Sentimental Analysis on Turkish Blogs via Ensemble Classifier

Sadi Evren SEKER
Dept. of Business Administration
Istanbul Medeniyet University
academic@sadievrenseker.com

Khaled Al-NAAMI
Computer Science Department
The University of Texas at Dallas
kma041000@utdallas.edu

ABSTRACT
Sentimental analysis on web-mined data has an increasing impact on most of the studies. Sentimental influence of any content on the web is one of the most curios questions by the content creators and publishers. In this study, we have researched the impact of the comments collected from five different web sites in Turkish with more than 2 million comments in total. The web sites are from newspapers; movie reviews, e-marketing web site and a literature web site. We mix all the comments into a single file. The comments also have a like or dislike number, which we use as ground proof of the impact of the comment, as the sentimental of the comment. We try to correlate the text of comment and the like / dislike grade of the proof. We use three classifiers as support vector machine, k-nearest neighborhood and C4.5 decision tree classifier. On top of them, we add an ensemble classifier based on the majority voting. For the feature extraction from the text, we use the term frequency – inverse document frequency approach and limit the top most features depending on their information gain. The result of study shows that there are about 56% correlation between the blogs and comments and their like / dislike score depending on our classification model.

Keywords
Data Mining, Sentimental Analysis, Big Data, Text Mining

1. INTRODUCTION
The data set on this study is collected from internet for one of the high-circulating newspapers, a movie review web page with highest comments, an e-marketing web site with highest comments and a literature web site holding poems and novels all in Turkish. The properties of the dataset will be explained in the experiments section. We have processed the comments with text mining approach called term frequency - inverse document frequency (TF-IDF), which will be explained in the methodology section. On the other hand, we have accepted the number of like or dislike as the ground proof of the impact of the comment. Finally we have investigated the correlation between the features extracted from text mining and signal processing to compare the effect of signal processing outputs into the economy news. During this correlation study, we have implemented k-nearest neighborhood (KNN), C4.5 decision tree (C4.5) and support vector machine (SVM) algorithms, which are discussed in the section of classification. Moreover we have implemented an ensemble classifier over those three classifiers, which is based on majority voting (MaVL), which will also be explained in the background section. Finally, this paper holds the implementation details and the methodology of evaluation over classification results, which are held in the evaluation section.

2. PROBLEM STATEMENT
This study is the first time to address the correlation effect of the comment text and like / dislike count of comments for Turkish data sources.

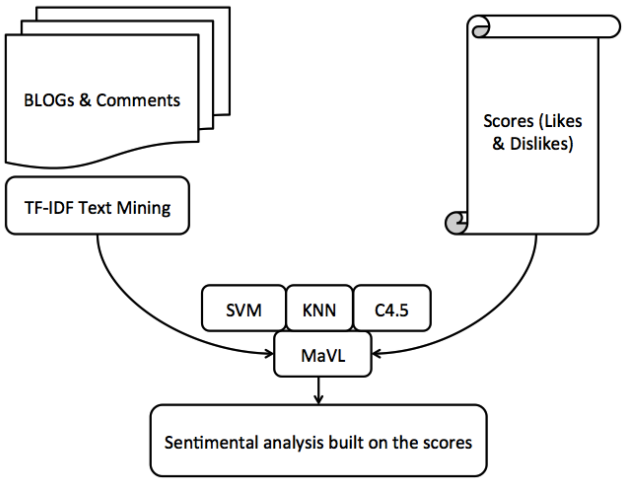

Figure 1. Overview of Study

One of the difficulties in this study is dealing with natural language data source, which requires a feature extraction. The other difficulty is dealing with large number of comments,

which can be accepted as big data problem. The dataset holds 131,248 distinct words and when the feature vector of each economy news item is collected, the total size of the feature vector is over 32.5 GByte, which is beyond the computation capacity of a single computer with these classification algorithms. For a simple SVM implementation the required RAM is slightly more than 1TB.

