Simple and Efficient Approach to the Aspect Extraction from Customers' Product Reviews

Nadezhda Chechneva
Saint Petersburg State University
Saint Petersburg, Russia
chechnevanadegda@mail.ru

Abstract— **In this paper we present an approach to the one of the most popular natural language processing tasks of automatic aspect extraction from product reviews. Our approach is based on using clustering of word embeddings, morphological features, information about syntax dependencies and word frequencies. We use these features in the well-known machine learning method – the Decision tree. The primary evaluation of our method quality for the task of identifying the explicit aspects from the reviews demonstrate good performance in the precision and recall for cross-domain aspect extraction task.**

I. Introduction

In recent decades, there has been growing interest in a special type of text – product reviews. They are texts which contain an author's assessment of a product or service, and are posted on the specialized Internet resources.

This interest is due to the fact that this type of text has a powerful communicative potential, often having a direct impact on the decision to purchase a product, choosing a service provider, or visiting some place.

According to some studies [1], the proportion of people who regularly read other people's opinions in reviews and comments is more than 80%. Most often, reviews are read before buying consumer goods, when choosing travel, medical services, as well as before visiting cafes, restaurants and cultural events.

On the one hand, reviews provide communication between consumers and producers. Consumers can evaluate any product or service, express their opinion about the company or brand, and producers, in turn, can monitor their reputation and the quality of their product through such a feedback.

On the other hand, reviews provide an opportunity for information exchange between consumers and potential consumers. Consumers can explicitly advise a product to purchase or, conversely, warn against the purchase of a product, or simply share the experience of using a product and their impressions. Potential consumers, getting acquainted with the reviews on the product they are interested in, can make a choice in favor of this product, or refuse to purchase it.

Thus, producers and service providers, as well as real and potential consumers, act as communicants in the reviews.

Despite the huge variety of subject relatedness of reviews, it is likely that all of them have common structural, semantic, lexical features that make it possible to attribute them to a separate genre - online review.

Structural characteristics include the presence of the following components:

- subject of review - its author,
- object of review - product, service and their components or features,
- date of writing,
- title,
- rating - most often on a 5-point or 10-point scale,
- advantages, disadvantages and comment,
- aspects, that are rated in reviews.

Some blocks may be present in a particular review, or may be absent. It depends on the given review structure on different resources.

From the semantics point of view, reviews not only contain information about the experience of using a product or service, but are also indicators of the deep value orientations of people.

What evaluations refer to is called aspect. The structure of the evaluated parameters, or aspects, depends on the subject area of the review. Different product categories will have their own set of aspects. For example, the category television will have the following aspects: 'screen', 'image quality', 'sound', 'screen size', etc. And for the category teapots, there will already be completely different aspects: 'volume', 'boiling speed', 'power indicator', etc. However, some aspects will be found in almost all categories. For example, the 'price' of a product.

In order to better understand the structure of aspects, to highlight the unique and universal aspects we need to extract them. In this article, we propose an approach to automatic extraction of aspects for different categories of products.

II. Aspect extraction approaches

First of all, it should be noted that the question of what is considered an aspect is ambiguous.

Three groups of aspects were distinguished by the expert group that was preparing data for the aspect-oriented competition SentiRuEval [2] held in 2015:

1) Explicit aspects. They clearly mention target aspects and do not contain their assessment (for example, 'screen', 'speed', 'coffee'). Subjective expressions (i.e. words and phrases expressing tonality, opinion or assessment) are not considered as aspect terms.

2) Implicit aspects: They are characterized by the fact that they contain both an indication of an aspect category and its assessment ('tasty', 'comfortable', 'convenient', etc.).

3) Tonal facts: They are mostly quite verbose expressions. They obviously belong to one of the aspect categories, but do not contain explicit appraisal words. Tonal facts carry some knowledge of the world that communicates negative or positive meaning (or example, 'a plastic case', 'mold was formed', 'instructions in Russian'). In some cases, tonal facts can be expressed in one word (for example, broke).

In our research, we will consider as aspects only terms of the first type, i.e. explicit aspects, since implicit terms and tonal facts, in our opinion, relate more to the sentiments, rather than to aspects.

In general, all approaches [3] to extracting aspects can be divided into four groups:

- Methods using word frequency information, including tf.idf, C-value measures;
- Methods based on relationship aspects with sentiment-related words (for example, double propagation method).
- Methods using the machine learning algorithm (supervised and unsupervised). These include Neural Networks, Conditional Random Fields, Support Vector Machine.
- Topic Modeling (Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Additive Regularization of Topic Models).

The paper [4] describes the aspect extraction system based on a conditional random field algorithm. For machine learning was used set of morphological features: word, POS, lemma. This system showed good results, especially precision metric.

Researchers in the paper [5] use a hybrid approach for aspect extraction. They first define an improved CNN architecture for aspect extraction which achieves comparable results against the current state-of-the-art systems. Then they combine the proposed improved CNN with an SVM that uses the manually engineered features.

