A Recommender System based on Collaborative Filtering using Polarity Improvement in Sentiment Analysis

Alaleh Sadat Hosseini Charyani[1*], Alireza Norouzi[2]
1-Department of Electrical and Computer Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: alaleh1987@gmail.com (Corresponding author)
2-Department of Electrical and Computer Engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran.
Email: nrzi.alireza@gmail.com

Received: September 2019 Revised: November 2019 Accepted: January 2020

ABSTRACT:
Sentiment Analysis, which is a new subfield of the processing of natural language and text mining, categorizes the texts based on the sentiment expressed in them. Sentiment plays a significant role in decision-making. So sentiment analysis technology has a broad scope for scientific applications. On the other hand, a huge amount of information in the world today is in the form of text. Therefore, text mining techniques are important. Exploring comments or analyzing sentiment as a branch of text mining, means finding the author's perspective on a specific subject. The Internet allows users to easily express their opinions and get informed about the opinions of others. The high volume and the lack of proper structure for the text of the comments provided on the web, make it difficult to use hidden knowledge within them. Therefore, it is important to provide methods that can prepare and provide this knowledge in a summarized and structured way. In this research, it has been tried to provide a fuzzy method for analyzing the following comments on news sites according to the text of the report. In this regard, it has been tried to investigate the relationship with the author's commentary and opinion in light of the subject of the text using the grammatical features of texts such as noun and verb, as well as sentimental load analysis of sentences. Subsequently, the method is evaluated by implementing it on the dataset collected from news and comments. The proposed method has 87% diagnosis accuracy.

KEYWORDS: Sentiment Analysis, Fuzzy Method, Grammatical Features of Texts, Text-Mining Techniques.

1. INTRODUCTION

Today, Recommender Systems are widely used to help users find their needs through a huge amount of information that is frequently available. The collaborative filtering recommender system finds users who are in agreement with the active users and then, recommends their favorite items to them. Recommender Systems are in fact a type of information filter. Today, these systems have become widespread, due to the growing volume of information and significant Internet growth, which in fact have been introduced to help Internet users find their favorite information. They have also solved the problem of data overflow. One of the most important and most widely used methods of designing the advisory system in recent years is the collaborative filtering. It was predicted that about 1.5 million articles have been published on Wikipedia encyclopedia site up to the beginning of 2007 or almost 250 million images are uploaded in Flickr images hosting and sharing service. Hence, it can be said that we are in the midst of a huge amount of data and information, which may cause ineffective or non-optimal choices without proper guidance and navigation. Recommender Systems affect user guidance among a huge amount of possible choices to achieve their favorite and preferred option, so that the process is personalized for the same user.

Recommender systems try identifying and suggesting the most suitable and closest product to users' taste by guessing the user's thinking (with the help of information that describes how he or his users are treated and their views). These systems are, in fact, the same process that everyone uses in everyday lives, and tries to find people with close proximity and consider them about our choices. Among the proposed techniques at the data level, the SMOTE algorithm is proposed by the paper [1], and it is sampled from the minority class. This technique is used in [2] to solve the problem of unbalanced data. In this paper, the data are first balanced and then extracted by SVM, MLP, and RBF classification algorithms. The models obtained from these algorithms were examined and compared after the balancing of the data from the point of view of accuracy. The results indicate an improvement in the accuracy of these models after balancing the data. The disadvantage of this article can be in assessing the performance of the model with an evaluation criterion. In this thesis, the effectiveness of the proposed method is reviewed with several criteria for the evaluation. Moreover, Smote's method is improved by a support degree technique in [3]. In this paper, it has been tried to help use to select a sample of minority classes and the sample of the new class by the proposed sampling method. The results indicate an increase in the performance of the classifier made using unbalanced data. Pasteur et al. [4] proposed a new approach called the Random Balance Method to balance the data to build a model by multiple classifiers. The innovation of this paper is that the class properties for each classifier are specified randomly. A new approach is proposed in [5] for the classification of data and the selection of samples from the majority class with a high imbalance rate using the sublevel method. The results show that the proposed method is more efficient than the weighed support vector machine algorithm for the classification of the unbalanced data set according to the sensitivity evaluation criterion. An online ensemble approach in [6] is proposed for nonsingular algorithms. This approach is two-layered. The results of the evaluation of the proposed method are based on three standard datasets and eight real-world collections indicating significant improvement over the other common online methods on the same features. Using the combination of cost-effective and cost-intensive learning techniques (CSL) in [7], it has been tried to solve the unbalanced class problem in educational data as well as the results of students' predictions. The results show that the proposed method works better than the base classification algorithms, decision trees, Bayesian networks, and supporting vector machines in order to counterbalance the unbalanced class.

