Leveraging Latent Economic Concepts for Trend Prediction in Financial Markets

Saeede Anbaee Farimani
Dept. of Computer Engineering
Islamic Azad University
Mashhad, Iran
anbaee@mashdiau.ac.ir

Majid Vafaie Jahan
Dept. of Computer Engineering
Islamic Azad University
Mashhad, Iran
VafaeiJahan@mshdiau.ac.ir

Amin Milani Fard
Dept. of Computer Science
New York Institute of Technology
Vancouver, Canada
amilanif@nyit.edu

Gholamreza Haffari
Faculty of IT
Monash University
Melbourne, Australia
gholamreza.haffari@monash.edu

***Abstract*—Fundamental analysis techniques that is based on news for financial market prediction, depends on the news document representation methods. Most of the existing market prediction techniques disregard latent relations in the news stream. In this work we consider the conceptual relationship between news documents and propose an approach based on latent topic modeling for document representation. We build a corpus of economic news related to the Foreign Stock Exchange Market (FOREX), and apply word embedding to represents semantic and syntactic relations among words. We then cluster the word vectors to create latent economic concepts and finally propose a document representation method based on the distribution of words on these concepts. Each document is labeled according to changes in the close prices in the market, and SVM classifier is used for prediction. Results show that the accuracy of the proposed approach is improved by X% compared to the state-of-the-art methods and this is achieved without using other features such as financial indices or sentiment analysis.**

***Keywords*— financial market prediction, topic modeling, latent economic concept, word embedding, document representation**

I. Introduction

In the past decades technical and fundamental analyses have been used in behavioral economics to study the influence of information on investors (Cutler, Poterba, & Summers, 1988; Fama, 1965). Existing predictive models for financial markets apply text mining on unstructured data and analyze sentiments in news documents without considering latent relations amongst news with similar topics methods(Ho, Damien, Gu, & Konana, 2017; Krishnamoorthy, 2018; X. Li et al., 2016; Van de Kauter, Breesch, & Hoste, 2015) . Different topics in the news have different impact and influence on investors. For instance, a political news about war does not have the same effect as an economic report by the World Bank (Tetlock, 2010). Therefore, topics and concepts should be incorporated in document representation so that conceptual similarity of documents help in the market prediction process.

Words are put together to make texts with different topics and what gives a meaning to a text is the semantic and syntactic relation that is made by the placement of phrases. Most of the market prediction techniques use Bag-of-words (BoW) model (Salton, 1989) for document representation (Hagenau, Liebmann, & Neumann, 2013; Schumaker, Zhang, Huang, & Chen, 2012; Seifollahi & Shajari, 2019) , which does not consider the semantic relation between the words and has limitations with high dimensionality and sparse representation. Other approaches utilize ontologies and tools for determining syntactic roles in order to model semantic relations between words (Dang & Duong, 2016), however, as data growth, this method suffers from the curse of dimensionality and fails to preserve words with different meanings in different domains in ontologies (Krishnamoorthy, 2018). One of the solutions to this challenge is semantic concepts modeling based on the probabilistic features of the occurrence of contextual words close to each other with word embedding technique (Hu, Zhang, Hou, & Li, 2017).

Semantic concepts modeling aims at the creation of low dimensional vector for text documents representation whereas the document vectors with similar contextual information are located close to each other in the embedded space (Hu et al., 2017; Kim, Kim, & Cho, 2017). In this method, at first vector representations of words are generated based on semantic relations between words using word embedding (Mikolov, Chen, Corrado, & Dean, 2013). The advantage of word embedding is in generating similar vectors for words that are close to each other in the conceptual window (Ma & Zhang, 2015). Then the corresponding word vector clustered and each cluster is considered as a latent semantic concept. Eventually, document vectorization is done based on its word distribution through latent semantic concepts.