3. RELATED WORK

Current studies on sentimental analysis on web-mined data has a great impact for the both content authors and publishers. For example the impact of a politician's speech can now be monitored real-time by the help of current studies[1]. For example, the researchs on Arabic Spring and the effect of social media on the Tunisian case [1] or French Presidental Election and social media research [2] or Iran Green Movement from the twitter data [3] or research on UK 2010 election and effect of social media [4] are only a few researches on the topic.

In most of the researches, the data is collected from the social media like Twitter [1-4] or Facebook [5] or e-learning environments mixed with social networks[6]. All of these studies have a text mining part. Zhai[7] shows that the studies based on TF-IDF has a higher success than suffix tress or n-gram based approaches for Chinese case with the SVM classifier.

Some of the reserachers prefers using the metrics built on the social network itself. For example in Twitter, it is possible to get the number of followers and following and such information may be useful to calculate the political views of people depending on who they follow as in UK Election research [4] where the feature extraction is built on the followers/following. Or on some other researches, text mining approaches like bag of words, interjection of emotics, part of speech tagging methods are implemented together [6].

4. BACKGROUND

We have implemented TF-IDF and classification methods as already explained in the introduction; this section will discuss these methods in detail. Also one of the difficulties is the number of words we are dealing with. We have implemented the information gain calculation for eliminating some of the features. Finally the evaluation and error calculation methods will be explained in detail.

4.1. Term Frequency – Inverse Document Frequency

TF-IDF is one of the text mining methods used for feature extraction from natural language data sources[7,8,9].

For the TF-IDF calculation is given in equation (1).

$$tfidf(t,d,D) = tf(t,d)x\ idf(t,D) \quad (1)$$

Where t is the selected term, d is the selected document and D is all documents in the corpus. Also TF-IDF calculation in above formula is built over term frequency (TF) and inverse document frequency (IDF), which can be rewritten as in equation (2).

$$tf(t,d) = \frac{f(t,d)}{\max\{f(w,d): w \in d\}} \quad (2)$$

where f is the frequency function and w is the word with maximum occurrence. Also the formulation of IDF is given in equation (3).

$$id(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

where $|D|$ indicates the cardinality of D, which is the total number of documents in the corpus.

4.2. Information Gain

The information gain of all the terms is calculated and ordered in descending order. Let *Attr* be the set of all attributes and E_x be the set of all training examples, *value(x,a)* with x $\in E_x$ defines the value of a specific example *x* or attribute $a \in Attr$, H , specifies the entropy. The information gain for an attribute $a \in Attr$ is defined as in equation (4).

$$IG(Ex,a) = H(Ex) - \sum_{v \in v(a)} \frac{|x \in Ex | v(x,a)|}{|Ex|} H\big(x \in Ex \big| v(x,a)\big) \quad (4)$$

Also entropy in the information gain calculation can be rewritten as in equation (5).

$$H(X) = \sum_{i=1}^{n} P(x_i)I(x_i) = \sum_{i=1}^{n} P(x_i)\log_b\left(\frac{1}{P(x_i)}\right) = \sum_{i=1}^{n} P(x_i)\log_b\big(P(x_i)\big) \quad (5)$$

4.3. K- Nearest Neighborhood (KNN)

The k, c-neighborhood (or k, c(x) in short) of an U-outlier *x* is the set of *k* class *c* instances that are nearest to *x* (k-nearest class c neighbors of x).

