ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 8, Issue 7, January 2019

Sentiment Analysis using Chi^2 Algorithm in Dari Language

Mohammad Hussain Sultani, Amir Hossein Javan Amoli, Nazifa Kazimi

Abstract— Opinion mining or sentiment analysis tries to express different people feelings and emotions toward different objects or entities. Opinion mining is used to classify sentiments of text. If the sentence is positive, it is considered positive and if that is negative, it will be classified as negative sentence. Today, much more attention has been paid to opinion mining domain. One of its main reasons is the enormous applications of opinion mining in different fields. Up to now there is more than 1500 scientific researches in this area. For sentiment analysis application we can refer to portfolio, and elections prediction. In this paper chi^2 is used for feature selection and Bayes Naïve is used in model creation. Evaluation results show that the generated model has higher accuracy than previous works results.

Keywords: **Sentiment Analysis, Persian Opinion Mining, Machine Learning, Data Mining.**

I. INTRODUCTION

By emerging Web 2.0 users can easily express their opinions in related to different entities and publish their agreement and disagreement for social media users. Nowadays, it is necessary for each one to know what other think about products which they want to buy. As there are large amount of comments in social media, it is not possible to analyze them manually. Opinion mining is trying to find a method for classifying these opinions. Opinion mining or sentiment analysis is a study field that attempts to express feelings, behaviors, opinions, and analysis of different individuals related to different entities such as product, services, organizations, individuals, events or topics.

Sentiment analysis comes with different names such as sentiment analysis, opinion analysis, opinion mining, opinion extraction and sentiment mining. Moreover, when you study in this domain, it is possible that you may see many other titles which are used in opinion mining.

The aim of opinion mining is analyzing user’s sentiments. These sentiments can be expressed in text, voice or video. The main work of opinion mining is done on texts; it can be considered a part of Natural Language Processing.

II. DATASET EXTRACTION

Hotel dataset has been extracted from comments of hellokish.com. This dataset includes 8499 records in which 6126 of them are positive and 2373 of them are negative. In this dataset, we have 5 fields named: comment, c, rate, name and date. C field shows the class of comment which is pos for positive comments and neg for negative comments. Missing value deletion and balancing are also done on dataset. The Beautiful Soup library is used for data extraction from HTML and XML web pages.

Manuscript received: 21 December 2018
Manuscript received in revised form: 17 January 2019
Manuscript accepted: 03 February 2019
Manuscript Available online: 10 February 2019

Afterwards, we select 2000 samples from each class and we will form a new dataset with 4000 records. Hazm library is used for normalization. In addition to deleting words with wrong syntax, these words can be detected by automatic technique. These words will have low term frequency and will not be selected as features. Stop words are deleted due to degrading efficiency, precision and accuracy.

III. FEATURE EXTRACTION

In order to modeling on unstructured text, firstly the significance and eligible features should be extracted. Many methods of feature extraction will be discussed later. In all methods, first step is the separation of meaningful units from the text. These meaningful units are called tokens and this process is called tokenization. Token can one word or can be a phrase of word. In the following, we will propose common corpus display methods as matrix of numbers.

A. Binary classification

In order to using numerical methods, we need to have numeric features. Bag of Word is one of the easiest ways for displaying numeral corpus or in brevity BOW. Following the presence and absence (Binary) of words and thereafter their term-frequency will be shown by BOW. An entity can be represented as a matrix with the columns of that piece and its rows of documents. In this method the presence of the token in the considered document is shown as 1 and the 0 value shows its absence.

We used scikit-learn for establishing Bag of Words which is an open text library for machine learning and many classification, clustering and regression algorithms can be implemented there.

B. Term Frequency

In Term frequency, the frequency of a term in a document is displayed instead of just displaying the presence and absence of the term by 0 and 1. This method is named term-frequency. In spite of its simplicity, its drawback comes with high word dimension. Suppose that if a corpus consists of 100,000 words per document, there will be a vector equivalent to 100,000.These large dimensions require high storage space, as well as complex algorithms and longer execution time. Not considering the word order is another drawback of this method [4].

C. Ngrams

In some cases, it is better to use bigram, trigram instead of unigram for creating BOW.

The use of n grams usually increases the accuracy of the models, but why does ngram have such an effect?