2. RESEARCH METHODS

2.1. Architecture of the Proposed Method

In this approach, abnormal and noisy data are identified by the clustering technique in the first stage so that they are not used in the learning process. Then, the set of data for each cluster is balanced and the learning model is created for each data group. In the second step, the weight of each model is determined by the fuzzy system. Finally, the weight-based majority voting system has been used to collect the results. Fig. 1 shows the architecture of the proposed method.

In general, the proposed method consists of two phases of training and testing that is as follows:

• Training phase

1) First step: initial data clustering

To increase the efficiency of the results of the proposed method, the data are initially filtered by the clustering approach presented in [8]. In this approach, abnormal and noisy data are identified and will not be used in the learning process. In addition, the system can be used to determine the amount of data noise in the final result.

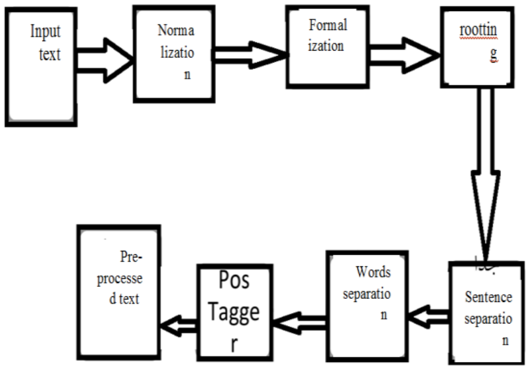

Fig. 1. Framework of the proposed method .

2) Second step: data grouping by secondary clustering

The re-cluster for a good data set is done after obtaining the noise data from the rest of the data. To this end, a clustering algorithm with 10 clusters is used to divide the data into 10 categories. At this stage, the clustering method has been used based on the K-Means method.

3) Step 3: Balancing the data of each cluster

At this stage, the data of each cluster is balanced by various balancing methods such as Sampling, Smote, Adaboost, etc. In this section, it has been tried use a different method of balancing for each category of data.

4) Step 4: Build Learning Models for each cluster

In this step, 10 learning models are created for the 10 clusters constructed from the previous step. In the proposed method, the Ensemble Classifier decision tree method is used. In this way, several learning models are made of an algorithmic category based on different data, and then the final result will be obtained by decision. The weight of each classifier is used to decide the majority voting method.

5) Step Five: Fuzzy Weighting System

In the training phase, weights can be assigned the classifier of each class. Accuracy is considered for weighting. The weight of each category is the same precision in prediction. In the proposed method, a fuzzy weighting system was used to weigh each of the 10 learning models. In a fuzzy system, its weight is determined based on the characteristics and behavior of the responses of each of the classifiers.

• Test phase

This phase initially enters into 10 learning categories by entering the new data and then, the closest class is identified. It also examined whether or not they are in the cluster of noisy data. Eventually all these inputs are ensembled and the result is determined.

2.2. Fuzzy weighting system

In this system, four variables are defined as inputs and one variable is defined as output. The internal structure of the fuzzy system used in the proposed method is shown in Fig. 2.

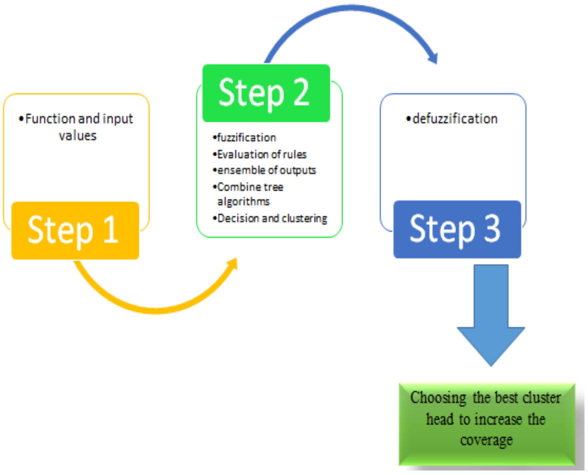

Fig. 2. The proposed method.

