Abstract— Online financial textual information contains a large amount of investor sentiment, i.e. subjective assessment and discussion with respect to financial instruments. An effective solution to automate the sentiment analysis of such large amounts of online financial texts would be extremely beneficial. This paper presents a natural language processing (NLP) based pre-processing approach both for noise removal from raw online financial texts and for organizing such texts into an enhanced format that is more usable for feature extraction. The proposed approach integrates six NLP processing steps, including a developed syntactic and semantic combined negation handling algorithm, to reduce noise in the online informal text. Three-class sentiment classification is also introduced in each system implementation. Experimental results show that the proposed pre-processing approach outperforms other pre-processing methods. The combined negation handling algorithm is also evaluated against three standard negation handling approaches.

I. INTRODUCTION

Financial investors trade on the basis of available information, particularly on the probability that items of information about corporate financial analysis will impact on the market. Online information related to finance plays an increasingly important role in financial markets and personal finance. Financial bulletin board forums, Twitter and blogs are popular channels for investors to exchange information and express opinions about different financial instruments and trading strategies. Sentiment analysis is the process of extracting the emotive content from such texts. Sentiment classification is an important component of sentiment analysis and aims to classify an opinion from text as expressing a positive or negative sentiment [1]. Machine learning based sentiment classification is performed in a series of sequential steps: pre-processing, feature extraction and selection, and actual classification. Pre-processing, an important step in text classification [2], is the process of converting the original online textual documents into a feature extraction-ready structure, where the key features and terms from online texts, that serve to differentiate between different sentiment classes, are identified. It has been proven that the time spent on pre-processing occupies most of the time in the entire classification process [2]. The importance of pre-processing is emphasized by two factors in recent years [3, 4]: (1) the volume of online text on stock related chat rooms and bulletin boards is growing rapidly as an increasing number of participants offer financial analyses and express different opinions, which increases the dimension of feature space; (2) online informal text requires more sophisticated methods to clean noise in raw text. Most of the existing research on sentiment analysis focuses on developing classifiers and feature extraction techniques for preprocessed textual data, but in contrast the development of more effective pre-processing approaches is still needed. Existing stop words removal approaches simply remove a list of common words which may carry sentiment information, and ignore other non-informative information and semantic relation in text. A number of negation handling methods have been developed and the effect of negation handling as one pre-processing step in sentiment analysis has been evaluated in [3, 5, 6, 7, 8]. However, these methods used syntactic-based and semantic-based approaches separately. Existing negation approaches and models use common negation words but do not include other lexical units, such as diminishers. Furthermore, an effective approach for modelling the sentiment impact of negation expressions in the sentence is needed.

The aim of this research is to develop an effective pre-processing approach to clean non-informative words and symbols in text, so as to enhance feature extraction from online unstructured text. Such an approach will improve the relevance between word and document and between word and sentiment classes for sentiment analysis. To achieve this aim, an effective NLP based approach is proposed for preprocessing online unstructured text. This method integrates five existing NLP steps with a new negation handling method to handle hyperlinks, additional punctuations, negation expression and lengthening words in online informal text. The proposed negation handling method combines semantic and syntactic techniques for text pre-processing to identify more negation terms in sentences and its sentiment impact on the sentences more accurately. The whole proposed pre-processing approach consists of six NLP steps to remove more non-informative information and handle the means of expression typically in online informal text.

The remainder of this paper is structured as follows: existing sentiment analysis related work is presented in Section II. Section III presents the proposed pre-processing approach in detail and a description of the financial text corpora used for evaluation is presented in Section IV. Section V presents the experimental evaluation of the proposed approach and comparisons with the selected three
existing sentiment analysis systems. Finally, conclusions and future research directions are presented in Section VI.