D. TF-IDF

Giving the highest importance to most frequent words is the main drawback of TF method. While in modeling these

DOI:10.17605/OSF.IO/UX4ED

features may have low importance. The aim of tf-idf is to give value for words which can more differentiate text groups. IDF is used to reduce the importance of words which are more repeated throughout the entire text. The aim of combining tf and idf is taking their advantages along with eliminating their disadvantages. The equation 1 is used to calculate the tfidf of t term [11].

(1) $\text{tf-idf}_{t,d} = \text{tf} * \text{idf}$ $\quad \text{idf}(t,D) = \log\left(\frac{N}{|\{d\in D: t\in d\}|}\right)$

- N is the total number of documents
- D is the number of whole document
- d is the number of document is containing term t

For calculating TF-IDF of t word in d document, the t word frequency in d document should be multiplied by t word IDF (which is calculated in all documents). The value of TF-IDF is high for a term, when each term appears in small number of documents but mostly appears in respected document. However it has the lowest value, when it is frequently appeared in all documents. Because in this case, it will not have high differentiation potential for use. It's important to note that TF, TF -IDF and ngram are actually methods for extracting features that are used to formulate document-term matrices.

IV. FEATURE SELECTION

Feature selection is the most important and effective section in data mining. At this point, irrelevant or noise features are eliminated. The features which increase the model's error rate are called noise features.

For example student's id and height are irrelevant features concerning to calculating probability of student's immigration to foreign countries. Studies have shown, using the entire features do not produce high precision [1]. There may not be all informative features, therefore; in reality higher numbers of features do not produce better accuracy. Selecting a collection of words as best features for training model is called feature selection. Increasing model performance and deleting noise are two main goals which can be achieved via feature selection. Suppose that an irrelevant word has been repeated in many records of a class. More repetition of this word cause the algorithm consider it as an important feature and generates a model that gives high precision on testing data which contains this word. However, in overall testing data will have low precision? (We call this over fitting of data) There are three generic methods for feature selection.

A. Filter

In this method, feature selection is performed prior to training classifiers. Irrelevant features are removed before giving them to classifiers. And selected features will be given to classifier for generating model. Here classifier only creates the model and does not play any role in feature selection. The features in Filter method are more general than the features in the Wrapper. It denotes that the features obtained in the Filter method can be given to various classification algorithms. However in Wrapper method, the features can be given to specific algorithm by which the features have been extracted. Low calculation time is another advantage of this method. This method's techniques:

1. Information Gain

Information Gain is another method for filtering features which is widely used in machine learning algorithm. The purpose of this technique is to eliminate features which are less likely to be useful. Here equation 2 is used to calculate the benchmark [2].

(2) $$IG(t_k, c_i) = \sum_{c\in\{c_i,\overline{c_i}\}} \sum_{t\in\{t_i,\overline{t_i}\}} P(t,c) \times \log \frac{P(t,c)}{P(t)\times P(c)}$$

P(c) expresses the probability of each class. For two class containing negative and positive, P(c) will be ½. P (t,c) studies the probability of t feature in c class. P (t) expresses the probability of the presence of the t word in the entire document.

2. Chi-Square

Another method of filtering inappropriate features is the CHI2 method, which is considered as one of the most reliable and famous method in statistics. In statistics CHI2 is used for dependency evaluation of two variables. In text mining CHI2 is used to measure the dependency of tk concerning to c_i class. CHI is calculated by equation 3 [2].

(3)

$$CHI(t_k, c_i) = \frac{N \times (a_{ki} d_{ki} - b_{ki} c_{ki})^2}{(a_{ki}+b_{ki}) \times (a_{ki}+c_{ki}) \times (b_{ki}+d_{ki}) \times (c_{ki}+d_{ki})}$$

1. N shows the number of documents or comments.
2. a_{ki} shows the number of t_k feature within c_i class.
3. b_{ki} shows the number of t_k feature within other classes without c_i.
4. C_{ki} is the classes which do not contain t_k feature.
5. D_{ki} is the amount of time when there is not t_k feature and c_i class [2].

This method is one of the most successful statistical and widely used methods in data mining. Indeed, the evaluation is done concerning to how a word and a document are related with each other.

3. Mutual Information

It is a concept related to information theory which measures the dependency relationship between random variables.