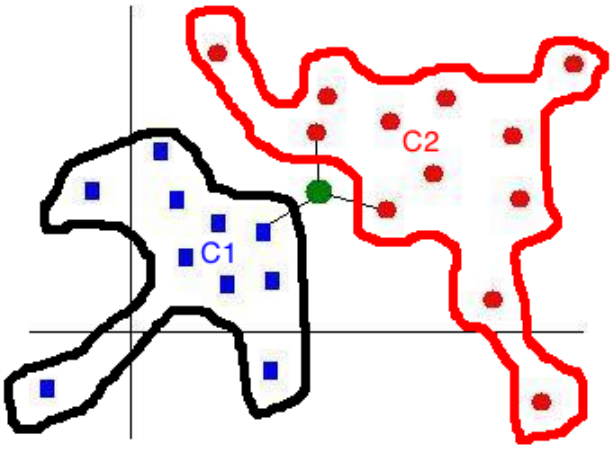

Figure 2. Visualization of K-NN

The K-NN [10] is explained in Figure 2. Here *k* is a user defined parameter. For example, *k* , $c_1(x)$ of an U-outliers *x* is the k-nearest class c_1 neighbors of *x*.

Let $\overline{D}_{Cout,q}(x)$ be the mean distance of a U-outlier x to its k-nearest U-outlier neighbors. Also, let $\overline{D}_{C,q}(x)$ be the mean distance from x to its $k,c(x)$, and let $\overline{D}_{Cmin,q}(x)$ be the minimum among all $\overline{D}_{C,q}(x)$, c ∈ {Set of existing classes}. In order words, k , c_{min} is the nearest existing class neighborhood of x. Then k-NSC of x is given in equation (6).

$$k-NSC(x)=\frac{\overline{D}_{C\min,q}(x)-\overline{D}_{Cout,q}(x)}{\max(\overline{D}_{C\min,q}(x),\overline{D}_{Cout,q}(x))} \qquad (6)$$

4.4. Support Vector Machine (SVM)

The reason of applying SVM method as in Figure 3 over the dataset is determining the boundaries between classes [11.

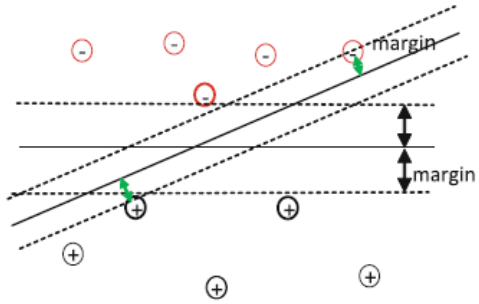

Figure 3. SVM boundary and margins

SVM aims to classify the samples into groups and define a boundary between the groups. SVM also tries to find out the maximum margin possibility between the groups [11].

$$w^{*}=\sum_{i=1}^{n} a_i y_i x_i \qquad (7)$$

The margin between the classes is symbolized by ω symbol in equation (7) and SVM seeks to maximize the value of ω. The above formula can be rewritten as below for the linearly separable classes [12].

$$\|\omega\|^2=\sum_{i=1}^{l}\alpha_i=\sum_{iSVs}\alpha_i=\sum_{iSVs}\sum_{iSVs}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle \qquad (8)$$

In the equation (8), all the possible cases of i and j are considered. Also SVM can use a radial basis function and one of the options is the Gaussian kernel function, quoted in equation (9) [12].

$$K(x,x')=\exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \qquad (9)$$

Finally, the class is determined by the result achieved from K function.

4.5. C4.5 Tree

C4.5 method [13] is a decision tree based classification algorithm. The tree is built by using the information gain of each feature in the feature vector.

The algorithm starts with a training data set S where S= { s_1, s_2 , … s_n } where each sample s_i has a p dimensional feature vector, FV.

For each sample s_i, FV = { x_{1i}, x_{2i}, … x_{pi} } and the information gain of each values would be IG = { $ig(x_{1i})$, $ig(x_{2i})$, … $ig(x_{pi})$ }.

The algorithm creates a decision tree where each node defines a decision to either side.

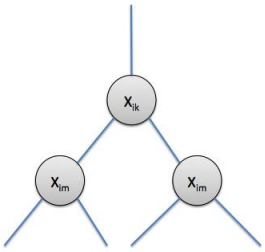

Figure 4. C4.5 Tree Demonstration

The highest information gain value is selected for the top most decision node and the second is get the decision criteria on the next level. Let $ig(x_{ik}) > ig(x_{im})$ for the

Figure **4**. The tree is constructed by following the similar approach for the next levels. Finally at the leaves, the samples are placed after the training.

