Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 84 (2016) 86 – 93

Procedia Computer Science

7th International conference on Intelligent Human Computer Interaction, IHCI 2015

Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering

Deepa Anand[a]*, Deepan Naorem[a]

[a]*CMR Institute of Technology,AECS Layout, Bangalore 560048, India*

Abstract

Aspect based Sentiment Analysis (ABSA) is a subarea of opinion mining which enables one to gain deeper insights into the features of items which interest the users by mining reviews. In this paper we attempt to perform ABSA on movie review data. Unlike other domains such as camera, laptops restaurants etc, a major chunk of movie reviews is devoted to describing the plot and contains no information about user interests. The presence of such narrative content may potentially mislead the review analysis. The contribution of this paper is two-fold: a two class classification scheme for plots and reviews without the need for labeled data is proposed. The overhead of constructing manually labeled data to build the classifier is avoided and the resulting classifier is shown to be effective using a small manually built test set. Secondly we propose a scheme to detect aspects and the corresponding opinions using a set of hand crafted rules and aspect clue words. Three schemes for selection of aspect clue words are explored - manual labeling (M), clustering(C) and review guided clustering (RC). The aspect and sentiment detection using all the three schemes is empirically evaluated against a manually constructed test set. The experiments establish the effectiveness of manual labeling over cluster based approaches but among the cluster based approaches, the ones utilizing the review guided clue words performed better.

Peer-review under responsibility of the Organizing Committee of IHCI 2015

*Keywords:*Aspect Based Sentiment Analysis; Feature Based Opinion Mining; Naive Bayes; Natural Language Processing

1. Introduction

The web provides an excellent platform for users to express their views and opinions on a gamut of topics varying from products in e-commerce, political reforms they desire to their feelings on various day to day happenings. Free flow reviews allow the users to be much more expressive than plain rating systems by providing the freedom to articulate the aspects about the topic which is most important to them. The various forms of such user expression in the form of free flowing text manifesting in various forms such as blogs, reviews, and tweets provide a goldmine of

Peer-review under responsibility of the Organizing Committee of IHCI 2015
doi:10.1016/j.procs.2016.04.070

information which can be exploited in various ways, for product Recommendations, text summarization etc and has given rise to the broad area of research known as Opinion Mining [1].

Sentiment Analysis is a sub area of Opinion Mining which aims at inferring the emotion of a user towards a particular item expressed through the reviews. Though sentiments could fall under several classes such as anger, misery, happiness etc, the primary emotions of interest in most cases is the positivity or negativity of the review. Plenty of research has gone into accurate classification of reviews[1][2] under these categories but a variation of sentiment analysis called aspect based or feature based sentiment analysis[1][3][4] delves deeper to understand the reason for positive or negative orientation and investigates the item features for which the user has expressed an opinion. For instance, in the movie review snippet given below one can infer that the user is happy with choice of cast but not the story.

... the might of Tom Hanks and Julia Roberts could not salvage this horrible joke of a story ...

Many techniques have been proposed for the problem using a combination of common sense knowledge, based on Natural Language Techniques normally encoded as rules, Supervised and unsupervised techniques[9][12]. The common sense knowledge encoded as rules may be domain dependent or domain independent. ABSA has been attempted in various domains such as laptops [3], restaurants [9], movies [6][7] etc. Though there exists labeled dataset for applying supervised techniques in the laptop and restaurant domains such a dataset is lacking in the movie domain and thus we attempt an approach in this direction. We use a combination of domain knowledge along with semi/un supervised techniques to build an aspect and sentiment detector. An important step in achieving this is to tackle the problem which is unique to the movie domain i.e. reviews containing narrative content which may mar the accurate computation of the actual polarity of the review and its aspects. However creation of manually labeled data for filtering out plot sentences is time consuming and we outline an approach to overcome this hurdle.

To detect aspects and sentiments we employ hand crafted rules based on the work presented in [6][9] to detect potential sentiment aspect pairs. The ABSA task is then two-fold: detect the aspects (map the potential aspect words detected to the corresponding aspect categories) occurring in the sentence and estimate the polarity of the user towards these aspects. To detect the aspect category we make use of words which would be indicators of the presence of the aspect - we call such words aspect clue words. Three approaches for selecting the clue words are attempted here: manual, semi-supervised clustering and review word guided semi-supervised clustering. Whereas the former requires manual effort and needs to be tailored for a domain the clustering approaches require minimal intervention. We thus justify the effort expended in manual selection by evaluating the performance improvement it offers over un/semi supervised approaches. We also demonstrate that a review word based supervised clustering is able to offer better performance than the method without.