This concept can be used to measure the essence of features. In feature selection the equation 4 is used for measuring dependency between t_k feature and c_i class [2]. The higher calculated value of this relation shows that the t_k feature is more informative for c_i class.

(4) $MI(t_k, c_i) = \log \frac{P(t_k, c_i)}{P(t_k)\times P(c_i)}$

P (t_k, c_i) is the probability of t_k feature in c_i class [2].
P (t_k) is the probability of t_k term in document.
P (c_i) is the probability of c_i class.

4. Frequency Based

In this method features with the highest frequency in a class are selected. The attributes which are repeated in all classes, are eliminated. Since they do not carry information about a specific class. Frequency can be expressed in two ways:

1. Document Frequency: The number of documents in a class in which the word is present
2. Collection Frequency: the number of tokens which are in documents of specific class

Document frequency is appropriate for Bernoulli models and collection frequency is fitting to polynomial models [3]. Methods concerning to Frequency can be used as an Unsupervised [1]. Means that we can calculate frequency without class dependency.

V. PREVIOUS WORK

According to Bing Liu's claim [2], about 1500 articles have been published in this area. There have been fewer considerations on Persian language rather than other languages such as English and Chines. Here we tried to summarize the articles published in Persian opinion mining. Persian opinion mining is at the beginning and limited endeavors have been made in this area. In addition, there are varieties of tools available for English language, such as lexicons which is not available in Persian language and for using these tools; researchers are bound up to translate them to Persian. The translation itself contains many mistakes which reduces the power of rendered methods. As well as Persian language has many problems and challenges for various reasons such as the difference in written and spoken language and the structure of the language, which makes Persian language researchers to struggle more powerful ways to achieve powerful methods.

Mr. Basiri and his friends [3] in 2014 seeked to provide a framework for Persian opinion mining using unsupervised methods. Writers presented two new features for their research.

1. Offering unsupervised approach lexicon-based for Persian opinion mining.
2. Offering two dataset for Persian opinion mining.

This research is a sentence level research. Initially the preprocessing step is done. Normalization and stemming are done in prepossessing, afterward lexical errors of text is corrected by Persian correction approaches. After preprocessing the classification of the sentences is completed. SentiStrength dictionary, which is a very popular dictionary in English, is used for extracting positive and negative words. In sentiStrenght each word is numbered between 1 to 5 according to sentiment strength. In overall, the sentiment of a sentence by summing up each words numbers. The supplied dataset contains two datasets of goods. The first dataset consists of 1100 comments and the second dataset contains 263 comments on the goods. The result is compared with the result obtained by the NaiveBayes algorithms, decision tree, and SVM and the reported accuracy obtained by the proposed method is 10% higher than the method mentioned in [7] and on the same set of data.

Mohammad Reza Shams and his friends [5] in 2012, in the sense of the first work of the Persian opinion mining, presented in an English article. In this article, the subjective lexicon contains 8027 words, the dictionary is originally in English and authors have translated in to Persian. The authors have considered the words as inputs and using the subjective dictionary to separate them as positive and negative words and send them to the LDASA algorithm. The output of the algorithm determines whether the comment is positive or negative. Dataset used 400 comments Positive and negative for 3 groups of goods including Mobile, digital camera and hotel. A total of 1200 comments have been reviewed. LDASA algorithm is used for evaluating model performance. The accuracy is improved between 7% and 15%. The best result is 78% which is achieved on hotel group.

Madem Ali Mardani and Mr Aguaee [6] in 1394 proposed hybrid approach for Persian opinion mining. In this research, initially the pre-processing step is done. Subsequently the words, which are so frequent in text, are separated. After translating separated words to English, the SentiWordNet dictionary is used to get the polarity of each word. Polarized words are fitted as features to SVM algorithm to create model for opinions classification. Senti Word Net lexicon is one the most effective lexicon in the field of opinion mining. The dictionary provides three numbers for each word. These numbers represent the polarity degree of positive or negative sense for each word. The dataset consists of 1566 hotel reviews. In evaluation of proposed method, the results are compared with results achieved by SVM and NaiveBayes algorithms. The research is implemented by Vika software. The best reported accuracy is 83.57%.

