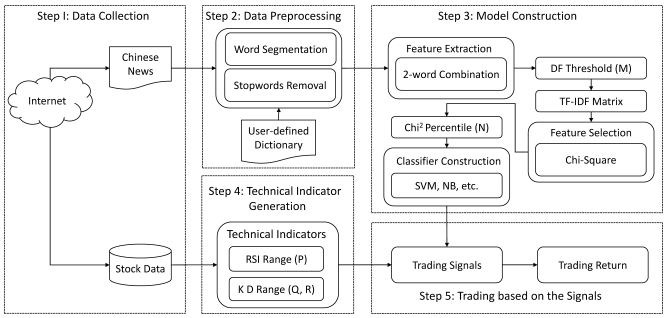

Fig. 1: The flowchart of the STP approach.

The flowchart can be divided into five steps. In Step 1, the stock data and Chinese news are collected from the Internet. Once the news corpus is ready, it segments the corpus into individual terms and removes stop words in Step 2. The Jieba is used to do text segmentation in this study. The reason is that it is equipped with the capability of incorporating a user-defined dictionary. The Chinese language is constantly evolving and changing, and the parlance in Traditional Chinese is not the same as in Simplified Chinese. Terms and "acronyms" are not the same among different domains. Hence, it is preferable to use a domain-specific Traditional Chinese user-defined dictionary in this study. The said dictionary is manually edited by observing Jieba's preliminary segmented results and if there are any miss-segmented terms, we add them to the dictionary. Each entry in the dictionary consists of one term and its weight.

In Step 3, the model construction is done as follows. The 2-word combination which is used in [18] as a feature extraction method is applied to the corpus. It is an extension of N-gram feature extraction, which captures word combinations that has a distance higher than zero. It can capture word combinations that separately connote less predictive power but possess high predictive power in pairs. We set the maximum distance at five, the same as used in [27], and examine the effectiveness of this method used on traditional Chinese text.

We choose features with a minimum document frequency of n to construct a TF-IDF model. Traditional sentiment analysis methodology maps the terms in the corpus with a predefined sentiment dictionary, usually segmented into positive, negative, and/or neutral, but whether it's a general usages sentiment dictionary like Harvard-IV-4 dictionary or finance specific sentiment dictionary like Loughran & McDonald financial dictionary, they are prone to human bias and can only apply on English corpus. We want to have the positive/negative value of the terms exogenously given so we label each article as positive/negative based on the price change of $t + d$ where t is the date that the article is released and d is the number of days we intend to hold said stock. If the closing price of $t + d$ is greater than the opening price of t, we label the article as positive otherwise as negative.

After a 2-word combination is applied, the number of features will be extremely large and unwieldy, hence the Chi-square feature selection method is employed to reduce the number of features and only retain features with the highest m percent Chi-square score. A higher Chi-square score suggests that the term appears more in one class (whether positive or negative) and less in the other.

The minimum document frequency threshold will determine the specificity of the features selected and Chi-square score percentage the relevancy and predicting power of the features. These two hyper-parameters both contribute to the number of features selected for training. In the best case, we would like to select the lowest number of features to reduce noise and the most relevant for the accuracy of the model. Hence, this presents an optimization problem that we adopt the GA to solve, which we are discussing in detail in the following section.

Next, the generated dataset is utilized to construct the classifier. We adopt two classifiers, namely SVM with RBF kernel and multinomial Naïve Bayes, because they are wildly used in previous studies, and we would like to examine its performance on traditional Chinese text. The classifier classified each article into two categories: positive and negative, and the outcome means the prediction of the stock movement for the following d days. In Step 4, to see in which period the textual features are the most efficacious, we also put technical indicators, the Relative Strength Index (*RSI*) [2] and *KD* [30], to determine the trading signals. They are described as follows.

The RSI illustrates the strength of stock in its current movement. The formula for RSI is given in Eq. 1 and Eq. 2.

$$RS = \frac{\text{Avg. gain}}{\text{Avg. loss}}, and \tag{1}$$

$$RSI = 100 - \frac{100}{1 + RS}. \tag{2}$$

RSI ranges between 0 to 100. When a *RSI* score is at 50, it indicates that the rises and falls of a certain stock are roughly the same; a *RSI* score higher than 80 indicates overbought and possible downturn; a *RSI* lower than 20 indicates oversold and possible upturn.

Raw Stochastic Value (*RSV*) is the preliminary step of calculating the *KD* value of a certain stock. It ranges from 0 to

100. The higher the *RSV*, the higher the closing price is within the specified period therefore indicates possible overbought of the aforementioned stock and vice versa. The formula for *RSV* is given in Eq. 3 where *n* is the period range and $Close_t$ denotes the closing price of the day *t*, $High_{[t-n+1,t]}$ denotes the highest price between $t-n+1$ to t. The same goes for $Low_{[t-n+1,t]}$.

$$RSV_t = \frac{Close_t - Low_{[t-n+1,t]}}{High_{[t-n+1,t]} - Low_{[t-n+1,t]}}. \quad (3)$$

Once we have the *RSV* ready, we can calculate the *K* value by the Eq. 4 and the *D* value by the Eq. 5.

$$K_t = \frac{2}{3} \times K_{t-1} + \frac{1}{3} \times RSV_t. \quad (4)$$

$$D_t = \frac{2}{3} \times D_{t-1} + \frac{1}{3} \times K_t. \quad (5)$$

Normally, the value of *K* and *D* follows the overall price trend while *K* moves faster and sharper than *D*. Although there are many ways of interpreting the *KD* value, it is commonly construed that a *KD* value higher than 80 indicates an overbought at a high relative price, and a *KD* value lower than 20 indicates an oversold at a low relative price. At last, in Step 5, the trading signals of the classifier and the two technical indicators are employed for trading. The pseudo-code for the STP approach is given in Algorithm 1.

Lines 1 to 5 describe the text preprocessing phase where unstructured articles are segmented, cleaned, and reorganized in 2-Word Combination. Lines 6 to 9 describe the removal of too specific features by setting a minimum document frequency threshold. Lines 10 to 15 calculate the Chi-square score for every remaining feature and retain only those within the highest *n*%. Lines 16 to 18 split the dataset into training and testing datasets and make a prediction using the classifier. Lines 19 to 26 make simulated trading based on the predicted movement and technical indicators.

B. GA-Based Stock Trend Prediction Approach

The GASTP is an enhanced version of the STP, which GA to obtain hyperparameters for optimizing the simulated trading return. The pseudo-code of GASTP is given in Algorithm 2.

Line 1 sets the initial population where it contains *pSize* chromosomes. Lines 2 to 14 describes the entirety of the evolution process, which it will run for *geneNum* of times. Lines 3 to 6 decode the picked chromosome into hyperparameters and call the STP algorithm for fitness evaluation. Lines 7 to 9 perform the crossover and mutation. Lines 10 to 13 select the chromosome to be passed to the next population. When reaching the termination condition, the best chromosome will be outputted. The encoding scheme is a binary string of twenty-five digits in the GASTP as shown in Fig. 2.