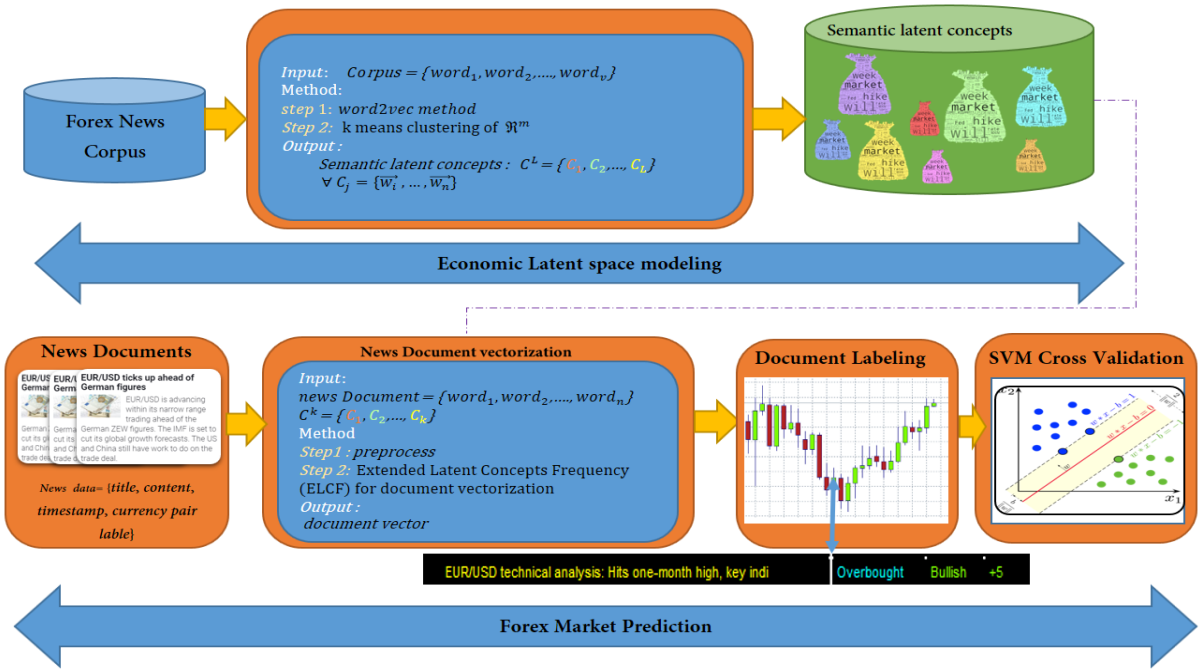

Figure 1- flowchart of proposed method

In this paper, given the dataset of forex news, we aim at predicting labels uptrend, downtrend or neutral for the EUR/USD currency pair. We proposed a predictive model for the Forex market based on currency-pairs related news. In the proposed method, with considering latent relations in he news stream and different influences of news topics on the market, we propose an extended latent economic semantic modeling method. Section II describes our approach and section III presents a case study on the impact of news on the Forex market and the base currency-pair EUR/USD. We discussed the evaluation method in section IV and in section V conclusion is mentioned.

II. LITRATURE REVIEW

Studies about investors' behavior by arriving information via various forms of news feeds are always important(Wang, Lu, & Zhao, 2019). After the presentation of the Efficient Market Hypothesis by Fama in 1965, the idea of complete randomness of the market is rejected and behavioral economics was motivated (Fama, 1965). So far, many methods used text mining techniques for knowledge extraction from unstructured text data and improvement of market prediction accuracy by machine learning methods (Ho et al., 2017; Khadjeh Nassirtoussi, Aghabozorgi, Ying Wah, & Ngo, 2015; Krishnamoorthy, 2018; X. Li et al., 2016; Van de Kauter et al., 2015). But most of which focus on the sentiment analysis of news headlines while very few consider the information hiding in the relationship between news stories which may be reported the different aspects of one event. A Few methods considered the different influences of political, social or cultural events on investor behavior, and use the information retrieval techniques for identifying important events and stock return prediction based on which (Mariana, Neves, & Horta, 2017; Wei, Lu, Chen, & Hsu, 2017; Wong & Ko, 2016), where (Birz, 2017; Gurin, Szymanski, & Keane, 2017; Tetlock, 2010, 2011) shows not only important events influence on the market, also common events like FED announcements have their impression. Besides, investors overreact to stale economic news and high media pessimism predicts downward pressure on the market prices followed by a reversion to fundamentals.

Hence, for the profit of both information in the content and relations between news items, it is better to structured news items based on its latent concepts deduct from news texts. A latent concept refers to a cluster of semantically related words that occur in similar contexts(Zhang, Li, & Wang, 2019). For

instance, the word 'Brexit' often occurs with the 'political', 'UK', 'exit' and ' parliament' in one concept. Document representation based on these economic concepts may capture the relation between news stories that are systematically related to each other. In many of the presented methods in the field of fundamental market analysis, the title of the news is analyzed (Blasch, 2010; Gurin et al., 2017; Long, Song, & Tian, 2019; Verma, Dey, & Meisheri, 2017), while there is some valuable information on the news content.

The essay of words in the news headline and content formed the news story, while words in the headline introduce the subject and the content words describe it. When the reader is skimming the text, the composition of these words induces a subjective conceptual in his mine. The news headline as its subjective core contains the words that most express the subject and the content is in some way a description of the subject of the news document. Document vectorization based on Latent semantic concept modeling reflects the subject in the process of text structuring whereas similar news in topics has vectors located close to each other in the embedded space.

The foreign exchange market, called Forex, is a financial markets in which investors trade based on a variety of currency pairs. Like other financial markets, the market is also affected by the news(Égert & Kočenda, 2014; Kočenda & Moravcová, 2018; Seifollahi & Shajari, 2019). There are a few method for fundamental analysis of forex market with text mining method (Arman Khadjeh Nassirtoussi, 2014), which may be due to the lack of a benchmark news dataset.