3. RESULTS

3.1. Class Distribution in the Learning Process by Considering Sentiment

It is proved that sentiment analysis is useful for recommender systems. A recommender system intends to predict the priority of a product for a target user. Mainstream recommender systems in an explicit data set, for example, group filtering operate on the matrix of content ranking and filtering on the metadata associated with a product. In many social networking services or e-commerce websites, users can write text reviews, comments, or feedbacks for a product. These user-generated texts provide a rich source of sentiment feedback from users about products and other cases. Such a text for a product can also reveal relevant features/aspects of the user for each feature. The feature/aspect that is described about a product in the text has a similar role with the metadata in the content-centered filtering, but the first one is more valuable to recommender systems. Since these features are widely criticized by users, they can be viewed as the most vital feature that effectively affects the user experience for a product, while the metadata of a product (usually provided by manufacturers not by consumers) may ignore features that are worrying for users. A user may have different sentiments for a variety of products with common features. Also, a distinctive feature of a product may get different sentiments from different users. The users' sentiments can be considered for a product as a multi-dimensional score rating that reflects their preference for a product. The most important unbalanced data challenge for a minority class is that it may identify its classifier as noisy data or outlier. As a result, the performance of the algorithm is low. On the other hand, noisy data or outlier must be identified so that a learning model is not created for them. For this purpose, clustering is used. In the process of clustering, abnormal and noise data are identified and will not be used in the learning process.

1) Balancing the data of each cluster:

After obtaining good and clean data, each cluster's data is balanced by various balancing methods such as sampling, SMOTE, AdaBoost, and so on. In this section, it has been tried to use a different method of balancing for each category of data.

2) Build a model for each balanced data group with multiple classifiers approach:

The model is created after getting balanced data. In the proposed method, the Ensemble method is used. In this case, several learning models are created from an algorithmic category based on different data and for each category.

3) Final decision to determine the output:

At this stage, the classifier of each class is first weighed and then, the fuzzy voting method is used to decide and ensemble the results.

- Percentage of majority output.
- Similarity of the output of the closest class to input data to the majority result.
- Is the output of the voting system from a minority class
- The similarity with the noisy cluster

The output of the fuzzy system is one of the following:

- Announce the majority result as output.
- Use existing weights for each category and re-calculate the result and announce it.
- Checking the noisy status of input data

In this research, the role of class distributions is investigated in learning a fuzzy classification of unbalanced data. This study comes from the fact that there is no guarantee that the information available for training represents (record) the distribution of test data. As a result, reducing the variance of output categories over the different distributions of the teaching class is a very important feature of classification.

3.2. Describing the Proposed Method

In order to analyze the approach of comment to data (for example, news), it is necessary to determine the subject of the discussion. To this end, the axes of discussion in the data text should be found. These axes are determined by identifying the names used in the data text. By specifying the subject of the discussion in the text of the data and finding a part of it in the text, it can be said that the text of the comment is related to the text. The author's perspective on issues can be reviewed by analyzing the sentiment load of sentences containing these names in the text and comments. By comparing these views, a final decision on the author's attitude toward the text can be achieved. In Fig. 3, the overall trend of the proposed method is presented. After the preprocessing operations are applied to news and comment texts, the proposed method is implemented in the next step as follows.

1. The sentences are separated by the expressed tools. This will be done by clustering the comments. The sentimental load of each sentence is extracted.
2. Using the "Decision tree" algorithm, the sentimental load of the sentence is determined. In this ensemble algorithm, the maximum of all values, regardless of the sign, is computed. User comments are fuzzificate with their sentimental load in each cluster and fuzzy rules are applied.
3. Now, names are identified and extracted in the context of the comment and the news.
4. For each sentence from the comment text, the total sentimental load of all the sentences in the text of the news with the current sentence of the same name is multiplied by the sentimental load of the sentence and the amount is allocated as the score of the sentence. A score is also considered for the sentimental load. This is a quarter-point advantage earned from the previous section. In fact, it is tried not to ignore the sentimental load of the comment in the case of lack of subscription because commenter expresses his opinion without referring to the data text in many cases. If most of these sentences are positive in content, the opinion will agree, and if most sentences have a negative value, the opposite comment will be recorded.

