International Journal of Investment Management and Financial Innovations
2017; 3(1): 1-9
http://www.aascit.org/journal/ijimfi
ISSN: 2381-1196 (Print); ISSN: 2381-120X (Online)

AASCIT American Association for Science and Technology

Keywords
Financial Sentiment,
Sentiment Analysis,
Text Categorization,
Text Classification

Received: March 25, 2017
Accepted: May 10, 2017
Published: August 23, 2017

Financial Sentiment Analysis Using Machine Learning Techniques

Sarkis Agaian, Petter Kolm

Department of Mathematics, New York University Courant Institute, New York, USA

Email address

sa1820@nyu.edu (S. Agaian), petter.kolm@nyu.edu (P. Kolm)

Citation

Sarkis Agaian, Petter Kolm. Financial Sentiment Analysis Using Machine Learning Techniques. *International Journal of Investment Management and Financial Innovations*. Vol. 3, No. 1, 2017, pp. 1-9.

Abstract

The rise of web content has presented a great opportunity to extract indicators of investor moods directly from news and social media. Gauging this sentiment or general prevailing attitude of investors may simplify the analysis of large, unstructured textual datasets and help anticipate price developments in the market. There are several challenges in developing a scalable and effective framework for financial sentiment analysis, including: identifying useful information content, representing unstructured text in a structured format under a scalable framework, and quantifying this structured sentiment data. To address these questions, a corpus of positive and negative financial news is introduced. Various supervised machine learning algorithms are applied to gage article sentiment and empirically evaluate the performance of the proposed framework on introduced media content.

1. Introduction

Online news, blogs, and social networks have become popular communication platforms to log thoughts and opinions about everything from world events to daily chatter. These opinion-rich resources attract attention from financial investors to understand the opinions of both businesses and individual users [1]. Market sentiment is the general prevailing attitude of investors as to anticipate price development in a market. This attitude is the accumulation of a variety of fundamental and technical factors, including price history, economic reports, seasonal factors, and national and world events. As more and more opinions are made available on websites, (such Twitter, Reddit, Facebook, Bloomberg Finance, Google Finance, Yahoo Finance, etc.) it is becoming increasingly difficult to analyze this large media content. For instance, the popular micro-blogging site, Twitter, has over 200 million active users, who post more than 400 million tweets a day [1].

Currently there is major interest in both industrial and academic research to use sentiment to analyze, classify, make predictions and gain insights into various aspects of daily life. A survey of sentiment analysis [2] has been cited over 5000 times in Google Scholar. Significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content [2] [3] [4] [5] [6] [7] [8] [9]. These works have laid the groundwork to address several challenges in developing a scalable and effective system for web dynamic sentiment analysis. These challenges include identifying useful information content, representing structured text under a scalable framework to determine sentiment, and extracting relationships between market trends and this quantified sentiment.

To address the challenges above one needs to develop a framework to automatically classify to classify financial news as positive or negative. Most previous research on

sentiment-based classification has been focused on non-financial content, such as movie reviews [7], travel and automobile reviews [10], and Amazon product reviews [11]. In the finance community, few papers [12] [13] have been published exploring sentiment in financial news. These works, however, use simple lexical algorithms to evaluate sentiment, and focus primarily on addressing whether the returns of a firm on a given day are connected to the news that was published about the firm on that day. This approach is unlike the strong connection between classification and text in a movie review [14].

This work focuses on developing the framework to perform more sophisticated sentiment analysis and learning on financial text. The performance of a sentiment analyses relies significantly on the qualities of the training and testing data. Unfortunately, the commonly used text collection such as the Reuters-21578, Amazon Product Review Data, and Cornell's movie review dataset cannot be used as a benchmark as they lack either sentiment or financial focus. As there is no publicly offered dataset with positive or negative financial news, a database of positive and negative financial texts is generated. The main remaining contributions of this work are to formally define the problem of using supervised sentiment analysis of financial texts to describe market trends and develop a supervised machine learning approach to gage financial sentiment. The goal of this work is to provide a prototype that can be leveraged to represent unstructured large financial texts that is efficient, accurate and scalable.

The remainder of this paper is organized as follows. Section 2 reviews existing literature related to this paper and formally define the problem of study. Section 3 proposes an automated supervised sentiment analysis framework. Section 4 introduces the generated financial database of positive and negative financial news. Sections 5 and 6 present the simulation and cross-validation results, respectively. Section 7 concludes and discusses directions for future research.