It uses the first five digits as minimum document frequency more *m*, digits 6 to 10 as the percentage *n*% of Chi-square score. The last fifteen digits are used as *RSI*, *K*, and *D* scores. In other words, m_1 to m_5 represents the minimum document frequency threshold, n_1 to n_5 the top percentage number of Chi-square score, p_1 to p_5 sets the range of *RSI*, q_1 to q_5 the range of *K* value and finally r_1 to r_5 the range of *D* value. Note that not only the RSI and KD that are commonly used technical indicators but also other indicators, e.g., moving average (MA), Bollinger Bands (BBands), can be used in the GASTP to find trading signals.

It converts the binary string into decimal numbers. This means that *m* and *n* will be two integers between 0 and 32. In the case when *m* or *n* is zero, it will set them as one because 0% or a document frequency of zero does not make any sense. We multiply the obtained integer *p*, *q*, and *r* by 3.125 so they

Algorithm 1: Proposed STP algorithm

Input: A set of news articles, $N = \{N_0, N_1, \ldots, N_n\}$ with released data $T = \{t_0, t_1, \ldots, t_n\}$; stock moving direction of t_{n+d} is $D = \{d_0, d_1, \ldots, d_n\}$, *RSI* of t_n is $P = \{P_0, P_1, \ldots, P_n\}$, *K* value of t_n is $Q = \{Q_0, Q_1, \ldots, Q_n\}$, and *D* value of t_n is $R = \{R_1, R_2, \ldots, R_n\}$; *m* is document frequency; *n* is Chi-Square score percentile; *p* is *RSI*, *q* is *K*, and *r* is *D*.

Output: Simulated trading returns.

1 **for** *i = 1 to N* **do**
2 set segment N_i;
3 remove stopwords from N_i;
4 apply 2-word combination in N_i;
5 set *X* := vectorized *N* by TF-IDF ;
6 **for** *j = 0 to X* **do**
7 **if** *document frequency of feature[j]* $< m$ **then**
8 remove *feature*[*j*] from *X*;
9 **for** q *= 0 to X* **do**
10 $Chi2\_score$:= calculate Chi-Square socre of *feature*[q];
11 split (*X*,*D*) into 6.5:3.5 as training data;
12 testing data := X_{tr}, D_{tr}, X_{ts}, D_{ts};
13 classifier = $buildingClassifier(X_{tr}, D_{tr})//SVM, NB, etc.$;
14 predicted = $classifier(X_{ts})$;
15 **for** s *= 0 to* X_{ts} **do**
16 **if** *predicted[s] == upward P[s], Q[s] and R[s] are between p+10,q+10,r+10 and p-10, q-10, r-10* **then**
17 *cost* = buy TWD 100K of related stock by opening price of (*t*+1);
18 *earn* = sell all the stocks with closing price of (*t*+1);
19 *profit* = *earn* - *cost*;
20 return $SUM(\forall profit)$;

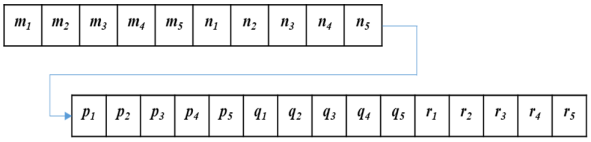

Fig. 2: Encoding scheme of GASTP.

Algorithm 2: Proposed GASTP algorithm

Input: *pSize*, popularion size; *pc*, crossover rate; *pm*, mutation rate; *geneNum*, number of generation; *trn*, tournament size.
Output: *bestChro*, best chromosome.

initial *population* := $initialPop(pSize)$;
for *iter = 0 to geneNum* **do**
  for *i = 0 to pSize* **do**
    m, n, p, q, r := covert *population*[*i*] to a set of 5 integers;
    call $STP(m, n, p, q, r)$;
  $nextPopulation \leftarrow crossover(pc, nextPopulation)$;
  $nextPopulation \leftarrow mutation(pc, nextPopulation)$;
  repeat to calculate fitness(*nextPopulation*);
  for *j = 0 to pSize* **do**
    select *trn* of chromosomes from population randomly;
    add chromosome with highest fitness score to *nextPopulation*;
bestChro = selectBestChro(*population*);

will be in the range between 3.125 and 100, and the *RSI*, *K*, and *D* range will be set to $p(q)(r) + 10$ to $p(q)(r) - 10$.

The fitness evaluation process consists of selecting features with a document frequency more than *m* and choose only the features with the highest *n%* of Chi-square score. If the machine learning algorithm predicts that the closing price of $(t + d)$ minus the opening price of $(t + 1)$ is positive, and *RSI* score of t is between $p - 10$ and $p + 10$ and $K(D)$ score of t is between $q(r) + 10$ and $q(r) - 10$, we purchase 100,000 TWD worth of stocks with the opening price of $(t + 1)$ and sell it with the closing price of $(t + d)$. The formula of total return is shown in Eq. 6:

$$TotReturn = \sum (P_{t+1} - P_t) \times \frac{100,000}{P_t}, \qquad (6)$$

where P_t denotes the closing price *P* of the date *t*. When a prediction returns positive and the *RSI*, *K*, *D* scores are between the boundaries set above, we multiply the total return with F1 score. The formula is shown in Eq. 7.

$$fitness(Cq) = TotReturn(C_q) \times F1(C_q), \qquad (7)$$

where the $F1(C_q)$ is the F1 score of chromosome C_q. The goal of the fitness function is to optimize the trading ROI, but if we simply put ROI score as the fitness score then the evolutionary process will target a few trades with high return and eliminate other profitable trading signals for it will "dilute" the already high ROI. To overcome this dilemma, we then multiply the total return with the F1 score, this way the effectiveness of a few extraordinary returns will be attenuated by the F1 score if it does not actually generate high predictive power but rather only works in a few situations. Another benefit of using this fitness function is that if we only targeted the F1 score, then the genetic algorithm will settle on chromosomes that achieve a high F1 score by making a mostly true prediction on a very scarce situation.

IV. Experimental Evaluation

This section presents the experimental results of the proposed approaches described in Section III. We conduct the experiments with real-world market data. Data description and the experimental environment is given in Section IV-A. The results for the STP are presented in Section IV-B and the GASTP in Section IV-C.

A. Data Description

Experiments were conducted with a real dataset to test the effectiveness of the proposed approaches. The news articles are captured from Wantgoo [1] which is one of the most popular stock market-related financial news websites in Taiwan. We collected stock market-related news articles starting from November 19, 2017, to January 30, 2019, and only choose articles that are related to six companies that are the subject of our experiment: Hon Hai Precision Industry Co., Ltd. (2317.TW), Yageo Corporation (2327.TW), Taiwan Semiconductor Manufacturing Company Limited (2330.TW), HTC Corporation (2498.TW), and Largan Precision Co., Ltd. (3008.TW), Catcher Technology Co., Ltd. (2474.TW). We only choose articles that are relating to at most five companies to avoid general trading market reports and commentaries. The six companies combined have an accumulated 1571 articles during the period. We have chosen these companies mainly for the reason that they generate the most news articles in this period. Coincidentally, they are all companies in the electronics sector, which may be a trait of the Taiwan Stock Exchange market. Stock price and volume data are downloaded from Yahoo Finance. We segmented news from these six companies into two folds. November 19, 2017, to July 31, 2018, as the training dataset, August 1, 2018, to January 30, 2019, as testing dataset. Summaries of the experimental datasets are given in Tables II and III.