II. LITERATURE REVIEW

Sentiment analysis of online unstructured text has attracted significant attention from researchers since the early 2000s [9], and research on extracting and classifying investors’ sentiments from online financial-domain text has been growing rapidly with the development of sentiment analysis. However, most existing research focuses on development of algorithms and computational techniques for feature extraction and textual sentiment polarity analysis. This paper focuses on machine learning-based sentiment analysis and related research work on machine learning sentiment analysis, especially in the financial domain, will be discussed in this section. In addition, NLP based text pre-processing, as one step in the process of sentiment analysis has been developed in many research works; this is also discussed in this section.

A. Machine learning based sentiment analysis approaches

Machine learning techniques are the main tools used in text classification. Popular approaches include K-Nearest Neighbour (kNN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Tree Learning, Naive Bayes (NB) and Maximum Entropy (ME) [9]. Machine learning approaches were first introduced into sentiment analysis by Pang and Lee [8]. NB, ME and SVM were used for sentiment analysis of movie reviews, and SVM produced an accuracy of 82.9% with a unigram feature set representation, NB classifier with an accuracy of 81.0% [8]. In 2004, Pang and Lee [6] developed their sentiment detection method by using sentence subjectivity classification in the same domain (i.e. movie review). As SVM performed comparably to or better than other machine learning methods, it was then widely used in sentiment analysis [1, 6, 10, 11, 12, 13, 14, 15]. However, NB, has also been used in [6, 8, 16, 17, 18]. Thelwall [18] used ME for sentiment classification and an NB classifier for robustness check, to extract sentiment strength from informal English text. Machine learning approaches have demonstrated the usefulness of sentiment extracted from web information. However, they use rather shallow statistics-based methods that typically classify at the document-level and do not analyze in detail the object of each expressed sentiment.

B. Financial domain specific sentiment analysis in short informal text

Sentiment analysis of online financial text has been performed in several research studies [5, 6, 11, 12, 15, 16, 18, 19, 20, 21, 22, 23]. Zhang and Swanson [23] pointed out that public online stock discussion boards provide a pool of analysed sentiment information with financial value. Machine learning techniques are widely applied to sentiment analysis of stock discussion boards; for example, Chua [6] employed a variation of NB classifiers to classify internet stock message boards associated with term frequency and information gain feature selection methods and produced an accuracy of 78.72%. O’Hare [21] developed a topic-dependency technique to extract topic-specific subdocuments from financial blogs in order to deal with the topic-shift problem in a document. Zhang and Swanson [23] showed that posts on financial stock message boards exist in the form of conversational discourse and demonstrated that the ME classifier performed well in classifying sentiments expressed in conversational discourse. Thelwall [18] created a new algorithm, SentiStrength, to extract sentiment strength from informal English stock messages. Klein [24] used a semantic technique, which integrated ontology-guided and rule-based web information extraction, and was employed for financial weblog sentiment extraction. Both [18] and [24] developed semantic feature extraction methods for sentiment classification. However, these approaches fall short in relating sentiment analysis to specific financial domains. As financial domain-specific knowledge can change the sentiment polarity in financial text, but financial terms and common financial phrases were split into single words and treated the same as other expressions, which damaged the sentimental implication in the financial-specific expressions.

C. Natural language processing and sentiment analysis

NLP can be applied both on the proposed pre-processing approach, and also feature extraction phases of the sentiment classification process. Linguistic features such as part-of-speech (POS) tagged features can be extracted from text and used as part of feature vectors [1, 3, 5, 17, 24]. Because of the informal and unstructured nature of the language used in Tweets, Cohen [17] introduced a series of pre-processing steps using the Natural Language Toolkit (NLTK) in Python to sanitize and normalize the users’ tweets; this has been proven to improve the quality of the features extracted from the messages and subsequently improve the performance of the classifier. The open source software, General Architecture for Text Engineering (GATE) [24], carries out sentiment and semantic annotation by means of gazetteers lists. GATE is a framework for developing and deploying software components that process natural language. NLP was performed based on GATE’s A Nearly-New Information Extraction System (ANNIE) to identify the sentences that potentially contain sentiment [24]. NLP techniques can also be used for both feature extraction and selection, such as POS tagged features [5].