The paper is structured as follows: the next section discusses the relevant background for ABSA; Section 3 presents the proposed approach whereas section 4 presents the experimental results. Section 5 concludes the work and describes the future enhancements planned.

2. Literature Survey

The ease of generating plain text content in different forms varying from blogs and tweets to reviews, discussion forum posts etc have opened endless possibilities in the ways in which these can be utilized. Ecommerce companies tap into the goldmine of user generated content available aplenty, in order to serve customers better and boost their sales. Sentiment Analysis [2] and particularly ABSA [1], which poses several challenges inherent to processing natural language, is an active area of research in this direction. Several challenges which still have not been adequately addressed are implicit aspect detection, mapping aspect words to categories, resolving anaphora references etc [1]. Researchers combine techniques from common sense rules, unsupervised and supervised techniques to perform these tasks.

The lack of labeled data has lead to several researchers to explore unsupervised learning techniques to learn both aspects and their sentiments expressed in plain text. Particularly the fact that aspects are normally described by opinion words and opinion words in turn will have a target aspect can be used to iteratively expand the sentiment and aspect lexicon. The expansion is done with the help of rules to associate aspects and sentiments [1]. Jo and Oh

[3] propose Sentence-LDA (SLDA), a probabilistic generative model to unify the aspect and sentiment modeling so that sentiments and the corresponding aspects can be extracted without supervision. A similar approach is followed in [8] where the authors explore both a supervised and unsupervised probabilistic technique in the domain of beer reviews. A double propagation method using propagation rules in the form of dependency relations and restrictions is explored in [5] for learning the aspect terms without supervision. The authors also filter out many of the irrelevant aspects by using stop words and PageRank algorithm on the word graph formed where an edge indicates that a word was discovered due to the word at the other end of the edge. A similar approach by the same authors [10] explores a graph based technique for extracting most relevant aspects and semantic word similarity based approach to detect sentiment words. Authors in [6] also follow an unsupervised learning technique albeit for a slightly different task i.e. for summarization of multiple reviews for a single product. They also employ several rules on the dependency relations between words in a sentence to detect aspects and sentiments. The aspect-sentiment phrases are then clustered in order to derive the summarized opinions of several users presented in the form of a count of number of people having a positive/negative polarity for different aspects.

Supervised techniques for aspect based sentiment analysis have obviously proven much better than unsupervised ones in most domains but are challenged by non availability of labeled data in certain domains. Authors in [9] combine common sense rules along with supervised techniques which use sentence level features such as 'Sentic' feature, part of speech feature etc to detect sentiments. The model learnt using supervised technique is employed only if the common sense rules do not apply. A conditional Random Field approach to find opinion targets is presented in [11] where the authors utilize features such as the token, part of speech, nearest noun, direct dependency and opinion sentence to determine eligibility of word as an aspect. The authors also incorporate anaphora resolution in this process. Methods for semi-supervised learning also abound. A distant supervision based technique for aspect-sentiment extraction is proposed in [12] where the authors claim to allow for a vast reduction in the supervision required and allows construction of large labeled data corpus. The idea is to use the section headings in reviews to capture aspects and overall sentiment for the entire review, normally available as a rating, to infer aspect level sentiments. In this work, we explore extracting aspects from text reviews by filtering out irrelevant aspects. However, unlike [10] the relevance of a word as an aspect is measured in terms of it being pertinent to reviews as opposed to plots, which should be effective in the movie domain. Section 3 explains our approach.

3. Proposed Work

The proposed approach attempts to perform aspect based sentiment analysis in two stages - filtering statements from the review pertinent to sentiment extraction, extracting sentiments from the reviews and associate it with the corresponding aspect categories. The first subsection explains our method to perform sentence filtering, subsection 3.2 then details the method to perform sentiment-aspect mapping.