In this paper, our main contribution is evaluating the effectiveness of leveraging latent concept modeling for trend prediction in Forex Market. Specific contributions include the following items, not previously reported in related data-mining literature: (a) organizing the news dataset related to Forex Market which we crawler and store in MongoDB dataset. (b) Structuring the news texts with latent concept modeling as the state of art model for deep learning based text classification method (c) we proposed Extended Economic Latent concept modeling for benefit all information in news title and content. (d) Experimenting the interpretability of proposed method.

III. The Proposed Approach

In this section we explain the five steps for our Extended latent Economic Concepts based predictive method. The first two steps are essentially identical to the bag of concepts method (Kim et al., 2017) for document classification procedure, while to the best of our knowledge this method was not motivated in the past fundamental news analysis. We present our novel idea in last 3 steps. Figure 1 depicts the steps that we follow in our methodology that contain the creation of embedding vector for words, latent semantic space modeling, document vectorization, news document labeling based on change level in the market (up/down) and finally, Support Vector Machine for the classifier as a baseline predictive model. Table 1 presents the notations and symbols that we use throughout the rest of this paper.

Table 1- notations and symbols that we use throughout the paper

Notation	*Description*
V	Vocabulary
V	Vocabulary Size
m	Word embedding dimension
$\Re^m$	M dimensional embedding space
$\vec{w}_i \in \Re^m$	Embedding vector for i'th word in vocabulary
K	Contextual window size
$C^L = \{c_1, \ldots, c_L\}$	Embedded latent concepts space
L	Number of latent concepts in Forex News Corpus or document embedding dimension
$c_i = \{\vec{w}_1 \ldots . \vec{w}_n\}$	$\mathbf{c_i}$ is i'th Economic Latent Concepts in Forex News Corpus
CF	Concept frequency based document representation method
ELCF	Economic latent concept frequency based document representation method
CF-IDF	Concept Frequency Inverse Document Frequency
ELCF-IDF	Extended Latent Concept Frequency Inverse Document Frequency

A. *Word Embedding*

This technique aims at creating a vector representation for each word that reflects syntactical roles and semantics (B. Liu, 2015). In word2vec method, an embedded vector representation for each word is generated using a Skip-gram neural network (Mikolov et al., 2013). The following procedure states steps for the word2vec method. In the skip-Gram network, each contextual neighbor for input word

predicted with equation 1. In the backpropagation phase, the loss computed based on equation 2.

Input

One hot encoding for all words in vocabulary

Method

Skip-gram Neural Net.

$$p(w_{c.j}|w_i) = \frac{\exp(w_{c.j}^T \cdot w_i)}{\sum_i \exp(w_j^T \cdot w_i)} \quad (1)$$

$$l = \frac{1}{2}\sum_j \left(p(w_{c.j}|w_i) - t_j\right)^2. \quad if\ term\ j\ observed\ in\ context\ window\ of\ term\ i \quad (2)$$

Notations:

$j = 1. \ldots . k$. $i = 1. \ldots . v$. k: context window size . v : vocabulary size

$w_{c.j}$: corresponding weight vector between input node and embeding node

w_i : corresponding weight vector between embeding node and output

Output

$\{\overrightarrow{w_1}. \ldots . \overrightarrow{w_v}\}$

$\vec{w}_i \in \Re^m$ Embedding vector for i'th word in vocabulary

In the proposed method, a corpus of Forex market-related news is developed, then preprocessing steps such as removing numbers and stop words and verb lemmatization is done. For word2vec method we used python Gensim package in Spyder environment.

B. *Semantic Latent Concepts Modeling*

After the formation of embedding space for words in the economic corpus, embedding vectors clustered to categorized conceptually related words (Hu et al., 2017). Each cluster corresponds to one latent concept. Accordingly, we encapsulate similar words vector in clusters as latent economic concepts. The following pseudo code describe the clustering method with k-means++ algorithm(Shi, 2005).

Input

$\{\overrightarrow{w_1}. \ldots . \overrightarrow{w_v}\}$. $\overrightarrow{w_t} \in \Re^m$

Method

Dimension Reduction with k-means++ clustering

Output

$Concepts\ clusters : \{\boldsymbol{c_1}, \ldots, \boldsymbol{c_l}\}, \boldsymbol{c_i} = \{\vec{\boldsymbol{w}}_1. \ldots . \vec{\boldsymbol{w}}_n\}$

$\mathbf{c_i}$ is i'th Economic Latent Concepts in Forex News Corpus

C. *Document Representation*

The text representation method is the basic step in the text classification process (Anbaee Farimani, Tabatabaee, & kaffashan kakhki, 2019). Each news document synthesize both the news title and description (Q. Liu, Cheng, Su, & Zhu, 2018). In the proposed method, the Extended latent Concept Frequency (ELCF) algorithm presented aims to benefit all information on both news title and description, besides the expansion of news title for strengthen of news topic. The following pseudo code describe the ECF procedure. In the first step, news documents would be preprocessed. Preprocessing steps include removing numbers and stopwords and verb lemmatizing. The next step is feature selection includes title expansion. Top n most similar words (Cos Similarity) to title's keywords extracted from word embedding space expressed in line 10. In the last step, keywords distribution through latent semantic concepts would be calculated which express in line 12. Therefore the dimension of the document vector is identical to the number of latent concepts in embedding concepts space.