D. Pre-processing techniques for sentiment analysis

Two very important pre-processing steps, stopwords removal and stemming in text categorization, were discussed and their effect on text classification analyzed by Sridividy and Anitha [25]. However, both Yu [26] and Saif [27] proved that stopwords removal, as a standard pre-processing step, reduced the sentiment classification accuracy as stopwords can be used as discriminative features for specific sentiment classification. Other sentiment analysis methods [5, 8, 10, 11, 14] used negation tagging, POS tagging and word tokenization. However, only Klein [24] used GATE’s ANNIE system to preprocess text in a systematic way. Jia [28] and Asmi [29] analyzed the effect of negation handling on sentiment analysis. Wiegand [30] systematically analysed natural language negation from linguistic perspective.

Text pre-processing puts emphasis on finding the scope of negation (where is negation) and its impact on the sentence (i.e. how does negation shift the sentiment polarity). The Bag-of–words (BOW) technique is a basic way to represent documents in a dataset [8, 11]. Pang [8] used BOW to extract features from documents and added a prefix NOT to each detected negation word as a new feature. However, this way to treat negation words in text cannot properly detect the scope of the negation impact. The first
III. PROPOSED PRE-PROCESSING APPROACH

Data pre-processing in sentiment analysis is the process of reducing the noise and preparing the text for sentiment classification. Online informal texts contain lots of noise and non-informative parts such as HTML tags, scripts and advertising. Furthermore, many words in the text do not have an impact on the actual sentiment. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult, as each word in the text is treated as one feature. Therefore, the goal of the text pre-processing is to reduce the noise in the text and help improve the performance of the classifier and speed up the classification process.

The proposed approach integrates six NLP processing steps. These are implemented in sequence as shown in Figure 1 and described below.

A. Removal of hyperlinks and numbers

Online informal text sometimes contains URL links which lead to another web paper to provide additional information to support the author’s opinion. However, the links themselves as content in text do not give any meaningful information; therefore we have decided to first remove URL links, such as ‘http://sentimentsymposium.com/’ in text. Some documents contain advertising hyperlinks which are not written by the document author, and the advertising hyperlinks are removed. A number is also a way to support the author’s opinion, but as a feature it does not contain semantic meaning, so numbers are removed after hyperlinks removal.

B. Abbreviation extending

Abbreviation is a common expression in informal text and extending abbreviations can make it easier for further text analysis, such as handling negation and POS tagging. Accordingly abbreviation extension is chosen as the second step in the pre-processing process. The extending abbreviation method is to replace the abbreviation by its full form, such as ‘aren’t’ replaced with ‘are not’, ‘we’re’ replaced with ‘we are’.

C. Additional punctuation and lengthening words extraction & replacement before tokenization

Online unstructured texts contain additional punctuations and lengthening words which sometimes represent strong sentiment or boost sentiment strength, so they cannot simply be removed. Thus processing these words is necessary for sentiment analysis and is selected as the third step in the proposed approach. Instead of removing the additional punctuations and lengthening words, the method to process them in this paper is as follows:

- replace additional punctuations by a suitable corresponding adverb word with a full stop to complete the sentence, such as ‘!!!!!!!’ → ‘extremely.’
- replace lengthened words by their correct spelling format with tagging _STR to emphasis its strong sentiment such as ‘Ahmmmmmm’ → ‘ah_STR’.

D. Negation identification and handling

Negation is a type of expression that can shift sentiment polarity in text, which needs to be taken into consideration in sentiment analysis. The method developed for handling negation is used to decide upon two aspects of scope: the scope of negation terms and the negation impact scope in a sentence which contains negation terms. To identify the scope of a negation term, the negative word list provided in

![Figure 1 Proposed pre-processing approach process](image-url)

- Removal of hyperlinks and numbers
- Abbreviation extending
- Additional punctuation and lengthening words extraction & replacement before tokenization
- Syntactic identification and negation words handling for negation sentences
- POS tagging and removal of pronoun, preposition and conjunction, and punctuations
- Lemmatizing remaining tokenized POS-tagged words