TABLE II: Summary of the training data.

Training		**2017/11/19**	**2018/07/31**	**ROI**
Stock No.	**No. News**	**Price at start**	**Price at end**	**Buy and hold**
2317.TW	378	131.9	104.75	-20.58%
2327.TW	125	231.5	781	237.37%
2330.TW	234	240.5	246	2.29%
2498.TW	134	67.5	53.9	-20.15%
3008.TW	100	5680	5150	-9.33%
2474.TW	57	331	376.5	13.75%
Total	1028		**Average**	33.89%

B. Experimental Results for STP

We set a minimum document frequency of 15 with a Chi-Square percentile of 5 as they used in [18], [39]; *RSI*, *K*, and *D* are set to 20, 30, and 20, respectively. The RSI score below 20 is usually viewed as a signal of oversold and therefore a turning point of the stock trend and *K* above *D* at around 20

[1] https://www.wantgoo.com

TABLE III: Summary of the testing data.

Training		2018/08/01	2019/01/30	ROI
Stock No.	**No. News**	**Price at start**	**Price at end**	**Buy and hold**
2317.TW	169	104.9	70	-33.27%
2327.TW	51	790	321.5	-59.30
2330.TW	173	247	221	-10.53%
2498.TW	41	54.4	36.55	-32.81%
3008.TW	68	5240	3780	-27.86%
2474.TW	41	378.5	231.5	-38.84
Total	543		**Average**	-33.77%

signifies an uptick at a relatively low price. The experimental results of the STP framework are shown in Tables IV and V. The column "2-word" suggests whether we apply a 2-word combination or not. "2-word = FALSE" signifies Bag-of-Words minus stop words; the column "TechUse" suggests whether the technical indicators how the technical indicators should be used. "TechUse = AND" signifies that trade will only be made when *RSI*, *KD* are both within the appointed range; "TechUse = OR" signifies that a trade will be made when either RSI or *KD* is within the appointed range.

TABLE IV: Experimental results of STP algorithm with SVM without technical indicator.

STP								Testing Data				
								Without technical indicators				
Target	**2-word**	**TechUse**	**MDF**	**Chi%**	**RSI**	**K**	**D**	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
t + 1	TRUE	AND	15	5	20	30	20	72	0.57	0.2	-6963	-0.0010
t + 1	TRUE	OR	15	5	20	30	20	72	0.57	0.2	-6963	-0.0010
t + 1	FALSE	AND	15	5	20	30	20	47	0.57	0.13	-11925	-0.0025
t + 1	FALSE	OR	15	5	20	30	20	47	0.57	0.13	-11925	-0.0025
t + 5	TRUE	AND	15	5	20	30	20	116	0.55	0.29	-212390	-0.0183
t + 5	TRUE	OR	15	5	20	30	20	116	0.55	0.29	-212390	-0.0183
t + 5	FALSE	AND	15	5	20	30	20	87	0.54	0.19	-198751	-0.0228
t + 5	FALSE	OR	15	5	20	30	20	87	0.54	0.19	-198751	-0.0228
t + 10	TRUE	AND	15	5	20	30	20	101	0.6	0.2	-608992	-0.0603
t + 10	TRUE	OR	15	5	20	30	20	101	0.6	0.2	-608992	-0.0603
t + 10	FALSE	AND	15	5	20	30	20	93	0.62	0.22	-506569	-0.0545
t + 10	FALSE	OR	15	5	20	30	20	93	0.62	0.22	-506569	-0.0545
Average			15	5	20	30	20	86	0.58	0.21	-257598	-0.0300

TABLE V: Experimental results of STP algorithm with SVM with technical indicator.

STP								Testing Data				
								With technical indicators				
Target	**2-word**	**TechUse**	**MDF**	**Chi%**	**RSI**	**K**	**D**	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
t + 1	TRUE	AND	15	5	20	30	20	0	0.1	0	0	0.0000
t + 1	TRUE	OR	15	5	20	30	20	20	0.56	0.17	-2993	-0.0015
t + 1	FALSE	AND	15	5	20	30	20	0	0.1	0	0	0.0000
t + 1	FALSE	OR	15	5	20	30	20	16	0.56	0.14	-722	-0.0005
t + 5	TRUE	AND	15	5	20	30	20	4	0.5	0.29	-9170	-0.0229
t + 5	TRUE	OR	15	5	20	30	20	58	0.55	0.38	-117208	-0.0202
t + 5	FALSE	AND	15	5	20	30	20	2	0.5	0	-8781	-0.0439
t + 5	FALSE	OR	15	5	20	30	20	44	0.54	0.3	-87767	-0.0199
t + 10	TRUE	AND	15	5	20	30	20	3	0.4	0	-12979	-0.0433
t + 10	TRUE	OR	15	5	20	30	20	55	0.48	0.2	-353629	-0.0643
t + 10	FALSE	AND	15	5	20	30	20	2	0.5	0	-11003	-0.0550
t + 10	FALSE	OR	15	5	20	30	20	50	0.5	0.2	-296236	-0.0592
Average			15	5	20	30	20	21	0.44	0.14	-75041	-0.0355

Findings from the experimental results of the STP algorithm are discussed as the following. First, as the results show, the STP algorithm although generates negative returns, still significantly outperforms the buy-and-hold strategy. When we only hold for one day $(t+1)$, the losses in terms of ROI are almost unseen. Compared to the overall bearish trend during the testing period, the result is very satisfactory. However, as we hold more and more days $(t+5, t+10)$ the return gets worse. This is because the overall stock price is on a falling trend and holding a particular stock longer means one incorporates more trading days in this plummeting market and therefore results in worsening returns. It would be a little far fetched to presume that this outcome serves as proof that news factors are more effective on short-term stock price movement prediction than on long-term, but as far as the results suggest, we can say that positive news articles are more useful for short-term stock trend forecasting than for long-term trends in a bearish market.

Second, we examine the use of technical indicators, namely *RSI* and *KD*. The results show that the return, ROI, accuracy, and F1 Score are not significantly higher when news articles are used with technical indicators. In some case even worsen. When technical indicators are used strictly, meaning that RSI and KD must both situated within the specified range, the STP algorithm generates very few trades and as a result, the ROI would be determined by only a few trading returns hence become very biased. Nevertheless, when used none-strictly, meaning that a trade will be made when either *RSI* or *KD* is in the specified range, the effect is still not seen. Since the ranges of the technical indicators are set upon common knowledge and not any existing literature, it is understandable that it does not return a superb result. This proves that the ranges of the technical indicators should be tuned for them to help predict the stock trends. The tuning of technical indicators and other hyper-parameters will be discussed in later sections.