1. ***Input***:
2. d ={title, content, timestamp, label}
3. $C^L = \{c_1, \ldots, c_L\}$, $c_i = \{\vec{w}_1. \ldots . \vec{w}_n\}$
4. $\Re^m = \{\overrightarrow{w_1}. \ldots . \overrightarrow{w_v}\}$
5. ***Method***
6. Extended Latent Concept Frequency (ELCF)
7. *Step1*:
8. Keywords = Preprocess(d)
9. Words = keywords corresponding vector from $\Re^m$
10. ExtWords = Words $\cup_{top\ 5\ similar\ of\ title's\ words}\{\overrightarrow{w_1}. \ldots . \overrightarrow{w_v}\}$
11. *Step2*:
12. $\vec{d}_j[t] = \#\ of\ words\ in\ ExtWords \in \boldsymbol{c_i}$,
13. i = 1, … , L , t = 1, … , L
14. ***Output***
15. $\vec{d}_j \in C^L$, vector of i'th document in Corpus

D. *News Document Labeling*

After news publishment, it may investors affecting the market up or downtrend or not and the level of close price would be neutral. Therefore trend prediction solution is accordant to the classification problem (!!! INVALID CITATION !!! [21, 24]). So far, some researchers have examined the different duration of time for analyzing the

impact of news on the market from twenty minutes to one week. Researchers in studies such as (Majid Vafaei Jahan & Akbarzadeh-T, 2012; M. V. Jahan & Akbarzadeh-Totonchi, 2010; Khadjeh Nassirtoussi et al., 2015; Seifollahi & Shajari, 2019) have argued that investors' reactions to the Forex market within an hour from the news release time can be examined. In this research, news labeling is done based on the equation 3. we perceive changes in the logarithmic return in equation 3, where Δt stands for the returns' time-lag the hourly close price of currency pair. After the news is released. If the change is greater than a threshold, news is assigned uptrend and otherwise if the change is smaller than a threshold, news is assigned downtrend label.

$$r_{\Delta t} = log\,(p(t + \Delta t)) - log\,(p(t - \Delta t)) \qquad (3)$$

If $r_{\Delta t}$ > threshold label = Up

If $r_{\Delta t}$ < threshold label = down

Else label = Neutral

E. *Model Training*

In the market prediction task, Support Vector machine (SVM) as a baseline predictive model is used (Agarwal, Kumar, & Goel, 2019; Q. Li, Chen, Wang, Chen, & Chen, 2018; Shynkevich, McGinnity, Coleman, & Belatreche, 2016). Given that the proposed method presents a document representation technique - which is an underlying task in text mining - it is justified to use this classifier. We exploit the version implemented in Sklearn package in Python and trained our SVM model with 80% of training samples. For estimating the time sensitivity of the model, we investigate two experiments. For training our model all the training samples shuffled to destroy chronological order and then, to address the dynamics of this process over time, we used a "sliding window" approach, in which we used four months' worth of data to train a model and then validated the model's performance on data from the succeeding month.

IV. CASE STUDY

For the case of study, we focus on Forex Market and conduct all the experiments on EUR/USD currency pair to determine the news impact on the volatility of market close price. In this section, the details about the dataset visualization of embedding spaces and parameter tuning of the proposed method are described. For evaluation of the efficacy of proposed method we used accuracy, ROC, F-Ssore, precision and recall measures.

A. *Datasets and statistical features*

The news text documents we use are all from Fxstreet (www.fxstreet.com), a well-known newsgroup which publishes specifically for the Forex market. The agency publishes an average of 120 news on business days. Figure 2 - distribution on news within a day shows the distribution on news within a day. The Forex News dataset containing over 40000 records of 18-month data from August 28, 2018, to February 7, 2020. The news is obtained by a web crawler we develop in python as the formation of Json item we store in unstructured MongoDB collections. Each record contains title, content, timestamp, URL, keywords and the author's name. For the training word embedding model, we use the Forex news dataset. Given we focus on the EUR/USD currency pair, we also explore the subset of news with keyword 'EURUSD' for training our predictive model.

Table 2 shows statistical features for these two datasets. There are no news releases on holidays, and indicators remain unchanged as the market closes. Therefore, for labeling the news item published in the first hour of Monday the price change will be calculated based on the last hour of last Friday. The EUR/USD dataset is slightly imbalanced. Therefore, We used a micro F1 score for 10-Fold Cross-validation in all evaluation details.