2. Background, Related Work, and Challenges

This section presents an overview of sentiment analysis and machine learning literature that is related to current work and the problem statement.

2.1. Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [2]. This problem of automatic text classification and categorization has spread to almost every possible domain and has grown to be one of the most active research areas in machine learning and natural language processing [2]. It has a very high commercial potential in fields like finance, where individuals seek analyze large texts of information of businesses and their customers - studies indicate, for instance, that 80% of company's information was contained in text documents such as emails, memos, and reports [1].

Sentiment analysis dates back to the 1990s [15] [16] [17]. Fama [18] previously demonstrated that emotions have an effect on rational thinking and social behavior. Hatzivassiloglou and McKeown [19] develop an algorithm for predicting semantic orientation. They classify positive, negative and neutral expressions in texts by using a small set of manually annotated seed words. Their algorithm performs well, but it is designed for isolated adjectives, rather than phrases containing adjectives or adverbs. Hatzivassiloglou and Wiebe [20] show the effects of adjective orientation and gradability on sentence subjectivity. Turney and Littman [21] present an unsupervised approach for classifying positive and negative terms. For additional works, readers are referred to: [22] [23] [24] [25] [26] [27] [28] [29] [30] [49] [50] [51] [52] [53] [54] as well as the extensive reviews [6] [31].

Sentiment analysis started being adopted in finance with the introduction of works such [13] [32] [33], which use sentiment analysis of weblog and news data to predict stock price moves. Nofsinger [33] demonstrates that the stock market itself can be considered as a measure of social mood. Gilbert and Karahalios [34] have found out that increases in expressions of anxiety, worry and fear in weblogs predict downward pressure on the S&P 500 index. Bordino et al. [35] show that trading volumes of stocks traded in NASDAQ-100 are correlated with their query volumes (i.e., the number of users' requests submitted to search engines on the Internet). Thelwall et al. [36] analyze events in Twitter and show that popular events are associated with increases in average negative sentiment strength. Nofer [37] Bollen et al. [38] have found that changes in a specific public mood dimension (i.e., calmness) can predict changes in stock price. Ruiz et al. [39] use time-constrained graphs to study the problem of correlating the Twitter micro-blogging activity with changes in stock prices and trading volumes. Smailović, et al. [40] use the volume and sentiment polarity of Apple financial tweets to identify important events and future movements of Apple stock prices.

The sentiment analysis can be divided into two key classes: supervised and unsupervised [7]. A conventional way to perform unsupervised sentiment analysis is the lexicon-based method [3] [8] [29]. This is the primary method employed in the financial work listed above, primarily due to its simplicity in algorithm and implementation. The lexicon-based methods employ a sentiment lexicon to determine overall sentiment polarity of a document. Since they disregard context and semantic structure, they are less accurate than other unsupervised and supervised methods; they have also become increasingly difficult due to the distinct language of social media, where short-unstructured texts with expressions such as "it's coooool" and "good 9t:)" are commonplace [5] [31] [42]. Thus, it is difficult to define a

universally optimal sentiment lexicon to cover words from different domains [42].

Most research on non-lexical sentiment-based classification not been focused on financial texts. Research has centered on: movie reviews [2] [7]; automobiles and travel destinations reviews [10]; and product reviews from Amazon [11]. Reviews have been used to generate datasets as reviewers often summarize their overall sentiment with a rating indicator, such as number of stars, thereby eliminating the need to hand label data [7].

Public, classified datasets have not been introduced for financial texts. Open research issues in sentiment analysis include [43]:

a) A need for better modeling of compositional sentiment. At the sentence level, this means more accurate calculation of the overall sentence sentiment of the sentiment-bearing words, the sentiment shifters, and the sentence structure.
b) A need design and implementation a dataset with positive or negative financial news.
c) A need to de-noise the noisy texts (those with spelling/grammatical mistakes, missing/problematic punctuation and slang)

This work addresses these issues by introducing a new dataset and framework to assemble financial text of single names and gage their sentiment. Traditional machine learning methods are first overviewed.

2.2. Machine Learning Methods

There are many classification systems that rely on machine learning methods including k-Nearest Neighbors (simple, powerful), Naive Bayes (simple, very efficient as it is linearly proportional to the time needed to read in all the data); Support-Vector Machines (relatively new, more powerful); K- Nearest Neighbor classification (simple, expensive at test time, high variance, non-linear); Vector space classification using centroids and hyperplanes that split them (simple, linear discriminant classifier); and AdaBoost (based on creating a highly accurate prediction rule using a weighted linear combination of other classifiers). Many commercial systems use a mixture of methods. The Naive Bayes and SVMs are currently among the best performers for a number of classification tasks including text data [44] [45] [46] [47]. This work uses the methods listed below.