Third, we look at the effectiveness of a 2-word combination on this task. Upon examination, we can see that in terms of F1 score, the results obtained using 2-word combination does outperform those that obtained without 2-word combination when we only hold for a shorter period $(t+1, t+5)$, although very slightly, but when we hold longer, the effect is not seen. This can also serve as evidence that news factors have minimal effect on long-term stock price prediction, no matter what features are inputted. However, we also observed that the F1 score of $t+5$ is better than $t+1$, showing that news factors may be useful for more than merely intraday trades. Another observation obtained from examining the effect of a 2-word combination is that it consistently generates more numbers of trades than its counterpart does, even though it does not return better performance because of it. In an overall bearish market, more trade means more risk of getting a worsening return and this phenomenon does appear in the result.

In continuation of this topic, we examine the features captured by a 2-word combination. Table VI shows part of 2-word combinations that would not have been captured if the 2-word combination process were not applied.

The translation is roughly made by this study. If the two terms combined express any meaning, it is translated into one phrase; otherwise, it is translated separately into two terms. Upon closer examination, we can see that apart from a few noises, most of the terms captured makes good sense. However, if we trace back to the actual news articles, these combinations can almost all be captured by using 2-gram feature extraction. We tried to explain this finding as to the following. In previous study [18], examined the use of a 2-word combination on English text and found it able to help improve the prediction performance by capturing semantics with a distance higher than zero. Combinations like "dividend (goes) up" or "reduction (of) capital" will not be captured by 2-gram because there exists a certain preposition, conjuncture, or "helping verb" between the two terms that jointly possess significant meaning. On the other hand, Chinese text does not have this many uses of prepositions in between terms

TABLE VI: Features captured by 2-word combinations.

Feature	Chi^2 Score	Translation
('現金','減資')	1.0675	(Reduction of cash capital)
('a','股')	0.7519	(A-share (China))
('小','股東')	0.7229	(Small shareholder)
('以','美元')	0.5913	('By', 'Dollar')
('召開','法說會')	0.5411	(Hold earnings call)
('也','因此')	0.5193	(Also because)
('代工','龍頭')	0.5058	(OEM leader)
('中國','a')	0.5039	('China', 'A')
('成長','幅度')	0.4808	(Growth rate)
('也','相當')	0.4734	(Also very)
('股東','臨時')	0.4673	(Interim shareholders meeting)
('台積電','今年')	0.4652	(TSMC this year)
('影響','下')	0.4644	(Under the influence of)
('兆','億元')	0.4634	('Ten billion', 'Hundred million Dollars')
('股東會','通過')	0.4631	(Shareholder's meeting approve)
('晶圓','龍頭')	0.3464	(Leader of wafer)
('鴻海','集團')	0.3459	(Foxconn Group)
('股東','們')	0.3434	(Shareholders)
('今年','成長')	0.3403	(Growth this year)
('大','數據')	0.3350	(Big Data)
('臨時','討論')	0.3323	(Other motions)
('外資','籌碼')	0.3271	(Foreign capital portion)
('人民幣','億元')	0.3259	(Hundred million CNY)
('認購','權證')	0.3244	(Call warrant)

and therefore defeats the purpose of a 2-word combination. Another reason lies in the syntactical structure. The process of word segmentation in English text is fairly simple comparing to Chinese text. One just separates all the whitespaces. Lacking the convenience of whitespace separators, Chinese word segmentation has to go through an arduous process like Jieba as used in this study. But by going through the more complicated word segmentation process, meaningful terms are already captured whether by predefined dictionaries or Markov model and presented as "one" term in the original feature set. For example, the term "shareholder's meeting approve" in English is presented with three words and the feature ('shareholder', 'approve') will be captured by a 2-word combination. On the other hand, the same phrase in Chinese is just "股東會通過". Even though "股東(shareholder)" and "會(meeting)" are two terms that individually has its meaning, during the word segmentation process, Jieba determines that they should be used jointly as one term, therefore, (shareholder's meeting, approve) can be easily captured by 2-gram feature extraction. In conclusion, due to the syntactical structure and difference in the word segmentation process, Chinese text does not necessitate the use of a 2-word combination to achieve the same semantical sophistication of a 2-word combination in English but still benefits slightly from adding another layer to the existing structure. Note that N-word combinations could be tried to find features when different types of articles could be collected and used for solving the problem.

C. Experimental Results for GASTP

In this section, we will discuss the findings of STP used with the Genetic Algorithm to tune its hyper-parameters. The goal is to acquire a set of hyper-parameters that will enhance the performance of the original STP algorithm. We set the geneNum to 200, which means that the evolutionary process will continue for 200 generations. After the evolutionary process, we will take the best performing chromosome as our final tuned hyper-parameters. The obtained hyper-parameters by GASTP are shown in Table VII.

TABLE VII: Obtained hyperparameters by GASTP.

GASTP							
Target	**2-word**	**TechUse**	**MDF**	**Chi%**	**RSI**	**K**	**D**
t + 1	TRUE	AND	3	26	50	68.75	68.75
t + 1	TRUE	OR	3	31	43.75	71.875	68.75
t + 1	FALSE	AND	3	24	78.125	87.5	75
t + 1	FALSE	OR	4	31	46.875	71.875	62.5
t + 5	TRUE	AND	4	26	43.75	34.375	34.375
t + 5	TRUE	OR	4	31	40.625	71.875	75
t + 5	FALSE	AND	3	22	62.5	75	71.875
t + 5	FALSE	OR	3	30	50	75	68.75
t + 10	TRUE	AND	3	22	34.375	37.5	53.125
t + 10	TRUE	OR	3	31	40.625	21.875	25
t + 10	FALSE	AND	5	21	34.375	34.375	59.375
t + 10	FALSE	OR	4	31	28.125	31.25	34.375
Average				27.17	46.09	56.77	58.07

First, we examine the hyperparameters obtained by GASTP. The first two parameters (document frequency threshold and Chi-square percentile) obtained via GASTP are surprisingly different from our benchmark. The first two parameters in the benchmark are set according to previous studies. The idea is that one should only select general textual features with a document frequency threshold of 15 times and choose those that are very high in the Chi-square score (5%). However, the hyperparameters acquired by our study suggest nearly the complete opposite with document frequency threshold less than 5 and most Chi-square score percentile over 25%. Next, we take a closer look at the technical indicators. For $t+1$ and $t+5$, the RSI score tends to be at around 50, which means that the buying and selling strength is precisely the same. This is quite different from common knowledge that a low RSI signifies an oversell hence the stock values are underestimated. In our study, the result shows that when the relative strength does not fall in any particular trend, a new coming news post disclosing new market information will have a stronger effect. In addition, for *KD* value, the best result comes when *K* value is slightly higher than *D* value at around 70. Because *K* moves faster than *D*, this observation signifies the situation in which the "long side" has started to move but not extendedly. The findings, albeit different from the benchmark, where it supposed that the best results happen at a relatively low *KD* value, still appears to be following the fact that *K* value should be higher than *D* value. This is because a news report is not a leading indicator. Market and industry experts usually obtain the information much earlier before it is published as news and therefore has already started to move their investments and therefore has already caused the *K* value to move upwards than the *D* value. Later, when the said information is made public, some investors will follow this trend and consequently generates another momentum for growth. On the other hand, for $t+10$, both *RSI* and *KD* appear to be at a relatively low point. Since we hold the stock for a longer period, the stock movement of $t+10$ will be affected more by the general trend rather than by news factors and so there the best performing technical indicator range goes back to a general knowledge of a low *KD* value.