3.1. Filtering Reviews

The domain of movies is fraught with reviews a major chunk which consists of narrative content. A natural technique for detecting the polarity of sentiments is to look for opinion words which may convey positive or negative sentiments on various aspects. The presence of plot sentences in the review might mislead the process of sentiment detection. Sample the review below:

*... Jumanji is a catalyst of fate, an **evil** entity that does not bring about awe as it did in Allsberg's book, but fear. The world that Allen Parrish is doomed to be incarcerated in the jungles of Jumanji is another one, but that period of time is nonexistent in the world he will eventually live in. This is the beauty of the movie...*

In the above review the words/phrases highlighted in green are words that might indicate negative sentiments whereas the word highlighted in blue conveys positive sentiment. We can observe that the first few sentences implying an overall negative polarity are words belonging to sentences describing the plot and thus must not be utilized in the review polarity determination.

It is evident therefore, that in the movie domain, particularly, review sentence filtering is required in order to ensure accuracy of sentiment detection. It is to be noted that this is different from subjectivity classification [1] where sentences which lack subjectivity are eliminated as not conveying any opinion. Here the filtering is done on relevance rather than subjectivity. It is also to be noted that other problems such as sentences denoting a target other than the item being reviewed exists (..*Terminator-I was good movie, better than this movie*...). However, in this work we only address the problem of filtering out the plot sentences since a larger percentage of sentences fall in this category.

The challenge however is the lack of labeled datasets for such classification. We build our dataset in the following manner: We extract plots of 1000 movies by crawling the IMDB website. We also make use of the Amazon dataset [13] containing around 7 million reviews. From the set of reviews we retain the reviews whose length is below a threshold (determined empirically) for training. The logic is that shorter reviews typically contain only the opinions expressed by the author and do not discuss the plot. However mandating that the reviews be very short would also affect the size of the training set and consequently the classification accuracy and therefore the threshold should be chosen carefully (The method employed to derive the threshold is explained in the Experiment section - Sec.4). The aim is to train a classifier to categorize a sentence into the class plot or review. Naive Bayes' Classifier is used for this purpose. Each item in the set of plots, *P*, and the set of reviews *R*, is split into n_p plot training sentences and n_r review training sentences.

Given a sentence S to classify we find the probability of the sentence belonging to a plot or review by computing;

where P(Plot/w) is defined as:

$$P(Plot/S) = \prod_{w \in S} P(Plot/w) \tag{1}$$

$$P(Plot / w) = \frac{P(w / Plot) \times P(Plot)}{P(w)} \tag{2}$$

P(w/Plot) is calculated as:

$$P(w / Plot) = \frac{n_{wp}}{n_p} \tag{3}$$

where n_{wp} is the number of plot sentences which contain the word *w*. The probability *P(Review/S)* is calculated in a similar manner. A sentence is classified as plot or review based on the comparison between *P(Plot/S)* and *P(Review/S)*. Detailed discussion on the procedure adopted to arr ive at the optimal value of the threshold *T* is presented in the Section 4.

3.2 Aspect and Sentiment Detection - Manual labeling vs. Unsupervised Techniques

Several papers in literature lay out rules for detecting aspect words which might be targets of opinions based on common sense knowledge of the natural language [9]. We chose to follow the rules laid out in [6][9] to detect aspects. The rules are applied to the output of dependency parsing by Stanford Parser [1] of each statement of the review statement. The Stanford dependency parser deduces the relations between words in a sentence. For instance, for the sentence "Mary walked home" it would derive the relations such as the one relating the words *walked* and *Mary* with the subject relation. An aspect-sentiment word pair identified using the rules, was considered only if the potential opinion word conveyed some sentiment. The polarity of a word is computed using SentiwordNet[1] which assigns three scores to each word in the range [0,1] - the positivity, negativity and the objectivity. The three scores

sum up to one. The scores are assigned to a (word,part of speech) pair and therefore the same word may have different scores according to the part of speech. For instance the word *good* may occur as an adjective as well as a noun and so may have differing scores accordingly. The actual polarity of the word is then computed as :

(4)

The word w is *score(w) = positivity(w)- negativity(w)* deemed to convey a sentiment only if the *score (w)* is non zero. Once the aspect sentiment pair is identified the main task in aspect detection is to map the detected aspect words to the correct aspect category. For instance considering the review sentence below:

'... The performances in the movie are great ...'

The opinion word "great" describes the aspect word "performances" which in turn indicates the aspect category "acting". The various aspect categories considered are listed in Table 1. One way to map potential aspects into categories is using aspect clue words which imply the corresponding aspect category. Thus for each broad aspect category we maintain a list of words which if described by an opinion word would indicate the user view about the category.