2.3. Classification Based on PMI-IR Algorithm

[10] Point-wise mutual information is a semantic word similarity measure between two words Word_1 and Word_2 and defined as

$$PMI(Word_1, Word_2) = \log \frac{\Pr(Word_1, Word_2)}{\Pr(Word_1)\Pr(Word_2)} \quad (1)$$

Where $\Pr(Word_1, Word_2)$ is the probability that Word_1 and Word_2 occur at the same, and where $\Pr(Word_i), i = 1, 2$ the number of is times that $Word_i$ appears in the corpus.

PMI, in other words, is the probability of observing words, $Word_1$ and $Word_2$, together. $PMI(Word_1, Word_2)$ is the amount of information that is acquired about the presence of one of the words when the other is observed. It is equal to zero if two words Word_1 and Word_2 are statistically independent

$$\Pr(Word_1, Word_2) = \Pr(Word_1)\Pr(Word_2) \quad (2)$$

of each other. Moreover, it is *positive* if they are positively correlated and negative if they are negatively correlated. For example words Word_1 and Word_2 could be:

Table 1. *Potential POS Pairs.*

	$Word_1$	$Word_2$
1.	Adjective	Noun
2.	Adverb	Adjective
3.	adjective	Adjective
4.	noun	Adjective
5.	adverb	Verb

Presented below is the unsupervised algorithm applied to movie reviews as introduced by Turney [10]. Turney's PMI-IR algorithm uses words *excellent* and *poor* as seed words. These seed words can be looked as proxies for the category labels of "positive" or "negative."

Table 2. *PMI-IR Algorithm.*

Input:	Text-review
Step 1:	Identify phrases that contain adjectives or adverbs by using a part-of-speech tagger. Define a distance measure d(t1, t2) ($PMI(Word_1, Word_2)$) between terms t1 (adjectives) and t2 (adverbs). Extract two consecutive words: one is an adjective or adverb, the other provides the context.
Step 2:	Estimate the semantic orientation of each phrase based on their association with database positive and seven negative words by using SO(phrase) = PMI(phrase, "positive") – PMI(phrase, "negative") (3) Note: Semantic Orientation is positive when phrase is more strongly associated with "excellent" and negative when phrase is more strongly associated with "poor".
Step 3.	Calculate the average semantic orientation (SO) of the phrases.
Step 4.	Classify the review as recommended if average SO is positive, not recommended otherwise. If hits(phrase NEAR "excellent") and hits(phrase NEAR "poor")≤4, then eliminate phrase

2.4. Classification Using Naive Bayes

Recalling the classification problem for the widely used algorithm:

Given a data: examples of the form (d, h(d))

where d are the data objects to classify (inputs)

and a fixed set of classes: H = { h(1) h(2),…,h(K)}or h(d) are correct class info for d, h(d)∈ {1,…K}

Determine: given d_{new}, provide $h(d_{new})$

Bayesian methods provide the basis for probabilistic learning methods that use knowledge about the prior probabilities of hypotheses and about the probability of observing data given the hypothesis. Naive Bayes Classifier is a Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attributes are independent given the class. The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier. It uses Bayes' Rule

$$p(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)} \tag{4}$$

where

d : data

h : hypothesis

$P(h)$: prior belief (probability of hypothesis h before seeing any data)

$P(d \mid h)$: likelihood (probability of the data if the hypothesis h is true)

$P(d) = \sum_{h} P(d \mid h)P(h)$: data evidence (marginal probability of the data)

$P(h \mid d)$: posterior (probability of the hypothesis h after having seen the data d)

The key approach to this type of text categorization is to assign to a given text d to class H = {h(1) h(2),…,h(K)}:

2.5. Classification using Support Vector Machines (SVMs)

SVM is a supervised learning algorithm developed by Vapnik and his co-workers [48]. SVM represents a powerful technique for general (nonlinear) classification, regression and outlier detection with an intuitive model representation. They are extremely effective in many applications including bioinformatics, signal/image recognition, and other fields. Moreover, SVM has been shown to be very effective at traditional text categorization, mostly outperforming Naive Bayes and maximum entropy classifiers [7] [46].