Another interesting observation is that this result does not

differ between the use of 2-word or not. This means that even though with a 2-word combination, the number of features increases drastically, the percentage of useful features does not change. The features selected from both feature sets after the genetic algorithm have a lot in common. An example of the selected features after the genetic algorithm is given in Table VIII.

TABLE VIII: Top-20 selected features from GASTP.

2-word	Translation	Bag-of-Words	Translation
東芝	Toshiba	減資	Capital reduction
減資	Capital reduction	東芝	Toshiba
牧德	Machvision Inc. Co., Ltd	精測	Precision test
精測	Precision test	法說	Earnings call
嘉聯益	Career Technology (Mfg.) Co., Ltd.	嘉聯益	Career Technology (Mfg.) Co., Ltd.
兆	100 million	點數	points
法說	Earnings call	何麗梅	Ms. He, LiMei (VP and CFO of TSMC)
點數	points	pc	PC
瑞祺電	CASwell Inc.	環保	Environment protection
pc	PC	銀行	Bank
何麗梅	Ms. He, LiMei (VP and CFO of TSMC)	洪水樹	Ms. He, LiMei (VP and CFO of TSMC)
銀行	Bank	exodus	Exodus (HTC smart phone)
環保	Environment protection	越南	Vietnam

In the example, the target date is $t + 1$ and the document frequency threshold and Chi-square percentage obtained from GASTP is $(3, 26)$ and $(3, 24)$ for 2-word combination and Bag-of-words. When we look at the top 20 features by Chi-square score, we can see that most of the features selected are the same and for 2-word combinations, only one combined feature appears in the top 20 list. This is again a sign to suggest that a 2-word combination is not necessary for Chinese text seeing that most features exclusively about 2-word combination do not appear to be among the highest performing features. Next, we look at the results for GASTP that are given in Tables IX and X.

TABLE IX: Results of GASTP (without technical indicators).

GASTP (without technical indicators)					
Algorithm	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
SVM (t + 1, TRUE, AND)	101	0.58	0.3	32303	0.0032
SVM (t + 1, TRUE, OR)	91	0.57	0.26	14694	0.0016
SVM (t + 1, FALSE, AND)	119	0.57	0.32	-771	-0.0001
SVM (t + 1, FALSE, OR)	121	0.57	0.32	12208	0.0010
SVM (t + 5, TRUE, AND)	174	0.53	0.37	-249315	-0.0143
SVM (t + 5, TRUE, OR)	176	0.53	0.37	-264358	-0.0150
SVM (t + 5, FALSE, AND)	163	0.55	0.37	-263851	-0.0162
SVM (t + 5, FALSE, OR)	161	0.53	0.35	-272534	-0.0169
SVM (t + 10, TRUE, AND)	122	0.59	0.24	-563787	-0.0462
SVM (t + 10, TRUE, OR)	126	0.58	0.25	-569541	-0.0452
SVM (t + 10, FALSE, AND)	144	0.57	0.27	-705616	-0.0490
SVM (t + 10, FALSE, OR)	147	0.57	0.28	-654103	-0.0445
Average	137.08	0.56	0.31	-290389.2	-0.02

TABLE X: Results of GASTP (with technical indicators).

GASTP (with technical indicators)					
Algorithm	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
SVM (t + 1, TRUE, AND)	2	0.33	0	-2819	-0.0141
SVM (t + 1, TRUE, OR)	34	0.59	0.27	1787	0.0005
SVM (t + 1, FALSE, AND)	2	0.77	0	-4042	-0.0202
SVM (t + 1, FALSE, OR)	54	0.57	0.37	18267	0.0034
SVM (t + 5, TRUE, AND)	27	0.44	0.43	-17822	-0.0066
SVM (t + 5, TRUE, OR)	58	0.54	0.37	-97488	-0.0168
SVM (t + 5, FALSE, AND)	2	0.25	0.25	-2611	-0.0131
SVM (t + 5, FALSE, OR)	63	0.54	0.33	-140870	-0.0224
SVM (t + 10, TRUE, AND)	5	0.55	0	-55055	-0.1101
SVM (t + 10, TRUE, OR)	16	0.68	0.33	-81286	-0.0508
SVM (t + 10, FALSE, AND)	2	0.25	0	-43863	-0.2193
SVM (t + 10, FALSE, OR)	76	0.57	0.32	-332522	-0.0438
Average	28.42	0.51	0.22	-63193.67	-0.0222

When we compare the results for holding different period lengths. Same as the findings from STP, when we hold for 5 days the result in terms of F1 Score is higher than hold for 1 or 10 days although due to the overall declining trend, this higher F1 Score does not necessarily contribute to a higher return ROI.

We then examine the use of technical indicators. Again, even after the evolutionary process, when technical indicators are used strictly, very few numbers of trades are generated. This is because the market situation between training and testing data is very different. For example, for the target date t + 1 the GASTP determines that the best performing technical indicator set is (50, 68.75, 68.75) for RSI, K value, and D value. In the training data, this market situation occurs 45 times while in the testing data, there are only nine news posts accompanied by this market situation hence creates a strongly biased result. On the other hand, when either one of the technical indicators appears in the set range, the result in terms of the F1 Score increases as we expected. This shows that the GASTP has done its job to tune the hyperparameters.

Next, we examine the performance of the GASTP algorithm on different machine learning classifiers. The results for GASTP on multinomial Naïve Bayes are shown in Tables XI and XII. As previously mentioned, the results for technical indicators used strictly generate very few numbers of trades and are very biased therefore here we only list the result for OR use.

TABLE XI: Results of GASTP with NB (without technical indicators).

GASTP (Without technical indicators)												
Algo.	**Target**	**2-word**	**MDF**	**Chi%**	**RSI**	**K**	**D**	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
NB	t + 1	TRUE	4	20	50	78.125	65.625	70	0.58	0.23	15749	0.0022
NB	t + 1	FALSE	3	24	56.25	75	62.5	121	0.59	0.35	18155	0.0015
NB	t + 5	TRUE	3	21	43.75	71.875	71.875	119	0.55	0.29	-210919	-0.0177
NB	t + 5	FALSE	3	30	50	71.875	68.75	169	0.53	0.36	-268070	-0.0159
NB	t + 10	TRUE	3	25	28.125	78.125	68.75	91	0.62	0.22	-416487	-0.0458
NB	t + 10	FALSE	3	31	40.625	31.25	53.125	93	0.61	0.2	-517908	-0.0557
Average			3.5	22	53.13	76.56	64.06	111	0.58	0.28	-229913	-0.0208

TABLE XII: Results of GASTP with NB (with technical indicators).