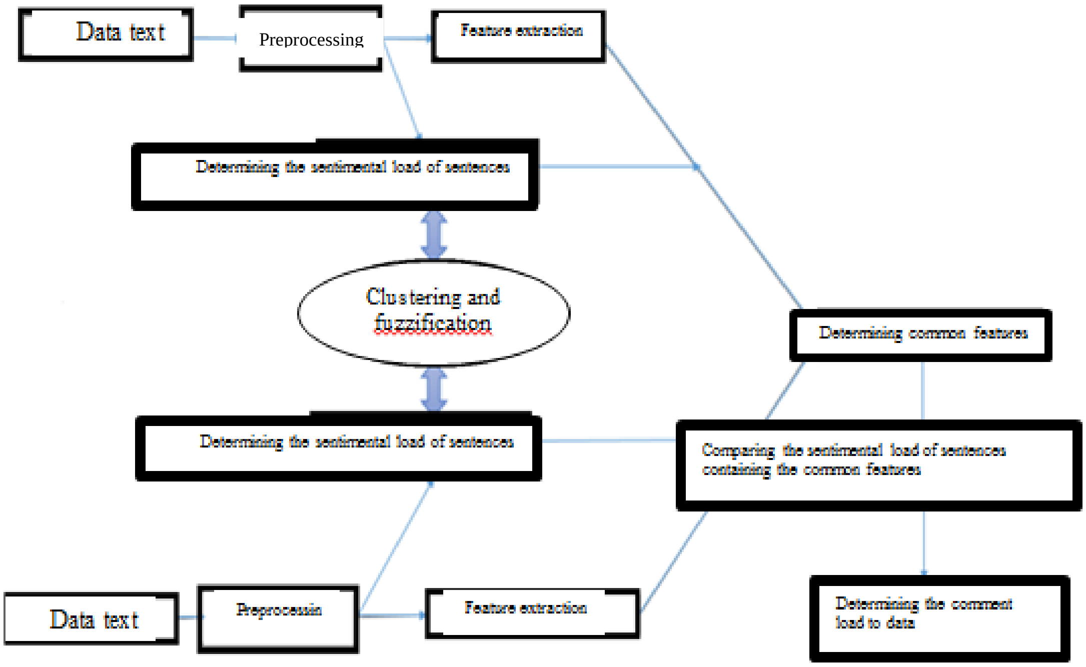
Data text
Preprocessing
Feature extraction
Determining the sentimental load of sentences
Clustering and fuzzification
Determining common features
Determining the sentimental load of sentences
Comparing the sentimental load of sentences containing the common features
Data text
Preprocessin
Feature extraction
Determining the comment load to data

Fig. 3. Framework of the proposed method.

The database used in this section consists of various comments and news collected from the four most popular news agencies. The information stored for each news item includes the news number, the news headline, the news text, the date and address of the news, along with the relevant comments. Considering the majority of votes, most of the labels were identified. Information about the news and the number of opinions in the newsletter is presented in neutral, positive, and negative, in Table 1.

Table 1. News agency comments in neutral, positive, and negative.

News agency	Number of news	Number of comments	Number of neutral comments	Number of agree comments	Number of agree comments
A	57	1379	512	132	735
Nuclear energy	56	1432	739	135	556
Online news	52	480	224	38	218
World news	52	550	211	58	136
Sports	47	356	199	21	136
Total	292	4631	2147	403	2081

In this section, the correct detection is shown with T and the false detection is shown with F. The correct diagnosis includes neutral detection with neutral label, positive detection with positive label, and negative detection with negative label. False detection is neutral detection with negative or positive label, positive detection with neutral or negative label, or negative detection with neutral or positive label. Using the proposed method, the data is placed in different clusters. Here, the number of clusters has a strong impact on the accuracy of the results. Fig. 4 shows the accuracy chart in terms of the number of clusters. It can be seen that the accuracy is very low in the absence of clustering and the accuracy is maximized for the 5 clusters.

Four aforementioned news agencies have been tagged with a fuzzy method based on the proposed method and the results are presented in Table 2.

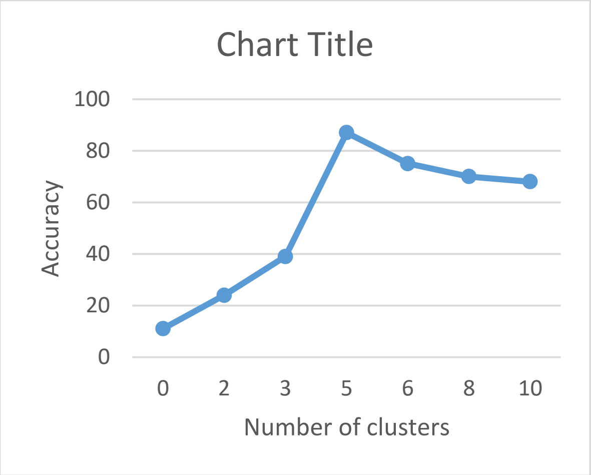

Fig. 4. Accuracy chart in terms of the number of clusters.