The basic idea behind SVMs is to find the optimal separating hyper-plane between the two classes by maximizing the margin between the positive and negative classes' closest points the points lying on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyper-plane. This can been achieved by solving an optimization problem: Letting training set $\{(x_i, h_i)\}_{i=1,2\ldots,n}$, $x_i \in R^m$, $h_i \in \{-1, 1\}$ where the h_i is either 1 or −1, indicating the class (corresponding to positive and negative) to which the text-data x_i belongs. The goal is to find the maximum-margin hyperplane that divides the points having h_i=1from those having h_i =-1. Any hyperplane can be written as the set of points x satisfying w.x-b=0, where . denotes the dot product and w the normal vector to the hyperplane. The parameter b/||w|| defines the offset of the hyperplane from the origin along the normal vector w.

The distance between these two hyperplanes is:

$$\text{Margin} = |(1/\|w\|) - (-1/\|w\|)| = 2/\|w\| \tag{5}$$

To minimize ||w|| and avoid data points from falling into the margin, the following constraint is used:

$wx_i - b \leq -1$ for x_i of the first class (corresponding let's say to positive news)

$wx_i - b \geq 1$ for x_i of the second class(corresponding to negative news) (6)

This can be rewritten as:

$$h_i\,(wx_i - b) \geq 1 \text{ for all i, i=1,2,….,n} \tag{7}$$

Consequently the optimization problem can be formulated as:

Minimize ||w||

Subject to

$$h_i\,(wx_i - b) \geq 1, \text{ for any i=1,2,…n} \tag{8}$$

It has been shown that the optimization problem solution can be expressed as a linear combination of the training vectors:

$$w = \sum_{i=1}^{n} \alpha_i h_i x_i \tag{9}$$

Where the α_i 's are obtained by solving a dual optimization problem and the corresponding x_i are exactly the support vectors, which lie on the margin and satisfy h_i $(wx_i - b) = 1$.

3. Proposed Supervised Binary Sentiment Categorization

The goal of this section is to introduce a framework for building a model to automatically classify sentiment of new, unlabeled financial documents. Presented are both a binary sentiment-categorization and multi-class sentiment-categorization method using machine learning methods such as the SVM algorithm illustrated above.

Methods and Problem Definition: Sentiment classification includes two key classes: binary sentiment classification and multi-class sentiment classification. The dataset of n instances D = $((f_1, d_1),..., (f_n, d_n))$, where f_i is the feature vector extracted from the i-th data instance, and d_i is the label for that particular data instance and a pre-defined categories. The labels in this work are "positive" or "negative" for the binary classification.

Below explains how features are selected in this work. A summary of the supervised learning algorithms/architecture implemented in this work is summarized in the following figure and table.