GASTP (With technical indicators)												
Algo.	**Target**	**2-word**	**MDF**	**Chi%**	**RSI**	**K**	**D**	**Trad. Signals**	**Acc.**	**F1**	**Return**	**ROI**
NB	t + 1	TRUE	4	20	50	78.125	65.625	30	0.56	0.23	1582	0.0005
NB	t + 1	FALSE	3	24	56.25	75	62.5	45	0.6	0.39	-1120	-0.0002
NB	t + 5	TRUE	3	21	43.75	71.875	71.875	53	0.53	0.32	-88647	-0.0167
NB	t + 5	FALSE	3	30	50	71.875	68.75	65	0.57	0.38	-114093	-0.0176
NB	t + 10	TRUE	3	25	28.125	78.125	68.75	24	0.59	0.19	-78115	-0.0325
NB	t + 10	FALSE	3	31	40.625	31.25	53.125	31	0.63	0.23	-187108	-0.0604
Average			3.5	22	53.13	76.56	64.06	41	0.58	0.29	-77917	-0.0189

As shown in the results, the five hyperparameters acquire using NB are very similar to those acquired by using SVM. The results in terms of F1 score, accuracy, and ROI are virtually the same. The most noticeable difference is that NB generally requires fewer features to achieve the same return. This helps tremendously in processing speed. As to with or without using technical indicators, the observations are stated as follows: (1) When using technical indicators, they could effectively reduce the number of trades when the market trend is a downtrend. In other words, the risk is reduced because a few trades can reach a similar return; (2) Using RSI and KD for finding trading signals still have spaces to be improved which means that other indicators can be employed to get a better result. For instance, the moving average (MA) or Bollinger Bands (BBands) can be examined. Besides, the sentiment indices that are constructed by different data can also be utilized to obtain trading signals. For example, Liang

et al. used social media, newspaper, and Internet media news to generate three sentiment indices for trading [25], and Reis *et al.* presented the EURsent that can provide investors to monitor the sentiment of the stock market in Europe [36]. The results of the comparison of the two classifiers are shown in Table XIII.

TABLE XIII: Compared results of the two classifiers.

GASTP										
	Without technical indicators					With technical indicators				
Alg.	Trad. Signals	Acc.	F1	Return	ROI	Trad. Signals	Acc.	F1	Return	ROI
SVM (t + 1, TRUE)	91	0.57	0.26	14694	0.0016	34	0.59	0.27	1787	0.0005
SVM (t + 1, FALSE)	121	0.57	0.32	12208	0.0010	54	0.57	0.37	18267	0.0034
SVM (t + 5, TRUE)	176	0.53	0.37	-264358	-0.0015	58	0.54	0.37	-97488	-0.0168
SVM (t + 5, FALSE)	161	0.53	0.35	-272534	-0.0169	63	0.54	0.33	-140870	-0.0224
SVM (t + 10, TRUE)	126	0.58	0.25	-569541	-0.0452	16	0.68	0.33	-81286	-0.0508
SVM (t + 10, FALSE)	147	0.57	0.28	-654103	-0.0445	76	0.57	0.32	-332522	-0.0438
Average	137	0.56	0.31	-288939	-0.02	50.17	0.58	0.33	-105352	-0.0210
	Without technical indicators					With technical indicators				
Algo.	Trad. Signals	Acc.	F1	Return	ROI	Trad. Signals	Acc.	F1	Return	ROI
NB (t + 1, TRUE)	70	0.58	0.23	15749	0.0022	30	0.56	0.23	1582	0.0005
NB (t + 1, FALSE)	121	0.59	0.35	18155	0.0015	45	0.6	0.39	-1120	-0.0002
NB (t + 5, TRUE)	119	0.55	0.29	-210919	-0.0177	53	0.53	0.32	-88647	-0.0167
NB (t + 5, FALSE)	169	0.53	0.36	-268070	-0.0159	65	0.57	0.38	-114093	-0.0176
NB (t + 10, TRUE)	91	0.62	0.22	-416487	-0.0458	24	0.59	0.19	-78115	-0.0325
NB (t + 10, FALSE)	93	0.61	0.2	-517908	-0.0557	31	0.63	0.23	-187108	-0.0604
Average	111	0.58	0.28	-229913	-0.0208	41	0.58	0.29	-77919	-0.0189

Finally, we compare the results of STP, GASTP with multinomial Naïve Bayes, and a simple buy-and-hold strategy (BHS). The compared results are shown in Table XIV. The formula of ROI is calculated as Eq. 8.

$$ROI = \frac{Earnings(Loss)}{\#\ of\ trades \times 100,000}. \tag{8}$$

TABLE XIV: Compared result of STP and GASTP with Naive Bayes and BHS.

STP										
	Without technical indicators					With technical indicators				
Alg.	Trad. Signals	Acc.	F1	Return	ROI	Trad. Signals	Acc.	F1	Return	ROI
SVM (t + 1, TRUE)	44	0.57	0.13	-5448	-0.0012	11	0.56	0.08	-392	-0.0004
SVM (t + 1, FALSE)	22	0.59	0.08	-3467	-0.0016	9	0.57	0.08	1490	0.0017
SVM (t + 5, TRUE)	116	0.55	0.28	-183485	-0.0158	54	0.56	0.37	-81715	-0.0151
SVM (t + 5, FALSE)	96	0.55	0.24	-209239	-0.0218	48	0.56	0.35	-101328	-0.0211
SVM (t + 10, TRUE)	92	0.6	0.18	-600594	-0.0653	47	0.51	0.19	-323905	-0.0689
SVM (t + 10, FALSE)	79	0.62	0.19	-506347	-0.0641	46	0.53	0.23	-315895	-0.0687
Average	75	0.58	0.18	-251430	-0.0336	36	0.55	0.22	-136958	-0.0382
GASTP										
	Without technical indicators					With technical indicators				
Algo.	Trad. Signals	Acc.	F1	Return	ROI	Trad. Signals	Acc.	F1	Return	ROI
NB (t + 1, TRUE)	70	0.58	0.23	15749	0.0022	30	0.56	0.23	1582	0.0005
NB (t + 1, FALSE)	121	0.59	0.35	18155	0.0015	45	0.6	0.39	-1120	-0.0002
NB (t + 5, TRUE)	119	0.55	0.29	-210919	-0.0177	53	0.53	0.32	-88647	-0.0167
NB (t + 5, FALSE)	169	0.53	0.36	-268070	-0.0159	65	0.57	0.38	-114093	-0.0176
NB (t + 10, TRUE)	91	0.62	0.22	-416487	-0.0458	24	0.59	0.19	-78115	-0.0325
NB (t + 10, FALSE)	93	0.61	0.2	-517908	-0.0557	31	0.63	0.23	-187108	-0.0604
Average	111	0.58	0.28	-229913	-0.0208	41	0.58	0.29	-77919	-0.0189
								Return of BHS		-0.33.77

For the buy-and-hold strategy, we buy $100,000$ TWD worth of stocks for each of our target companies at the start of our testing period and sell all of them at the end of said period. As the results show, when comparing the four blocks, the bottom-right, which is GASTP with technical indicators, outperformed the rest in terms of earnings, accuracy, and ROI. That is to say, that GASTP successfully enhanced the performance of STP, and both STP and GASTP significantly outperformed the buy-and-hold strategy.