After collecting goods comments the needed preprocessing steps are performed. In the prepossessing process, attempts have been taken to correct spelling mistakes and similize opinion text. After the prepossessing stage, the product features are extracted based on the product specification. For example, for mobile product the features such as appearance, battery and etc are the characteristics which should be extracted. After features extraction, the respected lexicon is established. To do this, the dictionary is associated with the descriptive dictionary feature and a negating dictionary. In this phase, the dictionary is created manually. Then, opinion patterns are extracted. This step is done semi-automatically, and at the end, around 1186 patterns are found. Finally based on the found patterns, the category of goods reviews is done. The accuracy of this proposed method is 89%. The strength of this work is considering to product features, The subject which was neglected in previous methods for Persian opinion mining before. Research dataset contains three model of mobile which was collected from DG product website. A total of 1520 comments, which includes 5853 sentences, have been reviewed.

Mr Haj Mohammadi and Ebrahim in 2013[7] proposed a method for Persian opinion mining based on supervised SVM algorithm. This study is conducted to examine two classic classification algorithms namely SVM and

NaiveBayes .Ngrams such as unigram, bigram and trigram are examined in feature selection. Dataset, which is a film reviews, contains 200 negative and 200 positive comments. The small size of dataset is a drawback of this research. The dataset is collected from montaqi website. Finally the highest result is reported 72.66% which was achieved by unigram features, binary classification and SVM algorithm. Their research proved the advantage of SVM in Persian opinions classification. Generally, the SVM algorithm is known as best classification algorithm.

Mr. Bagheri and Sarah [4] with focusing on the feature selection, tried to explore new method for opinion mining in Persian language. Initially in prepossessing, stemming is done for eliminating word redundancy since in Persian language words often take suffix and prefix which increase the complexity of words. In these circumstances, we can make better classifications of comments. After stemming the features such as word repetition, word frequency variance and feature provided by the MI are used and features are refined. The authors apply modified MI feature selection method. Finally, in this research the highest reported accuracy is 85% which is achieved by NaiveBayes algorithm and the MI-modified feature selection method. Their dataset contains 1020 comments in which 511 comments were positive and 509 comments were negative. These comments are about mobile, which is more attractive good for buyers. Due to variant models, users are more likely to share their opinions. The results in Persian language are summarized in Table 1. Comparing methods presented for Persian opinion mining is almost impossible because different datasets are used in each method.

Table 1: Achieved results up to 2015 in Persian opinion mining

No	Dataset	Approach	Year	Paper	Result
1	Hotel	Lexicon	2012	[8]	78% accuracy
2	Film	Machine learning	2013	[11]	72.66% F
3 4	Mobile phone	Machine learning lexicon	2013 2014	[7] [6]	85% F measure 95% F measure
5	Phone DJ Commodity	Lexicon	2014	[10]	89.5% F measure
6	Hotel	Hybrid Hybrid	2015 2015	[9] [11]	83.57% accuracy 87% accuracy

Due to have scientific comparison, the methods should be applied on same dataset. On of weaknesses in Persian opinion mining is the lack of trusty dataset.

VI. IMPLEMENTATION

In this section, important and best words are selected with chi-square feature reducing method. Modeling results are compared using the chi-square technique in feature selection. Each algorithm performance is depicted in which horizontal axis exhibits the increase of feature and vertical axis shows the model accuracy and finally for the number of distinct attributes, the graphs will be plotted to compare the algorithms. The NaiveBayes algorithm with a number of 19,000 features, with the Binary feature and ngram range of 1 to 3, has the highest accuracy of 94.3%.

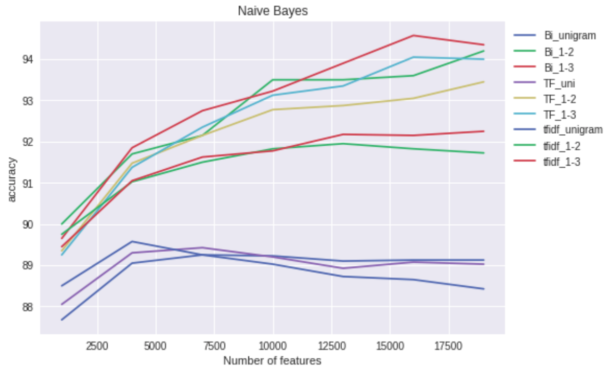

Fig 1: accuracy obtained by NaiveBayes and CHI2

TF-IDF is best method for SVM while using chi-square. This is well defined in Figure 2. The highest accuracy for this algorithm is 91%, which is obtained in the range of 19,000 features using IDF-TF and ngram in the range of 1 to 2.