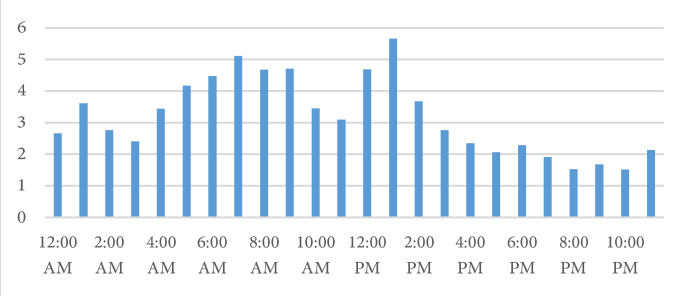

Figure 2 - distribution on news within a day

Table 2 – dataset statistical features

Corpus	*total news document*	*size*	*Keywords count*	*Label statics*

Forex News Corpus	40341	42 MB	22130[1]	-----
EUR/USD News Corpus	3016	4 MB	5146	Uptrend : 1092 Downtrend: 1925

We extracted our market dataset from FXTM (www.forextime.com) for the EUR/USD currency pair. Each record is time-stamped every 5 minutes and contains Ask, Bid, Open, Close price and trading Volume information. We available Crawler source code and the Forex News Dataset in our Github account[2].

A. *Visualization*

To better imagine the proposed financial news analysis, we visualized the embedding space with T-SNE algorithm (Maaten, 2014). Figure 3 - word embedding space for Forex News Datasetshows word Embedding Space. Each point correspond to word vector generated with word2vec method. In this figure, segregated segments can also be seen, represent the same concepts that arise together in those segments. For example, the semantic relationship that emerges from the common occurrence of words such as Open, Close, Low, High, Trend, Point, Pivot, Level in the contextual window of economic news is a price-related financial concept. This cause similar vectorization for these similar conceptually words.

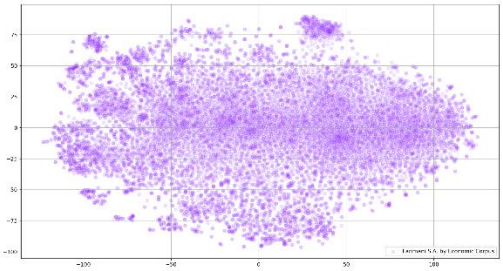

Figure 3 - word embedding space for Forex News Dataset

Figure 4- document embedding space through latent semantic concepts, visualization by T-SNE shows document embedding space through 30 latent semantic concepts. Each point correspond to document vector witch labeled according to the change in log return. The blue dots indicate the uptrend label and the red dot then downtrend label.

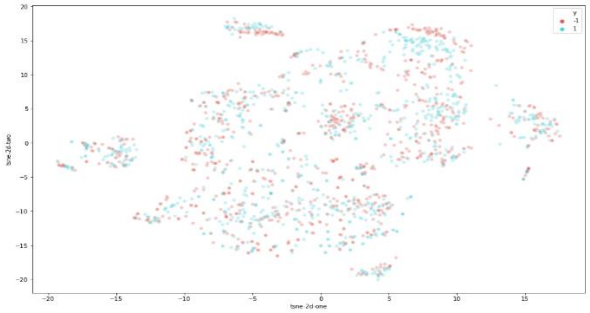

Figure 4- document embedding space through latent semantic concepts, visualization by T-SNE

Figure

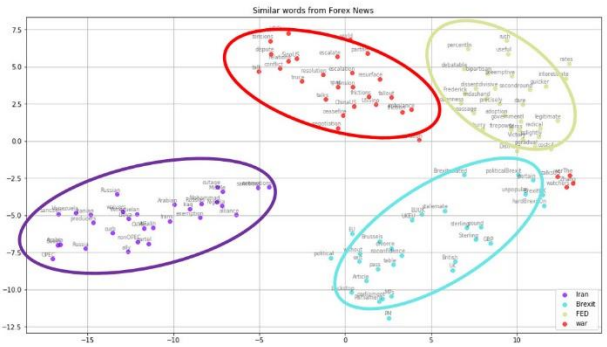

Figure 5 - similar words in embedding space with PCA Scatter Plot shows similar words to 'Iran', 'Brexit', 'Fed' and 'War' in the word embedding space.

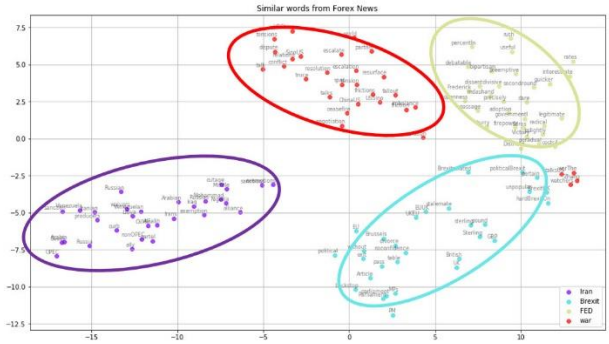

Figure 5 - similar words in embedding space with PCA Scatter Plot

We illustrate the similar words to each of these terms in Table 3 similar words for some economic related term3 . Words like Russia, OPEC, sanction have the most conceptual

[1] *After preprocessing steps, removing number, stop words, less than 3 times of occurance*

[2] https://github.com/anbaee

resemblance to the word Iran while the group of similar words with 'Iran' take far apart from 'Brexit' related terms. These distances infer the existence of independent concepts in the Forex news corpus.