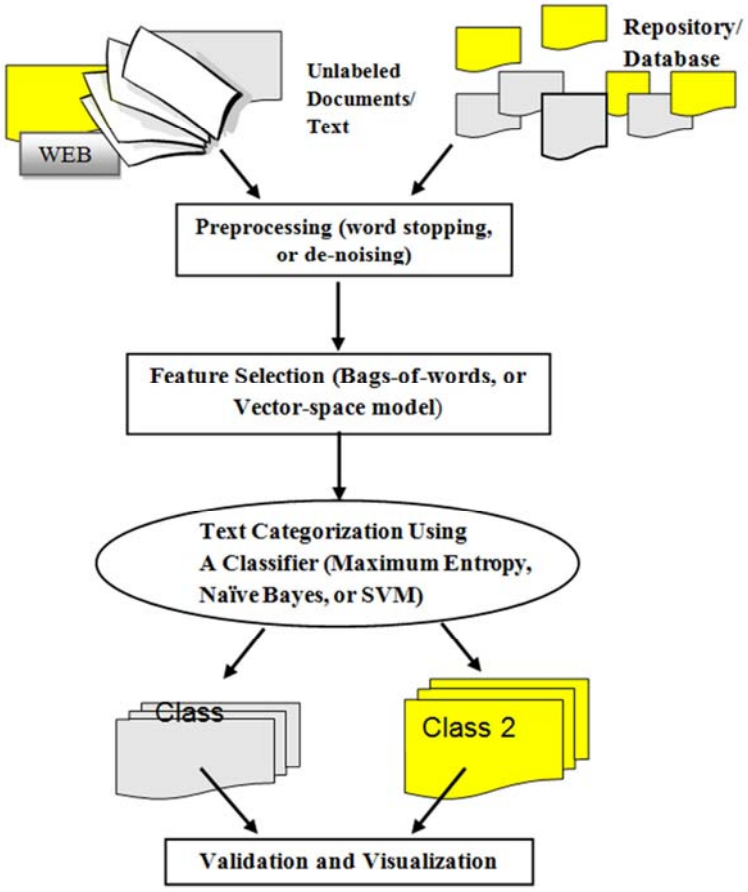

Figure 1. *Proposed Text Sentiment Analysis Framework and Processing Steps.*

Table 3. *Summary of Presented Approach.*

Steps	Proposed text sentiment analysis framework
1.	Data extraction and collection from unstructured data sources a) Gather articles or titles from various sources and time periods b) Create of training set using articles or titles from various sources and time periods c) Generate a database of set of words (including positive and negative; adjectives with positive and negative orientations)
2.	Preprocessing a) Apply dimensionality reduction algorithm b) Use of dictionary, stemming, and stop-words to filter corpus
3.	Feature selection and extraction a) Use of part-of speech tagger to identify phrases that contain adjectives or adverbs. b) Choose a similarity metric for pair words and phrases c) Assign the extracted phrases to a class, recommended or not recommended, based on measurements of the semantic orientation d) Classify articles/text as "Positive" and "Negative"
4.	Text categorization using a classifier (Maximum Entropy, Naïve Bayes, or SVM)
5.	Evaluate performance using *k*-fold cross validation

In the preprocessing step, stop-words (such as a, the, about, above, after, again, all, alone, along, already of, that, etc) are filtered out as they do not carry information. Also note that in feature selection and extraction step point-wise mutual information (PMI) maybe used as a features.

4. Financial Sentiment Corpus

As noted above there is no publicly offered dataset with positive or negative financial news. This section presents an implementation of a dataset with positive and negative financial news sourced from Seeking Alpha, where articles were selected where authors provide disclosure of either being or intending to be long or short a stock. Several factors are considered in using Seeking Alpha to implement the dataset: historical and future data perspectives, the diversity of the financial news user community, data integrity, and presence of sentiment. The constructed database has 501 documents from January 2011-January 2014 covering 125 companies in the SPX 500. A breakdown of the companies is provided below:

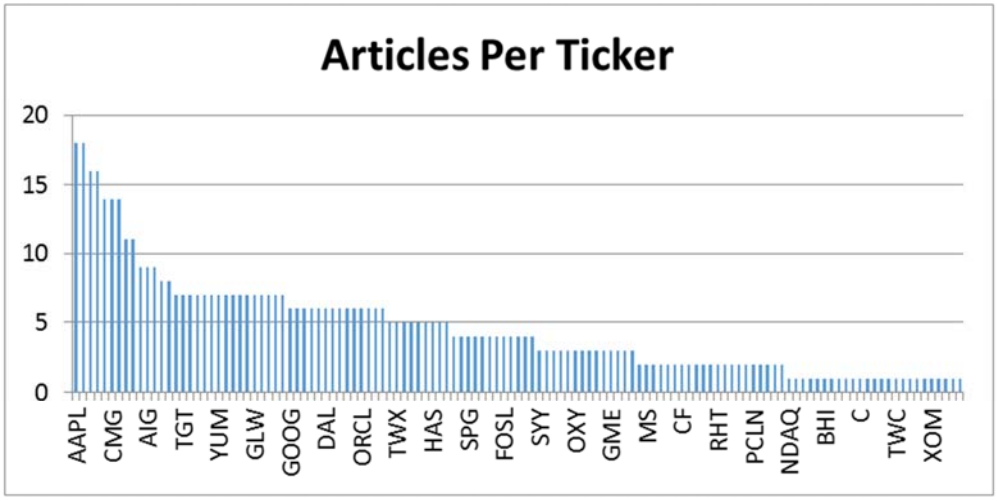

Figure 2. *Company Articles Breakdown.*

Six students, working in pairs, were asked to manually categorize the dataset into two sets of categories: positive and negative, on both an article and sentence level. Articles and sentences below a minimum consensus threshold were filtered out. The breakdown of this categorization is summarized in the table below:

Table 4. *Manual Categorization Breakdown.*

Level	Positive	Negative	Total
Article	251	250	501
Sentence	2743	2346	5089

An illustration of a sample article in XML format is provided below. The rating is defined as the median rating assigned by each of the three groups on a range of strong sell to strong buy [-3,3].