Table 1. Aspects and associated Aspect Clue words

Aspect Category	Aspect Clue words
Acting	Chemistry, performance, Charm, comedian ...
Direction	Director, filmmaker, vision ...
Screenplay	Sequence, script, lines, editing, screenwriting ...
Sound effect and music	Score, music, vocals, audio ...
Story	Mystery, spoof, thriller, twist, shock ...
Visual Effects	Effects, 3d, scenery, photography, camera, cinematography
Film on the whole	Flick, remake, sequel, classic, entertainment ...

In this work we explore different ways to frame the aspect clue words. The first method which is manual is constructed by first splitting the reviews into sentences. From each sentence (aspect, sentiment) pairs are extracted using the method outlined above. A count of the number of times an aspect has been described by various opinion words is computed and the aspect words having a citation count of more than a threshold value (100, in our case) are examined manually and assigned various aspect categories. We refer to this method as the manual method (M). A method generally explored in literature is some form of clustering [1][6] to associate aspect words with categories. We experiment with the one proposed in [6] where the authors use a K-Medioids algorithm for clustering and estimate word similarity based on the Jcn[6] semantic Wordnet based similarity metric. However, we make a few minor modifications to the proposed method - whereas the authors considered the entire aspect-sentiment phrase for similarity computation we only considered the aspect. In addition in our case the similarity between two aspect words is a combination of its semantic similarity as well as the co-occurrence degree, and is computed as:

$$sim(x, y) = \alpha\, Jcn(x, y) + (1 - \alpha)^{n_{xy} / \max(n_x, n_y)} \tag{5}$$

where n_{xy} is the number of different sentences in which the x and y have occurred together. Moreover, the initial seeds to the K-Medioids algorithm is given as the seven aspect category words listed in Table 1. The initial seeding of the cluster centers make it semi-supervised clustering. This method that we adopt will be henceforth referred to as Keyword Clustered (KC). On manual examination of the words clustered together with key aspect category terms we found a large percentage of terms which cannot act as an aspect word. For instance words such as 'queen', 'fairy', 'town' etc would seldom act as an aspect of a movie and more often than not would originate from plot sentences which would have been wrongly classified as review. We therefore use the criterion - P(Plot/w) > P(Review/w) (computed in Section 3.1), as a means to filter out potential aspect words which are more indicative of the plot than

reviews. The remaining aspect words are then clustered in a manner similar to KC method. This preprocessing of data before clustering would be henceforth referred to as Keyword Review Filtered Clustering (KRC). The next section compares these techniques experimentally.

4. Experimental Evaluation

In this section we present the results of experiments conducted to evaluate the proposed technique for review filtering and aspect-sentiment detection. We chose the Amazon movie review dataset [13] for evaluation. The dataset consists of 7,911,684 reviews provided by 889,176 users on 253,059 movies/TV shows. The dataset consisted of several duplicate items and we preprocessed the dataset by removing the duplicate items. For the purpose of obtaining plot data for training the classifier explained in section 3.1 we used plot data of 1000 movies crawled from IMDB.

The first task of tailoring a training dataset for filtering reviews was performed by using the plot texts from IMDB and the set of review data whose length was below the threshold (T) from the Amazon review dataset. The value of T was varied and the intention is to determine the optimal value of T which would maximize the classifier accuracy. To do this, we manually labeled a small set of sentences - 100 each of review and plot. The threshold, T, was varied from 100 to 1000 in intervals of 100 each and then from 1000 to 6000 in intervals of 1000. For each value of T, reviews having length less than T were chosen to be in the training set and the model trained thus, was used to predict the label of the manually labelled statements. The value of T, resulting in highest value of accuracy of classification was fixed as the threshold to be used. The estimate of P(Plot) and P(Review) was estimated by a different validation set T' consisting of manually labeled 100 randomly picked review statements. The quantity P(Plot) was roughly estimated at around 0.23 which was used in Eq. 2.