Table 3 similar words for some economic related term 3

#	word	10 Most Similar Words
1	Iran	Iranian, Saudi, sanction, waivers, Arabia, Venezuela, Russias, Iranian, Libya, OilMin
2	Brexit	Political, parliament, Article, table, divorce, exit, UK, British, EUUK, backstop
3	War	Spat, tensions, dispute, frictions, conflict, relations, ChinaUS, SinoUS, escalate, resolution
4	Fed	Powell, FOMC, message, BoC, Riksbank, pause, ECB, Echoing, Clarida,

B. *Simulation Results*

In this section we investigate the effect of some hyper parameters in our proposed approach. There are common hyper parameters for both bag of concept (CF) method and proposed ECLF. Therefore, we test all experiment based on micro F1 score measure for baseline CF method with 10-Fold cross validation. These hyper parameter are word embedding dimension, contextual window size, Vocabulary size in training word2vec, the number of latent concepts in the embedded latent space and the distribution of words through latent concepts. In the word2vec training process, after removing numbers and stop-words, we ignore the words with less than three-times of occurrences. For sensitivity evaluation of these hyper parameters, we test different values by F1 score measure*ment*.

- *Contextual Window size*

In training word2vec we test different value for k stands for conceptual window size. Starting from 1, the value of k increase by 1 until 15 for Forex News Corpus. The best value achieved when k set to be 3. Figure 6 represent micro F1 Score for different value of k.

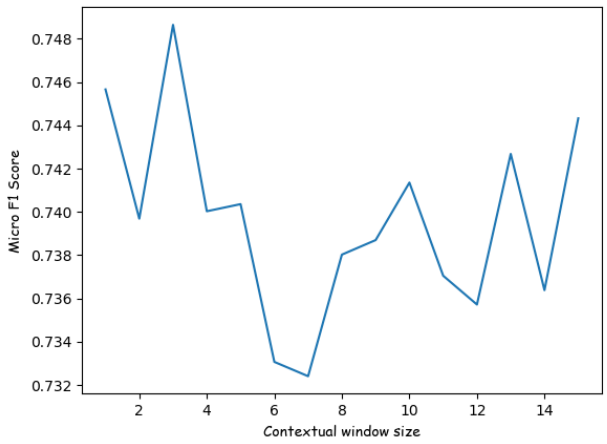

Figure 6 – Micro F1 Score for different contextual window size

- *Word embedding dimension*

For studying about the effect of word embedding dimention, we test different value for m. Starting from 50, the value of m increase by 20 until 300 on Forex News Corpus. Figure 7 depict micro F1 score for different embedding dimensions. The best value achieved when m is set to 210.

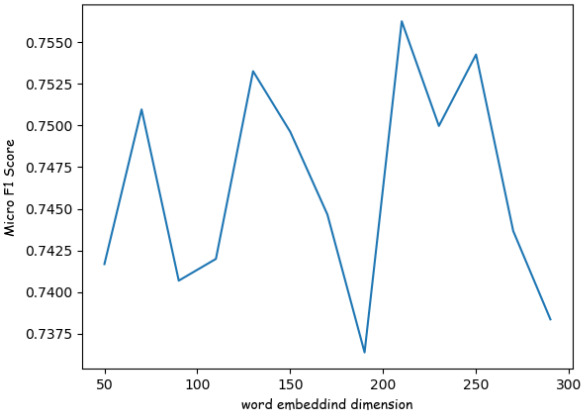

Figure 7- micro F1 score for different embedding dimensions

- *Number of latent concepts clusters*

Studying about the number of latent concepts is an open research domain (Koltcov, 2018). In order to observe its impact on the representation performance, we test several values for the number of concept clusters. The document embedding dimension is identical to the number of concepts

clusters (Kim et al., 2017). Starting from 10, the value of L increase by 10 until 300 on Forex News Corpus. Fig 8 depict micro F1 score for 10-fold cross validation under the different value of concepts cluster. The best result achieved when the concept cluster is identical to 210.