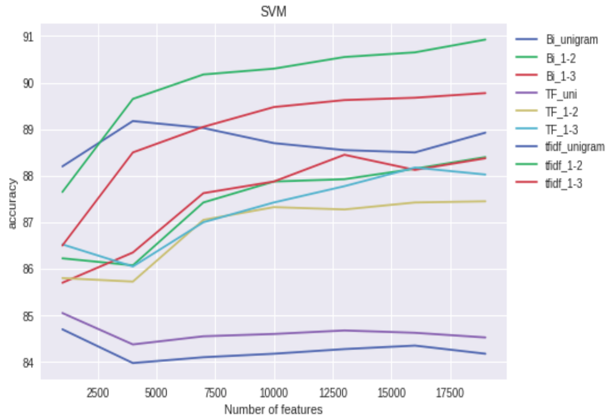

Fig 2: Accuracy produced by SVM algorithm and CHI2

The highest accuracy in the Logistic Regression algorithm is 89%, which is obtained via 19,000 features using Binary with ngram in the range of 1 to 3.

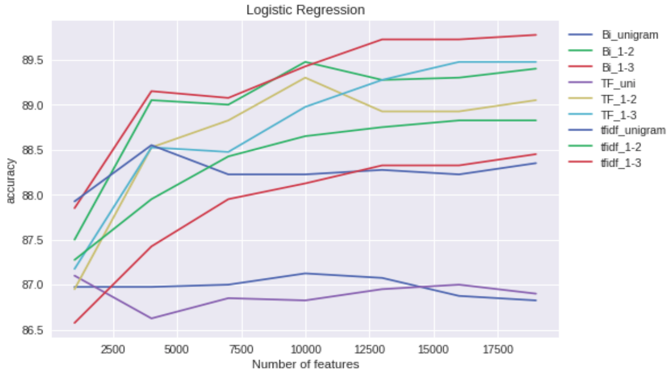

Fig 3: accuracy obtained by logistic Regression and CHI2

The weakest result in this section is the KNN algorithm, which showed an accuracy of 71% with k=5. Also, when the number of attributes increases, the behavior of the KNN algorithm is not very predictable.

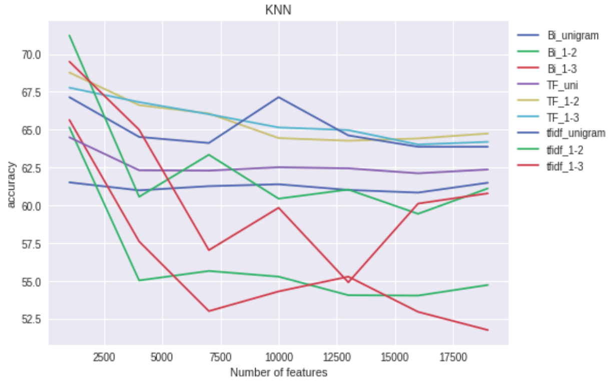

Fig 4: accuracy obtained by KNN and CHI2

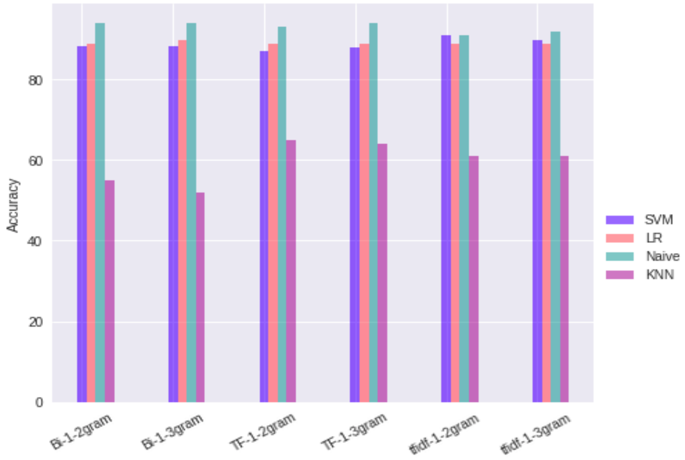

Fig 5: modeling performance produced by CHI2 and 19,000 features

VII. EVALUATION

After using different algorithms and comparing their performance, it is determined that NaiveBayes has the best result with Binary classification feature and unigram, bigram and trigram. Confusion table for the model obtained using CHI2 is shown in Table 4.