In the time of testing, the features extracted from test samples are questioned via the decision nodes in the tree from root to leaves. The final leaf is accepted as the class of the test sample.

C4.5 has an advantage on other decision trees, since it uses the information gain and normalization and also it uses the prunning for the time performance.

4.6. Ensemble Classification

We have implmented a majority vote learning (MaVL or Marvel) [16] based ensemble method to combine three different classification methods. MV can be considered as a meta classifier which works over the classifiers like KNN, C4.5 or SVM in our case.

Let $S_i \in S$ where S is the set of classifiers and let $C_i \in C$ where C is the set of classes,

$$C(x) = argmax_i \sum_{j=1}^{B} w_j I(S_j(x)=i) \qquad (10)$$

Where w_j is the weight of each indicator function I(•) which is added into the equation for normalization and the weights of each classifier is equal in our model.

Marvel, gets the summation for each of the classifier's vote and the sample is classified into the class with the highest vote.

4.7. Error Rate Calculation

The error rate of the system is calculated through root mean square error (RMSE). The calculation of RMSE is given in equation (11) [14].

$$x_{rmse} = \frac{\sqrt{x_1^{\,2} + x_2^{\,2} + \ldots + x_n^{\,2}}}{n} \qquad (11)$$

For this study, above x values are the results achieved from the implementation of the algorithm. The RMSE result of 0 is considered ideal and lower values close to 0 are relatively better.

By the results fetched from the output layer and the calculation of RMSE, the algorithm back propagates to the weight values of the synapses.

Also the results are interpreted by using a second error calculation method RRSE (Root Relative Squared Error) and the calculation is given in equation (12) [15].

$$x_{rrse} = \sqrt{\frac{\sum_{j=1}^{n}(P_{ij} - T_j)}{\sum_{j=1}^{n}(T_j - \overline{T})^2}} \qquad (12)$$

Where P_{ij} is the value predicted for the sample case j, T_j is the target value for sample case $_j$ and $\overline{T}$ is calculated by equation (13) [17].

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_J \qquad (13)$$

The RRSE value ranges from 0 to ∞, with 0 corresponding to ideal.

The third error calculation method is RAE (Relative Absolute Error) and the calculation is given in equation (14) [15].

$$E_i = \frac{\sum_{j=1} | P_{(ij)} - T_j |}{\sum_{j=1}^{n} | T_j - \overline{T} |} \qquad 1)$$

$P_{(ij)}$ is the value predicted by the individual program i for sample case j (out of n sample cases), T_j is the target value for sample case j, and $\overline{T}$ is given by the equation (15) [15]:

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \qquad (15)$$

For a perfect fit, the numerator is equal to 0 and E_i=0 So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal.

Also the success rate of prediction and expectation can be measured as the f-measure method. The f-measure method is built on the Table 1.

Table 1. f-measure method

	Predictions		
Expectations		Positive	Negative
	True	True Positive	True Negative
	False	False Positive	False Negative

The calculation of f-measure can be given as in equation (16) depending on the Table 1.

$$F_{measure} = \frac{2TP}{2TP + FN + FP} \qquad (16)$$

5. EXPERIMENTS

In this study the dataset is in natural language and some preprocessing for the feature extraction from the data source is required. The first approach is applying the TF-IDF for the all terms in the data source. Unfortunately the hardware in the study environment was not qualifying the requirements for the feature extraction of all the terms in data source which is 139,434.

5.1. Dataset

We have implemented our approach and Table 2 demonstrates the features of the datasets.

Table 2. Properties of the Dataset

	News
# of News	9871
Authors	6881
Texts per Author	Mean (**μ**) : 44.05 Stddev(**σ**) : 535.52
Average word length	~6.7

The above dataset is collected from the web site of a high-circulating newspaper in Turkey. The data is collected directly from a database so the noisy parts on the web page like ads, comments, links to other news, etc. are avoided. Another problem is the noise of HTML tags in the database entries for formatting the text of news. The data has preprocessed and all the HTML tags are removed from the news and also all punctuations and stop words are removed in the preprocessing phase.