---

Computational negation model, contextual valence shifting (CVS), was introduced in [31] and evaluated [18, 29, 31, 30]. The model assigned scores to polar expressions, i.e. positive scores to positive expressions and negative scores to negative expressions, respectively. If a polar expression is negated, its polarity score is simply inverted. Other researchers have tried to define the scope by defining lists of verbs, adjectives and adverbs and defining their relationships for sentiment analysis [30]. Lists of positive and negative terms and a set of lists for modifiers was proposed in [18] to define the scope of these modifiers as n-terms before and after positive or negative terms (where n remained a constant). This technique is better for negation identification in comparison to the BOW technique. However, the lists used for this technique may grow with time and can never be complete, as in any language there might be an infinite number of words and ways they can be used. Semantic composition refers to the relationship between concepts or meanings e.g. antonym and synonym. Semantic composition was used for negation identification in [2, 28, 29, 30]. In [28], a method to compute the polarity of headlines and complex noun phrases using compositional semantics is presented. The paper argues that the principles of the linguistic modeling paradigm can be successfully applied to determine the sub-sentential polarity of the sentiment expressed, demonstrating it through its application to contexts involving sentiment propagation, polarity reversal or polarity conflict resolution. A similar approach is presented in [29]. The main difference to [28] lies in the representation format on which the compositional model is applied. The advantage of this method is that it more accurately represents the meaning of the text it describes.

---

**Figure 1 Proposed pre-processing approach process**

![Figure 1 Proposed pre-processing approach process](image-url)
Negation shifting identification:

Assign a suffix to want as want\_NEG

Figu re 2 A example of negation recognition by using Stanford parser

[18] is used as a candidate negation word list with 93 words and terms. WordNet [32], which is an online lexical database to group English words into sets of cognitive synonyms, is used to search synonyms of negation words in the list. These include adverbs, suffix, prefix, verbs and nouns, extending the negation list to 394 words and terms, to produce a more complete negation word list, and search negation words in the new list from each sentence in the text. For a sentence with a negation term in new negation word list, the Stanford parser, one of the most popular POS parsers [30], is used to assign POS tags to each word, identify how different words interact within a sentence and identify the syntactic relationship within a sentence. When the word(s) which has/have relations with the negation term in a sentence is (are) decided by the dependency tree produced by the Stanford parser and shows in the same bracket, a \_NEG suffix is appended to the word(s). An example of negation recognition by using Stanford parser is given in Figure 2.

E. POS tagging and removal of pronouns, prepositions and conjunctions, and punctuations

There are many cases where a sentiment contrast exists between words that have the same string representation but different POS, so it may be worthwhile to apply a POS tagger on the sentiment text and then use the resulting word-tag pairs as features or components of features. Most sentiment analysis research [3, 6, 8, 11, 17, 22] indicate that only verb, noun, adjective, adverb and interjection contain sentiment and pronoun, preposition and conjunction with little sentiment information can be removed to reduce the dimension of features. So POS tagging and pronoun, preposition and conjunction removal is put in fifth step in the approach. The Penn Tree Bank parser [32] is employed to assign POS tags to each word in the sentence after sentence tokenization. A list of tagged tokens is returned, where a tagged token is a tuple of (word, tag), and then remove words with pronoun, preposition or conjunction tag. Remaining punctuations are removed in this step as well.

Table 1. Three financial sentiment datasets statistics

<table>
<thead>
<tr>
<th></th>
<th>No. of Buy-labelled post</th>
<th>No. of Hold-labelled post</th>
<th>No. of Sell-labelled post</th>
<th>No. of total labelled post</th>
<th>Average No. of words per post</th>
</tr>
</thead>
<tbody>
<tr>
<td>GKP</td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>1536</td>
<td>53.8</td>
</tr>
<tr>
<td>IFS</td>
<td>357</td>
<td></td>
<td>643</td>
<td>306</td>
<td>85.4</td>
</tr>
<tr>
<td>ST</td>
<td>191</td>
<td>377</td>
<td>581</td>
<td>1149</td>
<td>140</td>
</tr>
</tbody>
</table>