A simple estimation of accuracy computed as a ratio of number of correct predictions to the total size of test set would be biased towards methods predicting 'review' most of the times since 'review' statements constitute a major chunk of statements. Thus we estimate the accuracy while predicting plots and reviews separately. Accuracy for predicting plot statements, given n_{PP} is the number of plot statements predicted to be plot (true positive) and n_P being the number of actual plot statements, is given by the formula:

$$PAccuracy = \frac{n_{PP}}{n_P} \tag{6}$$

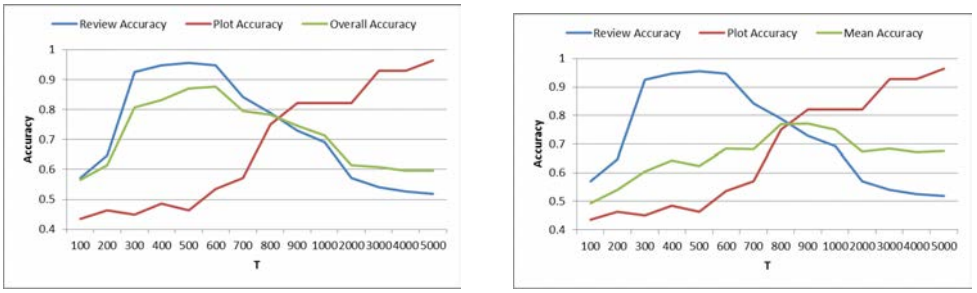

Fig 1: Estimation of T: (a) review accuracy, plot accuracy and Overall accuracy (b) review accuracy, plot accuracy and harmonic mean; with variation in values of T

In a similar manner we also estimate the accuracy while predicting reviews. The optimal classifier would be the one maximizing, both, review as well as plot accuracy and thus we chose a harmonic mean of the plot and review accuracy as the deciding factor for choosing T. Fig 1(a). contains all the plots showing overall accuracy and the plot and review accuracy in a single graph. We observe that as the threshold increases the accuracy of review prediction increases for review sizes from 100 to 400 and then shows a declining trend. The plot prediction accuracy on the hand increases with the increase in the size of T. We observe the overall accuracy follows a trend similar to review accuracy demonstrating the influence of review accuracy in the calculation of overall accuracy. Whereas Fig. 1(b)

shows the review and plot accuracy along with the harmonic mean of the two. We observe that the mean accuracy increases (with T in the range 100-900) with increase in review sizes and then decreases. The maximum value of the mean is obtained for T=800 and this is the final value of T chosen.

For detecting the aspect and the corresponding sentiments we compare the three methods for aspect detection - manual method (M), Keyword Clustered (KC) and Keyword Review Filtered Clustering (KRC). These methods frame the clue words to be used for associating the potential aspect words with the respective categories. For the clustering task we chose a random set of 20000 reviews from the Amazon dataset and extracted the aspects as outlines in Section 3.2 and the frequently mentioned aspects clustered using K-Medioids. The test data consists of 100 review sentences marked with the aspect categories and the corresponding sentiments. We measure the ability of the various techniques to derive clue-words w.r.t. their ability to detect an aspect in a review and correctly identify the sentiment associated with the aspect. The accuracy in detecting aspect is computed as the ratio of number of aspects categories detected and the total number of aspect categories across review statements. Fig 2(a). shows the result of accuracy of aspect detection for the three methods over 10 runs with the test data where a random 80% of the test data was used. The graph clearly shows the marked improvement in performance using manually assigned aspect words. Among the methods using clustering the one retaining select keyword (KRC) that is indicative of review only has a benefit and leads to better aspect detection accuracy than the plain KR technique.

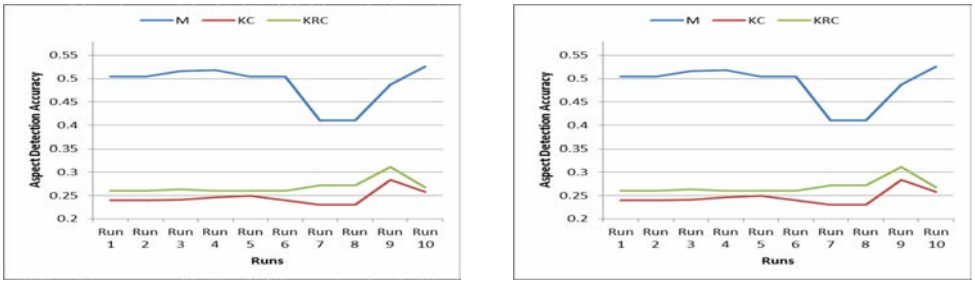

Fig. 2. Comparison of (a) Aspect Detection Accuracy of M, KC and KRC methods; (b) Aspect-Sentiment Detection Accuracy of M, KC and KRC methods. 5. Conclusions and Future Directions