V. Conclusion and Future Work

In this paper, two approaches, namely the STP and GASTP, have been proposed for finding trading signals. Through the experiments, the observations as stated as follows: (1) We found that Chinese financial news articles possess reasonable predictive power for stock trend prediction. They are most useful to be used on short-term trading while for long-term trading the effect diminishes over time; (2) We also found that news factors have the most effective either when the stock price does not fall in any particular trend hence any new information will stir the market for more movement or when a short-term trend has started to rise above the long-term trend and a confirmation of the beneficial news leak will further generate another momentum; (3) We also found that 2-word combination feature extraction technique although helps enhance the performance on English text, does not perform particularly well on the Chinese context. For future work, we suggest that: (1) The proposed algorithms can be tested on different kinds of corpus. For example, one can examine the performance of general news articles on market indices or sector-specific news on other commodities like energy, gold, etc.; (2) In addition, to make the feature selection process effectively, the important key sentences can be extracted instead of entire new to generate keywords, and the sentiment analytic can be utilized to generate more meaningful label of every article; (3) Furthermore, different types of classifiers, e.g., deep learning, decision tree, and indicators, e.g., MA, BBands, sentiment indices can be examined on the proposed algorithms.

ACKNOWLEDGMENT

This research was supported by the Ministry of Science and Technology of the Republic of China under grant MOST 109-2622-E-027-032.

References

[1] J. Boudoukh, R. Feldman, S. Kogan and M. Richardson, Which news moves stock prices? A textual analysis, *National Bureau of Economic Research Working Paper Series*, No. 18725, pp. 1–46, 2013.

[2] R. Bhargavi, G. Srinivas and R. Anith, Relative strength index for developing effective trading strategies in constructing optimal portfolio, *International Journal of Applied Engineering Research*, vol. 12, No. 19, pp. 8926–8936, 2017.

[3] M. Y. Chen, C. H. Liao and R. P. Hsieh, Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach, *Computers in Human Behavior*, vol. 101, pp. 402–408, 2019.

[4] M. Clatworthy and M. J. Jones, Financial reporting of good news and bad news: evidence from accounting narratives, *Accounting and Business Research*, Vol. 33, No. 3, pp. 171–185, 2003.

[5] W. Chen, M. Jiang, W. G. Zhang and Z. Chen, A novel graph convolutional feature based convolutional neural network for stock trend prediction, *Information Sciences*, vol. 556, pp. 67–94, 2021.

[6] R. Caruana and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, *The International Conference on Machine Learning*, pp. 1–8, 2006.

[7] A. Derakhshan and H. Beigy, Sentiment analysis on stock social media for stock price movement prediction, *Engineering Applications of Artificial Intelligence*, vol. 85, pp. 569–578, 2019.

[8] V. Dineshreddy and G. R. Gangadharan, Towards an "Internet of Things" framework for financial services sector, *International Conference on Recent Advances in Information Technology*, pp. 1–5, 2016.

[9] E. Fama, Random walks in stock market prices, *Financial Analysts Journal*, Vol. 21, pp. 55–59, 1965.

[10] E. Fama, Efficient capital markets: A review of theory and empirical work, *The Journal of Finance*, Vol. 25, pp. 383–417, 1970.

[11] S. Feuerriegel and H. Prendinger, News-based trading strategies, *Decision Support Systems*, Vol. 90, pp. 65–74, 2016.

[12] T. Geva and J. Zahavi, Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news, *Decision Support Systems*, Vol. 57, pp. 212–223, 2014.

[13] S. Gupta, R. Garg, N. Gupta, W. S. Alnumay, U. Ghosh, and P. K. Sharma, Energy-efficient dynamic homomorphic security scheme for fog computing in IoT networks, *Journal of Information Security and Applications*, Vol. 58, pp. 102768, 2021.

[14] T. Y. Lin, C. W. S. Chen and F. Y. Syu, Multi-asset pair-trading strategy: A statistical learning approach, *The North American Journal of Economics and Finance*, Vol. 55, 101295, 2021.

[15] M. Maiti and U. Ghosh, Next generation Internet of things in fintech ecosystem, *IEEE Internet of Things Journal*, 2021.