5.2. Feature Extraction

We have implemented a feature extraction algorithm 1 in order to extract two feature vectors.

Algorithm: Feature Extraction Methods

1. *Let E be Economy News Corpus,*
2. *Let C be Closings of Stockmarket,*
3. *For each* $E_i \in E$
4. *For each* $Term_j \in E_i$
5. *if(count(*$Term_j$*)>30)*
6. $T_j \leftarrow$ *TF-IDF of* $Term_j$
7. $C_i \leftarrow$ *closing_value(date(*E_i*))* $\in C$
8. $IG_{ij} \leftarrow$ *Information Gain (*$Term_j$*,*E_i*)*
9. $V_1 \leftarrow$*Top300(sort(IG))*
10. $V_2 \leftarrow C$

The above algorithm demonstrates the extraction of two vectors: one from the economy news corpus and another from the closing values of the stock market. We have limited the number of features to 300

and the Top300 function gets the topmost 300 features from the feature vector.

The V_2 feature vector is calculated easily by checking the closing value of the economy news on the date. There are some news items which are published during the time the stock market is closed like on weekends and we have considered these values as a third class besides the increase and decrease classes.

The correlation algorithms run over the two vectors V_1 and V_2 extracted via the Algorithm 1.

During the execution of algorithm, the execution requires more memory than the available hardware, where we run the algorithms on a intel 7 cpu and 8GByte of RAM. The required memory is calculated in equation (17).

Memory Requirement = 139,434 words x 9871 news x 6.7 average word length x 2 bytes for each character =~17GByte (17)

As a solution we have limited the number of words with the highest occurences. The number of occurences on our implementation is 30 and a word is taken into consideration after this number of occurences. The words appearing above this threshold value are 2878 and the memory required is reduced to 700MByte which is easier to handle in the RAM.

The feature vector extraction is about 56 minutes on average for the economy news.

5.3. Evaluation

The results of executions can be summarized in Table 3.

Table 3. Error and Success Rates of Classification Methods

	f-measure Average	RMSE	RAE	Correctly Classified
Random Walk	0.497	0.4182	0.9921	52.37%
RSI	0.501	0.4404	0.9930	50.70%
MACD	0.491	0.4174	0.9892	52.52%
Bollinger Band	**0.504**	**0.4141**	**0.9810**	**53.49%**

The success rate in Table 3 is the percentage of correctly classified instances. For example, the success rate of Random Walk with length=2 can be considered as the 37% of the instances are correctly classified to predict an increase, decrease or no change in the stock market value depending on the economy news processed.

The time series analysis method, “acceleration” should not be considered because of its unsuitable data output. The acceleration values calculated are either 0 or so close to 0, so the data set expectation was not realistic. This is the reason of high success rate on the acceleartion analysis. On the other hand rest 9 methods are suitable for the correlation and the highest success is achieved from the Bollinger Band with 52% correctly classified news. The success rate achieved in this study is much better than the previous studies[15].

The value of success is highly related with the market structure so the success rate here should not be understood as the success rate of the methodology or the classifier. The success rate in the table is the correlation between economy news and the stock market closing values.

6. CONCLUSION

During this study, it is first time the effect of time series anlaysis methods over the stock market closing values and their correlation with the economy news in the Turkey case has been studied. The feature extraction method and classification methods are kept simple and the study is mainly focused on the time series analysis. The analysis has shown that the success of Bollinger band is higher than the rest.

We believe this study would help to understand the market strength in Turkey from a financial perspective and also the study can help further research with other classification algorithms and feature extraction methodologies.