<?xml version="1.0" encoding="UTF-8"?><response><result start="0" numFound="501" name="response"><doc><str name="twitter_title">Bully For BlackRock</str><str name="keywords">NYSE:BLK</str><str name="url">http://seekingalpha.com/article/1345881-bully-for-blackrock</str><arrname="ratings_man"><int>0</int><int>3</int><int>0</int><int>2</int><int>2</int><int>0</int><int>0</int><int>0</int><int>0</int><int>0</int><int>0</int><int>0</int><int>1</int><int>0</int><int>2</int><int>0</int><int>2</int><int>2</int><int>2</int><int>2</int><int>0</int><int>0</int><int>0</int><int>0</int></arr><int name="sentiment_man">3</int><arr name="sentences"><str>Apr. 16, 2013 6:04 PM ET | About: BLK by: Bret Jensen I have owned and written about BlackRock (BLK) since October.</str><str>It is a core holding in my income portfolio, as it has a solid yield and has raised its payouts tremendously over the years -- even through the financial crisis.</str><str>The shares have gone from $187 to $257 in that time.</str><str>The company continues to show solid growth as its businesses rise along with the equity and credit markets.</str><str>Also, it has consistently beat earnings estimates.</str><str>The recently reported quarter was no exception.</str><str>Here are the quarterly earnings highlights: EPS came in at $3.65 a share, seven cents above consensus estimates.</str><str>Sales came in slightly above consensus led by iShares revenues, which were up more than 20% year over year.</str><str>AUM increased 7% year over year to $3.94 trillion.</str><str>Equity funds saw net inflows of over $33 billion.</str><str>Adjusted operating margins increased 140bps to 40%.</str><str>BlackRock is one of the largest investment managers in the world.</str><str>The firm provides its myriad services to institutional, intermediary, and individual investors.</str><str>Here are four reasons why BLK still has upside from $257 a share: Consensus earnings estimates for both FY 2013 and FY 2014 had consistently and significantly gone up before this earnings report.</str><str>FY 2014's projections are $1 a share above where they were 90 days ago.</str><str>I would look for further upward revisions after these quarterly results.</str><str>BlackRock is well positioned for the migration from mutual funds to ETFs.</str><str>It also should do well as cash starts to come back into the markets as the Fed continues to encourage investors to move into riskier assets.</str><str>Finally, it does not have a huge gold ETF like State Street (STT), which is seeing major outflows.</str><str>The company has now beat or met quarterly earnings estimates for 13 straight quarters (12 beats, one meet).</str><str>BlackRock is expected to grow revenues at a 10% CAGR over the next two years and is selling for just over 14x 2014's projected earnings.</str><str>BLK yields 2.6% and has quadrupled dividend payouts over the last six years or so.</str><str>The stock has a reasonable five-year projected PEG (1.25) for a dividend payer. </str></arr><arr name="disclosure_sa"><str>I am long BLK.</str></arr><arr name="author_sa"><str>Bret Jensen</str></arr></doc></result></response>

5. Simulation

Simulations were conducted using the supervised learning algorithms Naïve Bayes, Max Entropy, and SVM, with an average classification accuracy of ~75%. Detailed results along with traditional measures for accuracy are provided below.

Precision, Recall, and *F-measure* are traditional measures that have been widely used by text categorization algorithms for performance evaluation. The *F*-measure is defined as a combination of *Precision* and *Recall*:

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (10)$$

where *Precision* is the percentage of selected items that are correct and *Recall* is the percentage of correct items that are selected. Table 5 below summarizes the performance measures in terms of precision, recall, average F-measure for all article and sentence level classification for the SVM and Max Entropy classifiers.

Table 5. *Performance of SVM and Ma Entropy Classifiers on Dataset.*

		SVM			Max Entropy		
Level	**Category**	**Precision**	**Recall**	**FScore**	**Precision**	**Recall**	**FScore**
Sentence	Negative	71	65	68	70	69	69
Sentence	Positive	73	78	75	75	75	75
Article	Negative	79	76	77	75	77	76
Article	Positive	77	81	79	77	75	76

6. Cross Validation

Evaluating classification performance is important for several reasons: (1) when building classifiers, the parameters used for classification can be tuned. For example, at this point, several tests should be done in order to choose the best features, to explore data quality, and so on. (2) When evaluating given classifiers, it can be determined whether they are good enough for the purpose or whether they provide sufficient improvement over an existing method to merit switching.