[16] A. Handler, M. J. Denny, H. Wallach and B. O'Connor, Bag of what? Simple noun phrase extraction for text analysis, *Workshop on Natural Language Processing and Computational Social Science at the Conference on Empirical Methods in Natural Language Processing*, pp. 114–124, 2016.
[17] P. Y. Hao, C. F. Kung, C. Y. Chang and J. B. Ou, Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane, *Applied Soft Computing*, vol. 98, 106806, 2021.
[18] M. Hagenau, M. Liebmann and D. Neumann, Automated news reading: Stock price prediction based on financial news using context-capturing features, *Decision Support Systems*, Vol. 55, No. 3, pp. 685–697, 2013.
[19] Y. Kim, S. R. Jeong and I. Ghani, Text opinion mining to analyze news for stock market prediction, *International Journal of Advances in Soft Computing and its Applications*, Vol. 6, No. 1, pp. 1–13, 2014.
[20] C. Lee, A study of deep learning with different finance news providers for forecasting stock price trends, *Executive Master's Program of Business Administration in Information Management of Tamkang University*, 2016.
[21] K. Lim, R. D. Brooks and J. H. Kim, Financial crisis and stock market efficiency: Empirical evidence from Asian countries, *International Review of Financial Analysis*, Vol. 17, No. 3, pp. 571–591, 2008.
[22] J. Long, Z. Chen, W. He, T. Wu and J. Ren, An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market, *Applied Soft Computing*, vol. 91, 106205, 2020.
[23] I. Lee and K. Lee, The Internet of Things (IoT): Applications, investments, and challenges for enterprises, *Business Horizons*, vol. 58, No 4, pp. 431–440, 2015.
[24] N. I. Lomakin, I. A. Popov, A. B. Shohneh, M. C. Maramygin and A. B. Gorbunova, AI-System of Stock Exchange Trading Robot for Financial Risk Hedging, *Advances in Economics, Business and Management Research*, vol. 128, pp. 3273–3282, 2020.
[25] C. Liang, L. Tang, Y. Li and Y. Wei, Which sentiment index is more informative to forecast stock market volatility? Evidence from China, *International Review of Financial Analysis*, vol. 71, 101552, 2020.
[26] Q. Li, T. Wang, Q. Gong, Y. Chen, Z. Lin and S. Song, Media-aware quantitative trading based on public Web information, *Decision Support Systems*, Vol. 61, pp. 93–105, 2014.
[27] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong and Y. Chen, The effect of news and public mood on stock movements, *Information Sciences*, Vol. 278, pp. 826–840, 2014.
[28] E. Marsh and D. Perzanowski, MUC-7 Evaluation of IE technology: Overview of results, *Seventh Message Understanding Conference: Proceedings of a Conference Held in Fairfax, Virginia*, pp. 1–71, 1998.
[29] A. K. Nassirtoussi, S. Aghabozorgi, T. Wah and D. C. L. Ngo, Text mining for market prediction: A systematic review, *Expert Systems with Applications*, Vol. 41, pp. 7653–7670, 2014.
[30] M. Naved and P. Srivastava, Profitability of Oscillators used in technical analysis for financial market, *Advances in Economics and Business Management*, Vol. 2, No. 9, pp. 925-931, 2015.
[31] A. K. Nassirtoussi, S. Aghabozorgi, T. Wah and D. C. L. Ngo, Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment, *Expert Systems with Applications*, Vol. 42, No. 1, pp. 306–324, 2015.
[32] W. Nuij, V. Milea, F. Hogenboom, F. Frasincar and U. Kaymak, An automated framework for incorporating news into stock trading strategies, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 4, pp. 823–835, 2014.
[33] T. H. Nguyen, K. Shirai and J. Velcin, Sentiment analysis on social media for stock movement prediction, *Expert Systems with Applications*, Vol. 42, No. 24, pp. 9603–9611, 2015.
[34] V. Pestov, Is the -NN classifier in high dimensions affected by the curse of dimensionality?, *Computers and Mathematics with Applications*, Vol. 65, pp. 1427–1437, 2013.
[35] T. Poibeau and L. Kosseim, Proper name extraction from non-journalistic texts, *Language and Computers*, Vol. 37, No. 1, pp. 144–157, 2001.
[36] P. M. N. Reis and C. Pinho, A new European investor sentiment index (EURsent) and its return and volatility predictability, *Journal of Behavioral and Experimental Finance*, vol. 27, 100373, 2020.
[37] R. P. Schumaker and H. Chen, A discrete stock price prediction engine based on financial news, *Computer*, Vol. 43, No. 1, pp. 51–56, 2010.
[38] S. Sun, S. Wang and Y. Wei, A new ensemble deep learning approach for exchange rates forecasting and trading, *Advanced Engineering Informatics*, vol. 46, 101160, 2020.
[39] R. P. Schumaker, Y. Zhang, C. Huang and H. Chen, Evaluating sentiment in financial news articles, *Decision Support Systems*, Vol. 53, No. 3, pp. 458–464, 2012.
[40] T. T. Vu, S. Chang, Q. T. Ha and N. Collier, An experiment in integrating sentiment features for tech stock prediction in twitter, *The Workshop on Information Extraction and Entity Analytics on Social Media Data*, pp. 23–38, 2012.
[41] G. Wu, T. Hou and J. Lin, Can economic news predict Taiwan stock market returns?, *Asia Pacific Management Review*, Vol. 24, No. 1, pp. 54–59, 2019.
[42] J. M. T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C. W. Lin, A graph-based CNN-LSTM stock price prediction algorithm with leading indicators, *Multimedia Systems*, 2021.
[43] J. M. T. Wu, Z. Li, G. Srivastava, M. H. Tsai and J. C. W. Lin, A graph-based convolutional neural network stock price prediction with leading indicators, *Software: Practice and Experience*, Vol. 51(3), pp. 628–644, 2021.
[44] Y. Yu, W. Duan and Q. Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, *Decision Support Systems*, Vol. 55, No. 4, pp. 919–926, 2013.
[45] Y. Yang, S. Mo, A. Liu and A. A. Kirilenko, Genetic programming optimization for a sentiment feedback strength based trading strategy, *Neurocomputing*, vol. 264, pp. 29–41, 2017.

Chun-Hao Chen is an associate professor at Department of Information and Finance Management at National Taipei University of Technology, Taipei, Taiwan. Dr. Chen received his Ph.D. degree with major in computer science and information engineering from National Cheng Kung University, Taiwan, in 2008. He has a wide variety of research interests covering data mining, time series, machine learning, evolutionary algorithms, and fuzzy theory. Research topics cover portfolio selection, trading strategy, business data analysis, time series pattern discovery, etc. He serves as the associate editor of the International Journal of Data Science and Pattern Recognition, and IEEE Access. He is also a member of IEEE.

Ping Shih Ping Shih received his M.S. degree in Computer Science from the Department of Computer Science and Information Engineering at Tamkang University, Taiwan, in 2019. His thesis was published in the 2019 IEEE Congress on Evolutionary Computation. His research interests include Text mining and NLP with a focus on financial sector. He currently works as a Data Engineer in the Banking industry.

Gautam Srivastava was awarded his B.Sc. degree from Briar Cliff University in the U.S.A. in the year 2004, followed by his M.Sc. and Ph.D. degrees from the University of Victoria in Victoria, British Columbia, Canada in the years 2006 and 2012, respectively. He then taught for 3 years at the University of Victoria in the Department of Computer Science, where he was regarded as one of the top undergraduate professors in the Computer Science Course Instruction at the University. From there in the year 2014, he joined a tenure-track position at Brandon University in Brandon, Manitoba, Canada, where he currently is active in various professional and scholarly activities. He was promoted to the rank of Associate Professor in January 2018. Dr. G, as he is popularly known, is active in research in the field of Cryptography, Data Mining, Security and Privacy, and Blockchain Technology. He has published a total of 200 papers in high-impact conferences and in high status journals (SCI, SCIE). He is an Editor of several SCI/SCIE journals. He is an IEEE Senior Member.

Shih-Ting Hung received her B.S. degree in electrical engineering from Fu Jen Catholic University, Taiwan, in 2017. And received her M.S. degree in Data Science from Taipei Medical University, Taiwan, in 2019. Her research interests include genetic algorithms and data mining. She is currently pursuing her Ph.D. degree in the department of information and finance management from National Taipei University of Technology, Taipei, Taiwan.

Jerry Chun-Wei Lin received his Ph.D. from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan in 2010. He is currently a full Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has published more than 400 research articles in refereed journals (IEEE TKDE, IEEE TCYB, IEEE TII, IEEE TITS, IEEE TNSE, IEEE TETCI, IEEE SysJ, IEEE SensJ, IEEE IOTJ, ACM TKDD, ACM TDS, ACM TMIS, ACM TOIT, ACM TIST) and international conferences (ICDE, ICDM, PKDD, PAKDD). His research interests include data mining, soft computing, artificial intelligence and machine learning, and privacy preserving and security technologies. He is the Editor-in-Chief of the International Journal of Data Science and Pattern Recognition, the Guest Editor/Associate Editor of IEEE TFS, IEEE TII, ACM TMIS, IEEE JBHI, ACM JDIQ, ACM TOIT, ACM TALLIP, IEEE Access, JIT, PlosOne, IDA, AIHC, JCSC, and IJIMAI. He is the Fellow of IET (FIET), senior member for both IEEE and ACM.