Table 2. The four news agencies mentioned by the proposed method based on fuzzy method.

Software Human	Neutral	positive	Negative	Detection rate
Neutral	1896	132	58	٪73
positive	321	1345	84	٪88
Negative	474	113	1237	٪89
Accuracy rate	٪78	٪68	٪83	
Total: 4567 Correct: 3346 Accuracy: 87%				

Using this criterion, 87% is obtained for the table results, which represents the accuracy of the method. Fig. 5 shows the accuracy diagram of the proposed method by dividing the data. Fig. 5 compares the simulation results of this study with the previous work in this area. Reference [9] used only the decision tree method to optimize the accuracy of the sensory analysis, and ultimately reached an accuracy as much as 81%.

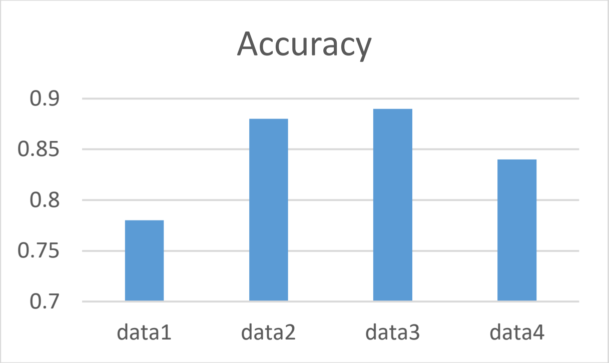

Fig. 5. The accuracy of the proposed method by dividing the data.

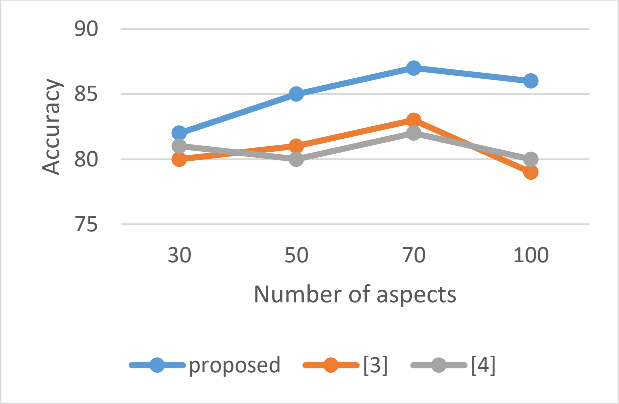

Fig. 6. Comparing the results of the simulation of this research with previous work.

In this paper, the runtime is compared in two algorithms i.e. algorithm [10] and the proposed algorithm. Fig. 7 shows the runtime graph for two different algorithms. It can be seen that the proposed run-time method has dropped sharply. Fig. 8 compares the memory consumption of the two methods. Here, the memory peaks of both algorithms are recorded for all data. Memory consumption results indicate that the proposed algorithm has a much lower memory consumption.

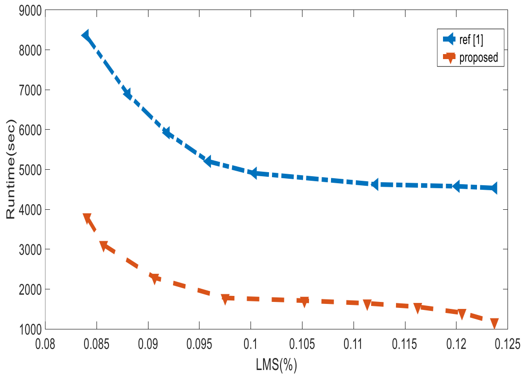

Fig. 7. Runtime graph for two different algorithms.

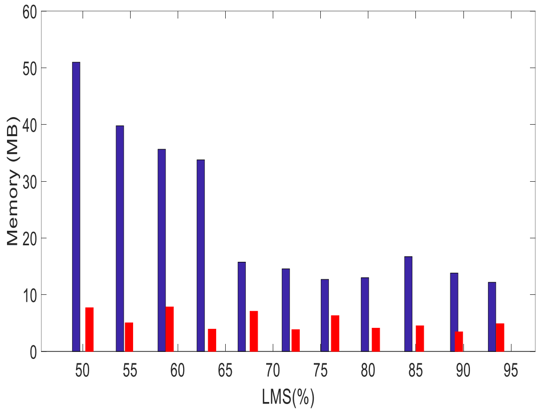

Fig. 8. Memory Consumption.