F. Lemmatisation

Lemmatisation is a technique for removing affixes from a word, ending up with the stem. The purpose of lemmatisation is to reduce the feature dimension, as sometimes words with different forms have the same meaning. Lemming is the last pre-processing step in the proposed approach. Lemmatisation is very similar to stemming, but is more akin to synonym replacement. A lemma is a valid root word, and the aim of word lemmatisation is to use one word form, such as ‘message’ and ‘messages’ or ‘believes’ and ‘belief’.

IV. ONLINE FINANCIAL TEXT CORPORA

Three sentiment annotated corpora of financial texts are used in this work to evaluate the proposed pre-processing approach for sentiment analysis. A summary description of the properties of the three datasets is shown in Table 1. Each dataset is introduced in more detail below.

A. GKP stock forum dataset (GKP)

This collection of financial posts was extracted from the Interactive Investor (iii.co.uk) stock discussion board about Gulf Keystone Petroleum stock by the authors. Interactive Investor is one of the largest UK-based communities of traders and investors, and provides financial products and services for investors. Its community discussion boards enable users to express their opinions about a specific stock. Gulf Keystone Petroleum Ltd is an independent oil and gas exploration and production British company, and it was incorporated in 2001 in Bermuda and listed on the Market AIM (Alternative Investment Market) of the London Stock Exchange in 2004 (stock quote GKP). GKP is the one of the most active stocks in the discussion boards of Interactive Investor.

Author-labeled posts discussed about GKP from GKP RSS feeds [http://www.iii.co.uk/rss/cotn:GKP.L.xml] were saved into an XML document for a half year period, from 1st July 2012 to 31st December 2012. The same numbers of posts from three classes—BUY, HOLD and SELL—were selected.
B. Irish financial sentiment dataset (IFS)

Both raw and preprocessed datasets were obtained from the University College Dublin Machine Learning Group [33]. The financial news sentiment analysis collection was retrieved from three online news sources (RTÉ, The Irish Times, the Irish Independent) during a three month period (July to October 2009). A subset of documents was annotated on a daily basis by a group of 33 volunteer users who labeled the articles as positive, negative, or irrelevant. The first month constituted a “warm-up” period, which provided an initial dataset containing 3858 articles, with 2693 user annotations covering 354 individual articles. This second “main” dataset comprises 12469 documents, with 6910 user annotations resulting in 1306 labeled articles. The “main” dataset was used for experiments in this paper.

C. Sanders-Twitter financial sentiment Corpus (ST Apple)

This corpus was obtained from the Sanders Analytics Company (http://www.sanalytics.com/lab/twitter-sentiment/), which is designed for training and testing Twitter sentiment analysis algorithms. It consists of 5513 hand-classified tweets. These tweets are classified with respect to one of 4 different topics: Apple, Google, Microsoft and Twitter, including product reviews, company news, company stock discussion, company market analysis etc. Each entry contains: Tweet id, Tweet text, Tweet creation date, Topic used for sentiment and Sentiment label includes: ‘positive’, ‘neutral’, ‘negative’, or ‘irrelevant’. Only the Apple-Topic tweet dataset was used in this paper and Irrelevant-labeled tweets were discarded.

V. EXPERIMENT SETUP AND RESULTS

The machine learning based sentiment classification process consists of four main stages as shown in Figure 3. To evaluate the effectiveness of the proposed pre-processing method, three published machine learning based sentiment analysis systems and their associated pre-processing methods were implemented [6, 12, 17] and used as comparisons. First the complete sentiment classification system from each research paper was implemented and evaluated on the three financial-domain datasets; subsequently the pre-processing method from each sentiment classification system was replaced by the proposed pre-processing approach. This means that only the datasets and the pre-processing step are replaced in the experiments reported here, as compared to the original research paper.