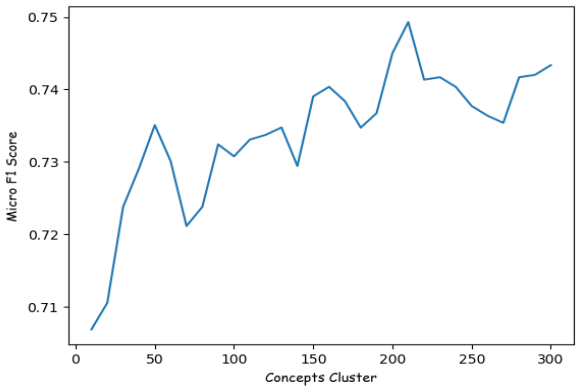

Figure 8 – Micro F1 Score for different concepts clusters number

- *Vocabulary size*

We experimenting the vocabulary size in the process of word embedding training. We test the accuracy of SVM model for both Forex News Corpus and EUE/USD corpus. Table 4 shows the accuracy of cross validation on SVM model for CF Document representation method over 210 concepts clusters. The results shows no significant difference. This may be due to the long interval of news dataset.

Corpus	*total news document*	*Keywords count*	*10-Fold Cross validation accuracy*
Forex News Corpus	40341	22130[3]	0.744994
EUR/USD News Corpus	3016	5146	0.739681195

- *distribution of words through latent concepts*

In figure 9 and 10, we plot words distribution over 20 and 30 concepts. The distribution of words in concepts in not uniform. Often density of words counts in two or three concepts is high. Given that the corpus of financial news is full of the economic-related words, it is a challenging task to distinguish between these words and other non-economic terms. In traditional Bag-of-words often applies term frequency-inverse document frequency (TF-IDF) (Salton, 1989) for removing the effect of unimportant item. In the proposed method for eliminating the effect of these dense concepts, we also use concept weighting based on the frequency of occurrence in documents embedding space.

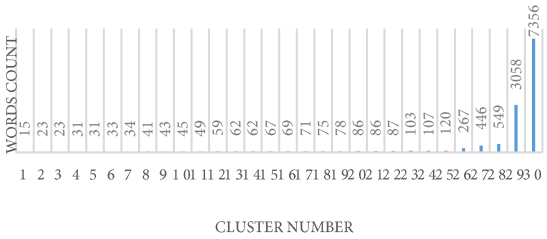

Figure 9- words distribution over latent concepts

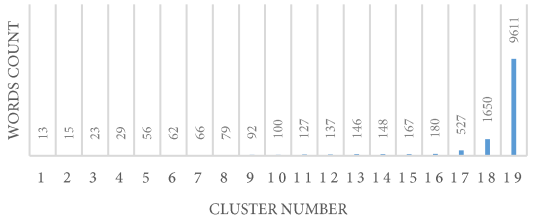

Figure 10- words distribution over 30 number of latent concepts

V. EXPERIMENTAL RESULTS

[3] *After preprocessing steps, removing number, stop words, less than 3 times of occurance*

. (!!! INVALID CITATION !!! [21, 24]).

Agarwal, S., Kumar, S., & Goel, U. (2019). Stock market response to information diffusion through internet sources: A literature review. *International Journal of Information Management, 45*, 118-131. doi:https://doi.org/10.1016/j.ijinfomgt.2018.11.002

Anbaee Farimani, S., Tabatabaee, H., & kaffashan kakhki, M. (2019). An Investigation into the Process of Organizing and Retrieving Web Texts Based on the Integration of Semantic Concepts In order to organize knowledge. *IranDoc, 34*(4), 1879-1904.

Arman Khadjeh Nassirtoussi, T. Y. W., Saeed Reza Aghabozorgi, David Ngo Chek Ling. (2014). Text Mining for Market Prediction: A Systematic Review. *Expert Systems with Applications, 41*(16), 7653-7670.

Birz, G. (2017). Stale economic news, media and the stock market. *Journal of Economic Psychology, 61*, 87-102. doi:https://doi.org/10.1016/j.joep.2017.03.002

Blasch, E., Valin, P., & Bosse, E. (2010). *Measures of effectiveness for high-level fusion*. Paper presented at the In 2010 13th Conference on information fusion

Cutler, D., Poterba, J., & Summers, L. (1988). What moves stock prices? *The Journal of Portfolio Management Spring, 15*(3), 4-12.

Dang, M., & Duong, D. (2016, 14-16 Sept. 2016). *Improvement methods for stock market prediction using financial news articles.* Paper presented at the 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS).

Égert, B., & Kočenda, E. (2014). The impact of macro news and central bank communication on emerging European forex markets. *Economic Systems, 38*(1), 73-88. doi:https://doi.org/10.1016/j.ecosys.2013.01.004

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business,, 38*(1), 34–105.

Gurin, Y., Szymanski, T., & Keane, M. T. (2017, 7-8 Sept. 2017). *Discovering news events that move markets.* Paper presented at the 2017 Intelligent Systems Conference (IntelliSys).