4. DISCUSSION AND CONCLUSION

In this study, a recommender system was investigated based on collaborative filtering using polarization improvement in sentimental system. Recommender systems are widely used today to help users find their needs through a huge amount of information that is frequently available. Today, these systems have been used extensively due to the growing volume of information and the significant growth of the Internet, which in fact have been introduced to help Internet users find their favorite information. . They have also solved the problem of data overflow. One of the most important and most widely used methods of designing the advisory system in recent years is the collaborative filtering. It was predicted that about 1.5 million articles have been published on Wikipedia encyclopedia site up to the beginning of 2007 or almost 250 million images are uploaded in Flickr images hosting and sharing service. Hence, it can be said that we are in the midst of a huge amount of data and information, which may cause ineffective or non-optimal choices without proper guidance and navigation. Recommender Systems affect user guidance among a huge amount of possible choices to achieve their favorite and preferred option, so that the process is personalized for the same user.

All aspects that a person likes or does not like about a particular product should be understood from online investigations. For example, user A and user B both care about the color of a cellphone device. If user A is very happy about a mobile phone but is not happy with the color, this cellphone should be not recommended to user B. Meanwhile, if the score of user B for the color of the cellphone is less than the average of all users, then user B is more stringent about color. So it's reasonable to consider the color of a cellphone more when offering to user B. Considering the points mentioned and the importance of quality and accuracy of offers, research in this regard is very important. In this research, the role of class distributions in learning a fuzzy classification of unbalanced data is investigated. This study comes from the fact that there is no guarantee that the information available for training represents (record) the distribution of data from the test. As a result, reducing the variance of output categories over the different distributions of the teaching class is a very important feature of classification. The correct detection is shown with T and the false detection is shown with F. The correct diagnosis includes neutral detection with neutral label, positive detection with positive label, and negative detection with negative label. False detection is neutral detection with negative or positive label, positive detection with neutral or negative label, or negative detection with neutral or positive label. Using the proposed method, the data is placed in different clusters. Using this criterion, 87% is obtained for the table results, which represents the accuracy of the method.

REFERENCES

[1] J.J. McCauley, J. Leskovec, "**Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text**", *in: Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China,* October 12–16, 2013, pp. 165–172, 2013.

[2] D.M. Blei, A.Y. Ng, M.I. Jordan, "**Latent Dirichlet Allocation**", *J. Mach.Learn. Res. 3* pp. 993–1022, 2003.

[3] Y. Jo, A.H. Oh, "**Aspect and Sentiment Unification Model for Online Review Analysis**", *in: Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China,* February 9–12, 2011, pp. 815–824, 2011.

[4] A. Popescu, O. Etzioni, "**Extracting Product Features and Opinions from Reviews**", *in: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference,* HLT/EMNLP, 2005.

[5] A. Pang, L. Lee, S. Vaithyanathan," **Sentiment Classification using Machine Learning Techniques**", *in: Proceedings of EMNLP,,* pp. 79-86, 2002.

[6] http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

[7] A. Titov, R.T. McDonald, "**A Joint Model of Text and Aspect Ratings for Sentiment Summarization**", *in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008, June 15–20, 2008, Columbus, Ohio, USA*, pp. 308–316, 2008.

[8] http://www.aclweb.org/anthology/P08-1036.

[9] Y. Wu, M. Ester, Flame,"**A Probabilistic Model Combining Aspect Based Opinion Mining And Collaborative Filtering**", *in: Eighth ACM International Conference on Web Search and Data Mining*, pp. 199–208, 2015.

[10] B. Pang, L. Lee, "**Opinion Mining and Sentiment Analysis**", *Found. Trends Inf. Retr*. 2 (1–2), pp.1–135, 2007.

[11] W. W. H. Wang, "**Opinion-enhanced Collaborative Filtering for Recommender Systems Through Sentiment Analysis",** *New Review of Hypermedia and Multimedia*", 2015.

[12] G. K. Mukund Deshpande, "**Item-based Top-N Recommendation Algorithms**," *ACM Transactions on Information Systems*, Vol. 22, pp. 143-177, 2004.