A. Sentiment classification system introduction

A brief summary of each sentiment classification system in the three research papers utilised as comparisons is given below.

1. Chua’s sentiment classification system [6]:

Details of each process step in Chua’s system are listed and two additional features used in the system are introduced

- Pre-processing: Thread volatility measure and post removal
  - ↓ stopwords removal (stopwords list in NLTK) ↓ non-informative words filtering out with word correction (PyEnchant Package) ↓ word stemming (Porter algorithm)
- Feature extraction and selection: unigram feature extraction method with top 10,000 feature frequency (TF) ranking selected and Information Gain (IG) feature selection methods
- Feature term weighting: Term frequency- Inverse document frequency (TF-IDF)

Classifier: NB Classifier Performance measure: Average accuracy, precision, recall, F-score

In addition, two kinds of domain specific features, bigram and trigram financial terms feature set (examples are shown in Table 2) and stock price alert feature, are added into the unigram feature set; a sample of the complete feature set structure is provided in Figure 4. As all datasets were not collected on the basis of thread, thread volatility measure and post removal is discarded in all experiments.

2. Cohen’s sentiment classification system [17]:

Cohen’s system used NLP techniques to preprocess textual data which has four pre-processing steps similar to proposed pre-processing approach.

- Pre-processing: Detecting capitalized words ↓ punctuation replacing and removal ↓ lower casing ↓ replacing emoticons ↓ replacing URLs ↓ removing repeating characters ↓ replacing platform-specific characters ↓ stopwords removal ↓ replacing URLs ↓ removing repeating characters
- feature extraction: BOW model
- Feature term weighting: Feature presence method
- Classifier: NB classifier
- Performance measure: Average accuracy

3. Smailovic’s sentiment classification system [12]:

Smailovic selected several pre-processing steps and combined them in different ways to preprocess financial tweet dataset.

<table>
<thead>
<tr>
<th>Bigram features</th>
<th>interactive investor, firm decisions, capital management, absolute return, balance sheet, corporate profit, market sentiment, Add Oil, gross dividend, project management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigram features</td>
<td>Bad debt recovery, market value added, green field investment, present value interest, target risk fund, stock exchange price, weak form efficiency, wealth added index, real time quote, days working capital</td>
</tr>
</tbody>
</table>

Table 2. bigram and trigram financial terms feature example
• Feature extraction and selection: unigram and bigram feature extraction method with feature appearance threshold feature selection method
• Feature term weighting: TF-IDF method
• Classifier: SVM Classifier
• Performance measure: Average accuracy, precision, recall

This system implemented the sentiment classification in six combined pre-processing ways:
1) Bigram & Mini word frequency being 2 & link replacement
2) Bigram & Mini word frequency being 2
3) Bigram & Mini word frequency being 2 & username replacement
4) Bigram & Mini word frequency being 2 & username replacement & replacement
5) Bigram & Mini word frequency being 3
6) Unigram & Mini word frequency being 2

B. Sentiment classification experimental results

The proposed pre-processing method and the pre-processing methods in each baseline paper were implemented in NLTK in Python [34]. 5-fold cross validation sentiment classification was conducted. Average accuracy, precision, recall and F-score performance metrics given in [25] are employed in this paper for three-class sentiment classification performance analysis.

Several experiments were conducted to assess the performance of the proposed pre-processing approach and compare it against the selected existing approaches. The results achieved were compared with the different types of features resulting from each pre-processing methods and data transformation. Table 3, Table 4 and Table 5 summarize the experimental results.

The proposed pre-processing approach improved sentiment classification results as compared to the existing preprocessors by around 3% - 5% before some feature selection methods were applied, while the classification performance in Table 5 is slightly different from the other two tables when processing the ST Apple Twitter dataset on Smailovic’s sentiment classification system. Some sentiment classification performance achieved by using the proposed approach is slightly worse than the performance resulting from existing pre-processing approach in Table 5; the reason for this is that the tweets data has special means of communicating, such as shortening words, extensive use of emoticons, and the use of informal language expressions. The proposed pre-processing approach does not currently have steps to process these special language expressions in tweets. Smailovic’s sentiment classification system takes the characteristics of tweet text more into account and cleans more noise in the tweets.