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697. doi:https://doi.org/10.1016/j.dss.2013.02.006

Ho, C.-S., Damien, P., Gu, B., & Konana, P. (2017). The time-varying nature of social media sentiments in modeling stock returns. *Decision Support Systems, 101*, 69-81. doi:https://doi.org/10.1016/j.dss.2017.06.001

Hu, L., Zhang, B., Hou, L., & Li, J. (2017). Adaptive online event detection in news streams. *Knowledge-Based Systems, 138*, 105-112. doi:https://doi.org/10.1016/j.knosys.2017.09.039

Jahan, M. V., & Akbarzadeh-T, M.-R. (2012). Composing local and global behaviors: Higher performance of spin glass based portfolio selection. *Journal of Computational Science, 3*(4), 238-245. doi:https://doi.org/10.1016/j.jocs.2012.04.004

Jahan, M. V., & Akbarzadeh-Totonchi, M. (2010). From Local Search to Global Conclusions: Migrating Spin Glass-Based Distributed Portfolio Selection. *IEEE Transactions on Evolutionary Computation, 14*(4), 591-601. doi:10.1109/TEVC.2009.2034646

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications, 42*(1), 306-324. doi:https://doi.org/10.1016/j.eswa.2014.08.004

Kim, H. K., Kim, H., & Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing, 266*, 336-352. doi:https://doi.org/10.1016/j.neucom.2017.05.046

Kočenda, E., & Moravcová, M. (2018). Intraday effect of news on emerging European forex markets: An event study analysis. *Economic Systems, 42*(4), 597-615. doi:https://doi.org/10.1016/j.ecosys.2018.05.003

Koltcov, S. (2018). Application of Rényi and Tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications, 512*, 1192-1204. doi:https://doi.org/10.1016/j.physa.2018.08.050

Krishnamoorthy, S. (2018). Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems, 56*(2), 373-394. doi:10.1007/s10115-017-1134-1

Li, Q., Chen, Y., Wang, J., Chen, Y., & Chen, H. (2018). Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective. *IEEE Transactions on Knowledge and Data Engineering, 30*(2), 381-399. doi:10.1109/TKDE.2017.2763144

Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., . . . Deng, X. (2016). Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications, 27*(1), 67–78.

Liu, B. (2015). *Opinions, Sentiment, and Emotion in Text*.

Liu, Q., Cheng, X., Su, S., & Zhu, S. (2018). *Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News*. Paper presented at the Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy.

Long, W., Song, L., & Tian, Y. (2019). A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. *Expert Systems with Applications, 118*, 411-424. doi:https://doi.org/10.1016/j.eswa.2018.10.008

Ma, L., & Zhang, Y. (2015, 29 Oct.-1 Nov. 2015). *Using Word2Vec to process big text data*. Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).

Maaten, L. V. D. (2014). Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res., 15*(1), 3221–3245.

Mariana, D., Neves, R. F., & Horta, N. (2017). Company event popularity for financial markets using Twitter and sentiment analysis. *Expert Systems with Applications, 71*, 111-124. doi:https://doi.org/10.1016/j.eswa.2016.11.022

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR, abs/1301.3781*, 1301-3781.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*: Addison-Wesley Longman Publishing Co., Inc.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464. doi:https://doi.org/10.1016/j.dss.2012.03.001

Seifollahi, S., & Shajari, M. (2019). Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *Journal of Intelligent Information Systems, 52*(1), 57-83. doi:10.1007/s10844-018-0504-9

Shi, Z. (2005, 31 July-4 Aug. 2005). *Efficient online spherical k-means clustering.* Paper presented at the Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems*, *85*, 74-83. doi:https://doi.org/10.1016/j.dss.2016.03.001

Tetlock, P. C. (2010). Does Public Financial News Resolve Asymmetric Information? *The Review of Financial Studies*, *23*(9), 3520–3557.

Tetlock, P. C. (2011). All the News That's Fit to Reprint: Do Investors React to Stale Information? *The Review of Financial Studies*, *24*(5), 1481–1512.

Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, *42*(11), 4999-5010. doi:https://doi.org/10.1016/j.eswa.2015.02.007

Verma, I., Dey, L., & Meisheri, H. (2017). *Detecting, quantifying and accessing impact of news events on Indian stock indices*. Paper presented at the Proceedings of the International Conference on Web Intelligence, Leipzig, Germany.

Wang, H., Lu, S., & Zhao, J. (2019). Aggregating multiple types of complex data in stock market prediction: A model-independent framework. *Knowledge-Based Systems*, *164*, 193-204. doi:https://doi.org/10.1016/j.knosys.2018.10.035

Wei, Y.-C., Lu, Y.-C., Chen, J.-N., & Hsu, Y.-J. (2017). Informativeness of the market news sentiment in the Taiwan stock market. *The North American Journal of Economics and Finance*, *39*, 158-181. doi:https://doi.org/10.1016/j.najef.2016.10.004

Wong, C., & Ko, I. (2016, 13-16 Oct. 2016). *Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices.* Paper presented at the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI).

Zhang, W., Li, Y., & Wang, S. (2019). Learning document representation via topic-enhanced LSTM model. *Knowledge-Based Systems*, *174*, 194-204. doi:https://doi.org/10.1016/j.knosys.2019.03.007