Table 4: confusion matrix for contracted model by CHI2

	Precision	Recall	F1-Score	Support
Positive negative	0.95 0.93	0.92 0.96	0.94 0.94	632 688
Avg/total	0.94	0.94	0.94	1320

VIII. CONCLUSION

Using the chi-square feature reduction methods, important words are selected and the modeling is done. Modeling results are compared using the chi-square technique in feature selection. Finally, for the number of specific attributes, the graphs will be plotted to compare the algorithms.

In contrary to UCF, when chi-square is used in feature selection, the NaiveBayes algorithm shows different results regarding to various type of feature extraction methods such as binary, term frequency, and tf-idf. This difference is well illustrated in Fig 1. The interesting point in this graph is that the binary feature, which is one of the simplest features for creating BOW, has far better results than other more common feature, such as tf-idf. Importantly in Unigram, when the number of attributes become over than 2000, the accuracy will drop sharply.

It indicates that the features, which are over than 2000, are noise features and they should be eliminated. However it will not happen in bigram, trigram since the number of features produced are greater than features in unigram. The highest accuracy with Naive Bayes algorithm using the Binary and the ngram range of 1 to 3 is 94.3% with 19,000 features, there was published an English paper in which reported precision on balance data was 87% [11]. As there was described in previous sections, in this paper the model performance is 94.3% on balance data (2000 negative sample and 2000 positive sample).

IX. RECOMMENDATION

Using the same dataset, it would be possible to do more opinion mining tasks in Dari language such as using evolutionary algorithms like genetic algorithm in feature selection. In addition, other modeling algorithms can be combined with other feature selection algorithms to produce more accurate result. Due to the fact that the same data set is used for selecting the attributes and training the model, the possibility of bias in feature selection may be a problem [10], so it is better to choose two data sets completely different from each other, to obtain more accurate and scientific results.

REFERENCES

[1] J. Yang, Z. Qu, and Z. Liu, “Improved feature selection method considering the imbalance problem in text categorization”, The Scientific World Journal, vol.2014, pp: 1-24, 2014.

[2] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, vol.5, no.1, pp.1–167, 2012.

[3] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, “A framework for sentiment analysis in persian,” Open Transactions on Information Processing, vol 1, no.3, pp.1–14, 2014.

[4] M. Saraee and A. Bagheri, “Feature selection methods in Persian sentiment analysis”, Natural Language Processing and Information Systems, pp.303–308, Springer, 2013.

[5] M. Shams, A. Shakery, and H. Faili, “A non-parametric lda-based induction method for sentiment analysis,” Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium on, pp.216–221, 2012.

[6] Saeedeh Ali Mardani and Abdollah Aghaee, "Providing a Supervisory Method for Farsi Dictionary in Persian using Dictionaries and Algorithms", SVM Journal of Information Technology Management, Volume 7, Issue 2, PP: 345-362, 2015.

DOI:10.17605/OSF.IO/UX4ED

[7] M. S. Hajmohammadi and R. Ibrahim, “A svm-based method for sentiment analysis in Persian language,” International Conference on Graphic and Image Processing, pp.876838–876838, 2013.

[8] S. Alimardani and A. Aghaei, “Opinion mining in Persian language using supervised algorithms,” Journal of Information Systems and Telecommunication, vol.3, no.11, pp.135–142, 2015.

[9] S. K. Singhi and H. Liu, “Feature subset selection bias for classification learning”, 23rd ACM international conference on Machine learning, pp.849–856, 2006.

[10] J.-C. Na, H. Sui, C. Khoo, S. Chan, Y. Zhou, “Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews”, Conference of the International Society for Knowledge Organization (ISKO), pp. 49 54, 2004.

[11] A.H.J Amoli, M.H Sultani, M Mohammadi “Enhanced Opinion Mining Approach using SVM in Persian Language”, International Journal of Advances in Electronics and Computer Science, Vol 5,Issue 5,pp:46-50, May 2018.