Let R be the set of reviews. For a review r, let P_r and N_r denote the set of aspect categories which have been rated positively and negatively in r and if P'_r and N'_r denote the set of aspects predicted to be positive and negative for the same review then we compute the overall accuracy of sentiment prediction as:

$$SAccuracy = \frac{\sum_{r \in R} \frac{|P_r \cap P'_r| + |N_r \cap N'_r|}{\max(|P_r \cup N_r|, |P'_r \cup N'_r|)}}{|R|} \quad (7)$$

Fig 2(b). shows the result of the sentiment accuracy obtained on 10 runs with the test dataset using all the three methods. The accuracy of all three methods is low but again it is observed the manual method (M) outperforms the other two. Among the clustering methods, again the ones filtering out non review based aspect words (KRC) outperforms the simple clustering method (KC)..

The main contributions of this paper are two-fold. For the problem of aspect based sentiment analysis in the movie domain we lay out a technique to solve the problem unique to the movie domain - i.e. of weeding out narrative content from reviews so that the sentiment and aspect extraction is more focused and not misled by irrelevant statements. To this end, we introduced a technique to train a classifier with minimal supervision and conducted experiments to tune the parameter involved. We also introduced a preprocessing technique to use the relevance of words computed above to decide on the set of words given to the clustering algorithm for deriving automatic classes for aspect category mapping. Experiments conducted proved that though automatic methods using

the preprocessing steps are able to offer an improvement over plain clustering ones, the manual method for mapping aspects to categories is the most effective for both aspect as well as sentiment prediction tasks.

In this work we have only tested the effectiveness of the methods on individual statements. However the effectiveness of filtering plot sentences out should have a good impact on the overall aspect-sentiment extraction on the review as a whole but it needs to be coupled with effective methods for aggregating opinions across different review sentences. Moreover the lack of large datasets for testing the aspect-sentiment extraction task in the movie domain can be mitigated by applying the user interests derived through these methods, to the domain of Recommender Systems and measuring the relative effectiveness of various methods [11]. We plan to explore these directions in the future.

References

1. Liu B. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers; 2012.
2. Mukherjee S, Bhattacharyya P. WikiSent: Weakly Supervised Sentiment Analysis through Extractive Summarization With Wikipedia, *Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science*. 7523; 2012. p. 774-793.
3. Jo Y, Oh A, Aspect and Sentiment Unification Model for Online Review Analysis. In: *Proceedings of the fourth ACM international conference on Web search and data mining*; 2011. p. 815-824.
4. McAuley J, Leskovec J, Jurafsky D. Learning Attitudes and Attributes from Multi-Aspect Reviews, In: *Proceedings of IEEE Conference on Data Mining 2012*. 2012. p. 1020-1025.
5. Garcıa-Pablos A, Rigau G. V3: Unsupervised Generation of Domain Aspect Terms for Aspect Based Sentiment Analysis. In: *Proceedings of 8th International Workshop on Semantic Evaluation*; 2014. p. 833-837.
6. Bancken W, Alfarone D, Davis J, Automatically Detecting and Rating Product Aspects from Textual Customer Reviews .In: *Proceedings of DMNLP, Workshop at ECML/PKDD*. 2014. p. 1-16.
7. Thet TT, Na J, Khoo CSG, Aspect-based sentiment analysis of movie reviews on discussion boards. *J Inf Sci* 2010; 36 (6), p. 823-848.
8. Brody S, Elhadad N. An Unsupervised Aspect-Sentiment Model for Online Reviews. In: *Proceedings of HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*; 2010. p. 804-812.
9. Poria S, Ofek N, Gelbukh A, Hussain A, Rokach L. Sentic Demo. A Hybrid Concept-Level Aspect-Based Sentiment Analysis Toolkit. In: *Proceedings of 11th European Semantic Web Conference 2014 (ESWC2014)*; 2014. p. 41-47.
10. Garcıa-Pablos A, Cuadros M, Rigau G, V3. Unsupervised Aspect Based Sentiment Analysis for SemEval-2015 Task 12. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*; 2015. p. 714–718.
11. Jakob N, Extracting Opinion Targets from User-Generated Discourse with an Application to Recommendation Systems. PhD Thesis; 2011.
12. Broß J. Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques. PhD Thesis. 2013.
13. Leskovec J. (n.d.). Web data: Amazon movie reviews. Retrieved August 15, 2015, from https://snap.stanford.edu/data/web-Movies.html.