The proposed pre-processing approach outperforms the NLP pre-processing method containing stopwords removal in comparison to Cohen’s sentiment classification system (see Table 4), as the proposed approach cleans more noise and leave more robust features in preprocessed text. Negation handling features and sentiment-shift-tagged features which were processed in the proposed pre-processing approach, combined with the existing feature matrix, build up robust features, enhancing the feature quality.

C. Negation handling methods implementation and results comparison

Three existing negation handling methods have been discussed earlier in the literature review section, including the BOW method [6, 7], contextual valence shifters [30]

Table 6 The classification performance in percentages using different negation handling approach

<table>
<thead>
<tr>
<th>BOW</th>
<th>CVS</th>
<th>Semantic</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>77.51</td>
<td>78.24</td>
<td>77.17</td>
</tr>
<tr>
<td>Precision</td>
<td>77.32</td>
<td>78.58</td>
<td>77.29</td>
</tr>
<tr>
<td>Recall</td>
<td>77.59</td>
<td>78.89</td>
<td>76.98</td>
</tr>
<tr>
<td>F-score</td>
<td>77.57</td>
<td>78.21</td>
<td>77.04</td>
</tr>
<tr>
<td>5-fold CV</td>
<td>77.91</td>
<td>78.43</td>
<td>77.39</td>
</tr>
</tbody>
</table>

and semantic composition [28, 29]. The three existing negation approaches were applied to the GKP stock forum dataset as one step in pre-processing process using Chua’s sentiment classification system 1 [5], as well as the developed negation handling approach. The sentiment classification performance of four negation handling approaches is provided in Table 6.

The difference in classification performance between the four negation handling methods is very slight. This can be explained by the fact that a large text containing many polar opinions may only have a small number of negation expressions, and the influence of negation is not strong enough to shift the sentiment polarity dramatically. The developed Semantic and syntactic combined negation approach achieved better classification performance in terms of precision, recall and 5-fold cross validation average accuracy shown in Table 6. The reason is that the approach extracts more negation terms by using WordNet to help identify more implicit negation expressions in text; and then the syntactic parser is used to identify how the negation terms interacts with other words within a sentence and which word(s’) meaning is/are shifted by the negation term(s) in the sentence. The developed negation handling approach provides a more accurate method to identify the sentiment impact of negation term(s) in one sentence.

VI. CONCLUSIONS AND FUTURE WORKS

Machine learning based sentiment analysis has been used for detecting opinions expressed in online user generated texts with applications in marketing and branding, social media monitoring, business analytics, financial decision-making, politics, etc. This work investigated the problem of extracting financially-relevant sentiment information from various sources such as news, message boards and microblogs. NLP techniques were used to preprocess online informal text. The paper proposed a pre-processing approach, integrating six NLP processing step, to reduce the noise in online informal text. The approach was employed in three existing sentiment classification methods from the literature. The results obtained show that better sentiment classification accuracy was achieved with the proposed text pre-processing approach applied on three datasets, in comparison to the pre-processing methods used in the original papers. The success of this model relies largely upon the fact that the proposed
approach cleans more non-informative noise in raw text and better prepares raw documents for feature extraction.

A future direction for this area of research is to investigate domain-specific knowledge in datasets, as sentiment polarity analysis is a domain-dependent task. It would be beneficial to integrate semantic understanding in sentences.

ACKNOWLEDGEMENT
Fan Sun is supported by a Vice-Chancellor’s Research Scholarship (VCRS) from the University of Ulster as part of the Capital Markets Engineering (CME) project. The CME project is supported by the companies and organizations involved in the N. Ireland Capital Markets Engineering Research Initiative (InvestNI, Citi, First Derivatives, Kofax, Fidessa, NYSE Technologies, ILEX, the University of Ulster and the Queen’s University of Belfast).

REFERENCES