One of the most used methods for estimating classification performance is cross-validation. Cross-validation can be done by applying three different schemes namely *k*-fold, hold-out, and leave-one-out. The basic form of cross-validation is *k*-fold, and the other schemes are special cases derived from *k*-fold. Implementation of cross-validation methods is done as follows:

a) *k*-fold validation: To implement this method, the data is randomly divide into *k* equally (or nearly equally) folds. Next, *k* iterations of training and test are carried out, such that at each iteration a different segment is held out for validation and *k*–1 folds are used to fit the model.
b) Hold-out validation: For this scheme, the dataset is split into two non-overlapped segments: one for training and the other for testing. Hold-out avoids some samples to be used for both learning and validation, yielding a better estimation for the generalization performance of the algorithm used for recognition.
c) Leave-one-out validation: This is a special case of *k*-fold cross-validation, where *k* is the number of data. In this scheme only one sample is held-out for testing. The results of leave-one-out cross-validation are considered to be almost unbiased, but they have large variances.

In order to obtain reliable performance estimation of a given classifier, it is recommended to have a large number of iterations. Table 6 below provides the *10*-fold cross validation results for the SVM classifier on both an article and sentence level:

Table 6. *Cross Validation Results.*

	Accuracy %	
Fold	**By Article**	**By Sentence**
1	78.33	70.02
2	84.21	72.84
3	90.24	75.00
4	83.33	74.95
5	82.35	72.04
6	83.72	74.65
7	75.56	74.42
8	72.73	68.81
9	78.26	71.24
10	80.70	74.14

7. Conclusion

This study demonstrates the promising attempt to incorporate supervised sentiment classification into financial article analysis. More specifically, a database of positive and negative financial news is generated and used to train supervised machine algorithms to gage article sentiment. Simulation and cross validation results are provided to evaluate sentiment categorization. This work may be leveraged as a prototype to represent unstructured large financial texts in an efficient, accurate and scalable sentiment analysis framework. Future work can build more robust categorization systems based on new classifiers and features and enhance classification accuracy through the incorporation of more news sources and authors.

Acknowledgements

The following students are acknowledged for their hard work in helping create and analyze this dataset. In alphabetical order, Vidya Akavoor, Vivaan Dave, Terrance Liang, Jing Lin, Sahr Singh, Jonathan Turner, and Lena Woo.

References

[1] X. Hu, J. Tang and H. Liu, "Unsupervised Sentiment Analysis with Emotional Signals," in *International World Wide Web Conference Committee*, Rio de Janeiro, 2013.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval,* vol. I, no. 2, pp. 1-135, 2008.

[3] B. Connor, R. Balasubramanyan, B. Routledge and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of ICWSM*, 2010.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2004.

[5] X. Hu, N. Sun, C. Zhang and T. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proceedings of CIKM*, 2009.

[6] B. Liu, Handbook of Natural Language Processing, Boca Raton: CRC Press, Taylor and Francis Group, 2010.

[7] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of ACL*, 2002.

[8] J. Wiebe, T. Wilson and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation,* vol. 39, no. 165, pp. 165-210, 2005.

[9] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of HLT and EMNLP*, 2005.

[10] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the Association for Computational Linguistics*, 2002.

[11] K. Dave, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *WWW2003*, 2004.

[12] P. Tetlock, M. Saar-Tsechansky and S. Macskassy, "More Than Words: Quantifying Language to Measure Firms," *Journal of Finance,* vol. 68, pp. 1437-1467, 2008.

[13] P. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance,* vol. 62, no. 3, pp. 1139-1168, 2007.

[14] P. Azar, "Sentiment Analysis in Financial News," Harvard College (Thesis), Cambridge, Massachusetts, 2009.

[15] S. Argamon-Engelson, M. Koppel and G. Avneri, "Style-based Text Categorization: What Newspaper Am I Reading?," AAAI, 1998.

[16] B. Kessler, G. Nunberg and H. Schautze, "Automatic Detection of Text Genre," in *ACL*, 1997.

[17] E. Spertus, "Smokey: Automatic recognition of hostile messages," in *Proceedings of Innovative Applications of Artificial Intelligence*, 1997.

[18] E. Fama, "Random Walks in Stock Market Prices," *Financial Analysts Journal,* vol. 21, no. 5, pp. 55-59, 1965.

[19] V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," in *Proceedings of ACL*, 1997.

[20] V. Hatzivassiloglou and J. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," in *Proceedings of International Conference on Computational Linguistics*, 2000.

[21] P. Turney and M. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus," 2002.

[22] P. Chaovalit and L. Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches," Proceedings of Annual Hawaii International Conference on System Sciences, 2005.