7. REFERENCES

[1] Younus, A.; Qureshi, M.A.; Asar, F.F.; Azam, M.; Saeed, M.; Touheed, N., "What Do the Average Twitterers Say: A Twitter Model for Public Opinion Analysis in the Face of Major Political Events," Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on , vol., no., pp.618,623, 25-27 July 2011
doi: 10.1109/ASONAM.2011.85 [2] Braun, H. 1987. Predicting stock market behavior through rule induction: an application of the learning-fromexample approach. Decision Sciences, vol. 18, no. 3, pp. 415-429.

[2] Wegrzyn-Wolska, K.; Bougueroua, L., "Tweets mining for French Presidential Election," Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on , vol., no., pp.138,143, 21-23 Nov. 2012
doi: 10.1109/CASoN.2012.6412392

[3] Khonsari, K.K.; Nayeri, Z.A.; Fathalian, A.; Fathalian, L., "Social Network Analysis of Iran's Green Movement Opposition Groups Using Twitter," Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on , vol., no., pp.414,415, 9-11 Aug. 2010
doi: 10.1109/ASONAM.2010.75

[4] Boutet, A.; Hyoungshick Kim; Yoneki, E., "What's in Twitter: I Know What Parties are Popular and Who You are Supporting Now!," Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on , vol., no., pp.132,139, 26-29 Aug. 2012
doi: 10.1109/ASONAM.2012.32

[5] Neri, F.; Aliprandi, C.; Capeci, F.; Cuadros, M.; By, T., "Sentiment Analysis on Social Media," Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on , vol., no., pp.919,926, 26-29 Aug. 2012
doi: 10.1109/ASONAM.2012.164

[6] Martin, J.M.; Ortigosa, A.; Carro, R.M., "SentBuk: Sentiment analysis for e-learning environments," Computers in Education (SIIE), 2012 International Symposium on , vol., no., pp.1,6, 29-31 Oct. 2012

[7] Zhongwu Zhai; Hua Xu; Jun Li; Peifa Jia, "Sentiment classification for Chinese reviews based on key substring features," Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on , vol., no., pp.1,8, 24-27 Sept. 2009

doi: 10.1109/NLPKE.2009.5313782

[8] Zhai, Y., Hsu, A., and Halgamuge, S. 2007.Combining News and Technical Indicators in Daily Stock Price Trends Prediction. Lecture Notes in Computer Science. 1087-1096.

[9] Fung, G., Yu, J., and Lam, W.2002. News sensitive stock trend prediction. Lecture Notes in Computer Science, vol. Volume 233, 481– 493.

[10] Masud, M. M., Al-Khateeb, T. M., Khan, L., Aggarwal, C. C., Gao, J., Han, J., and Thuraisingham, B. M.2011. Detecting recurring and novel classes in concept-drifting data streams. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*.(Vancouver, Canada, December 11-14, 2011)IEEE Computer Society Washington, DC, USA , 1176–1181. DOI= http://dx.doi.org/10.1109/ICDM.2011.49

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterlinga and B. P. Flannery , Numerical recipes: the art of scientific computing, Cambridge University Press, New York, 2007.

[12] S. R. Gunn, Support vector machines for classification and regression, University of Southamtpon, Technical Report, 1998.

[13] Yahia, M.E. and Ibrahim, B. A. 2003. K-nearest neighbor and C4.5 algorithms as data mining methods: advantages and difficulties. In Proceedings of Computer Systems and Applications, 2003. Book of Abstracts. ACS/IEEE International Conference on. (Tunis, Tunisia, July 14-18, 2003)

[14] K. V. Cartwright, Determining the effective or RMS voltage of various waveforms without calculus, Ph.D. Thesis, School of Sciences and Technology College of the Bahamas, Bahamas, 2007.

[15] Seker, S. E.;Ozalp N. ; Al-Naami, K. ; Mert C. ; Khan, L. , "Correlation Between Turkish Stock Market and EconomyNews", *Reliability Aware* Data Fusion, held along with SIAM International Conference on Data Mining 2013 (*SDM* 2013), May 2013, Austin, TX, USA