[23] A. Finn, N. Kushmerick and B. Smyth, "Genre Classification and Domain Transfer for Information Filtering," in *Proceedings of European Colloquium on Information Retrieval Research*, 2002.

[24] S. Durbin, D. Warner, J. Richter and Z. Gedeon, "Information Self-Service with a Knowledge Base That Learns," *AI Magazine,* vol. 23, no. 4, pp. 41-50, 2002.

[25] M. Efron, "Cultural orientations: Classifying subjective documents by cocitation analysis," in *Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music*, 2004.

[26] D. Inkpen, O. Feiguina and G. Hirst, "Generating more-postive and more-negative text," in *Computing Attitude and Affect in Text: Theory and Applications*, Dordrecht, The Netherlands, Springer, 2005, pp. 187-196.

[27] M. Gamon, "Sentiment classification on customer feedback data noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[28] J. Wiebe and E. Riloff, " Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," in *Computational Linguistics and Intelligent Text Processing*, 2005.

[29] T. Wilson., J. Wiebe. and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," *Computational Linguistics,* vol. 35, no. 3, pp. 399-433, 2009.

[30] Y. Dang, Z. Yulei and H. Chen, "A lexicon enhanced method for sentiment classification: An experiment on online product reviews," *IEEE Intelligent Systems,* vol. 25, no. 4, pp. 46-53, 2010.

[31] B. Liu, Sentiment Analysis and Opinion Mining, Claypool Publishers, 2012.

[32] P. Tetlock, M. Saar-Tsechansky and S. Macskassy, "More than words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance,* vol. 63, no. 3, pp. 1437-1467, 2008.

[33] G. Mishne, "Prediciting Movie Sales from Blogger Sentiment," in *Computational Approaches to Analysing Weblogs*, 2006.

[34] J. Nofsinger, "Social Mood and Financial Economics," *Journal of Behavioral Finance,* vol. 6, no. 3, pp. 144-160, 2005.

[35] E. Gilbert and K. Karahalios, "Widespread Worry and the Stock Market," in *Proceedings of the International*, 2010.

[36] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen and I. Weber, "Web Search Queries Can Predict Stock Market Volumes," *PLoS ONE,* vol. 7, no. 7, p. e40014, 2012.

[37] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 12, pp. 2544-2558, 2010.

[38] M. Nofer, Using Twitter to Predict the Stock Market: Where is the Mood Effect?, New York: Springer, 2015.

[39] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2, no. 1, pp. 1-8, 2011.

[40] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis and A. Jaimes, Correlating financial time series with micro-blogging activity, ACM Press, 2012.

[41] J. Smailovic, M. Grcar and M. Znidaršic, "Sentiment analysis on tweets in a financial domain," in *International Postgraduate School Students Conference*, 2012.

[42] Y. Lu, M. Castellanos, U. Dayal and C. Zhai, "Automatic construction of a context-aware sentiment lexicon: an optimization," in *Proceedings of WWW*, 2011.

[43] R. Feldman, "Techniques and Applications for Sentiment Analysis," *Communications of the ACM,* vol. 56, no. 4, pp. 82-89, 2013.

[44] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning,* vol. 29, pp. 103-130, 1997.

[45] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the Tenth European Conference on Machine Learning*, Berlin, 1998.

[46] T. Joachims, "Estimating the generalization performance of a SVM efficiently," LS VIII-Report, Universit at Dortmund, Germany, 1999.

[47] B. Schlkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.

[48] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[49] M. Nardo, M. Petracco and M. Naltsidis, "Walking Down Wall Street with a Tablet: A Survey of Stock Market Predictions Using the Web," *Journal of Economic Surveys,* vol. 30, no. 2, pp. 3556-369, 2016.

[50] B. Agarwal and N. Mitta, " Machine Learning Approach for Sentiment Analysis," in *Prominent Feature Extraction for Sentiment Analysis,* Springer, 2015, pp. 21-45.

[51] M.-Y. Day and C.-C. Lee, "Deep learning for financial sentiment analysis on finance news providers," in *IEEE/ACM International Conference*, 2016.

[52] S Das and A. Das, "Fusion with sentiment scores for market research," in *Information Fusion International Conference*, 2016.

[53] A. Akansu, S. Kulkarni and D. Malioutov, Financial Signal Processing and Machine Learning, John Wiley & Sons, 2016.

[54] D. D. Wu and D. L. Olson, "Financial Risk Forecast Using Machine Learning and Sentiment Analysis," in *Enterprise Risk Management in Finance*, Palgrave Macmillan, 2015, pp. 32-48.