In the article [6], the authors propose a generative model for extracting aspects based on the LDA. It is based on the idea that a consumer in his recall mentions aspects of both a certain group of goods and aspects of a generalized category of goods. For example, for the notebook category, electronics will be the generic parent category. Laptops have their own specific aspects, but they will also have common aspects that will belong to the category of electronics in general (for example, not only laptops have a screen, but also smartphones, tablets, computers). That is, aspects of any product are a mixture of aspects from their parent category and aspects unique to themselves.

First the authors attach each product to the nearest category in the category hierarchy. For each sentence with manual annotations of aspects, a model that uses category hierarchy information is used to find the topic that is closest to this aspect word.

Then they select 3 words from the sentence with the highest probability under the detected topic, because they are the best words to describe the topic. So topic models built in this way can successfully balances the aspects of a product itself and its parent category.

The paper [7] presents the Ontology-Based Product Sentiment Summarization framework for complex solution of two tasks such as aspect extraction and polarity detection. To improve performance, the authors propose the use of ontology to reinforce aspect extraction process by identifying features which relate to implicit entities, and reduce the errors of Sentiment analysis based on lexicons, which, in turn, improves the quality of analysis. They propose to use ontologies to go beyond the Sentiment analysis at the level of words and move to the level of concepts. Ontologies here act not just as a lexicon, but as a semantic knowledge base.

One of the most challenging problems in aspect based sentiment analysis is cross-domain aspect extraction. This problem is of practical importance, since good training sets are available only for certain review domains, while aspects in different categories are wildly varying. Cross-domain aspect extraction is studied for more than decade, in the paper [8] the authors proposed an approach based upon the use of Conditional Random Fields. In more recent papers alternative approaches such as an Interactive Attention Transfer Network (IATN) model [9] and Selective Adversarial Learning (SAL) method [10] based on LongShort Term Memory neural networks, and lexicon-based DomSent [11] were suggested. Nevertheless, cross-domain aspect extraction for Russian language reviews is less studied due to the lack of annotated data.

III. Data and preparation

For the study in this paper, we use the set of reviews for products placed on the Internet resource Yandex.Market. For automatic extraction of reviews, the program in Python, that uses the Yandex.Market API, was developed. Reviews were extracted in json format; meta-information was saved for further analysis. Thus, we've constructed a corpus of 41913 reviews (4739010 word usages) on 28 categories of products.

For representativeness, we have chosen a variety of product groups that are unlike each other. We have selected several groups of products from the categories: household appliances, electronics, health and beauty goods, goods for children, pet goods, goods for hobbies and leisure.

The distribution of the number of reviews by categories is presented in Table I.

TABLE I. THE DISTRIBUTION OF THE NUMBER OF REVIEWS BY CATEGORIES

Category	Number of reviews
Vacuum cleaning robots	2089
Smart watches and bracelets	2060
Electric teapots	2050
Pet food	2045
TV	2020
Cameras	2020
Mobile phones	2020
Electronic books	2020
Refrigerators	2020
Headphones & Bluetooth Headsets	2010
Washing machines	2010
Coffee machines	2010
Baby strollers	2010
Laptops	2010
Shampoos for hair	2010
Universal external batteries	2010
Bicycles for adults and children	1720
Curling irons and straighteners	1510
Child car seats	1440
Gaming consoles	1390
Steamers	1050
Mascara	590
Flea and tick remedies for cats and dogs	440
Exercise bikes	400
Quadcopters	400
Proteins for Athletes	339
Baby drinkers	140
Electric toothbrushes	80
Total	**41913**

As discussed above, our task is to elaborate a method for automatic extraction of aspects from reviews of various product categories. Before beginning our experiments, we have implemented the following steps:

1) We split the texts of reviews first into sentences, and then into words using UDPipe [12], pre-trained on the Syntagrus dependency treebank [13]. Words are delimited by whitespace characters. Punctuation marks are tokenized as separate tokens (words). We chose UDPipe because it performed well in CoNLL Shared task in recent years [14][15].

2) We got sentences that consist of one or more word lines, and word lines were then automatically annotated in CoNLL-U format [16].

3) Also, each token was assigned Part-of-speech tag using morphological analyzer pymorphy2 [17].

4) Besides, for our future model, we wanted to use the information about the meaning and generalized contexts of words. Therefore, we decided to apply vector model.

At the first stage, we have converted a text data array into word vectors. To represent words as vectors, we use an open source tool, Word2Vec [18]. Each word in the document is represented by a multidimensional vector containing semantic information about the word in the document. We have trained the model on our corpus using the Gensim [19] library in python, and as a result we've got words embeddings. For each word, we can find the words that are closest in meaning to our vector.

The Fig. 1 shows lists of the most similar words for words krasivyy (beautiful), shirokiy (wide), chaynik (teapot).

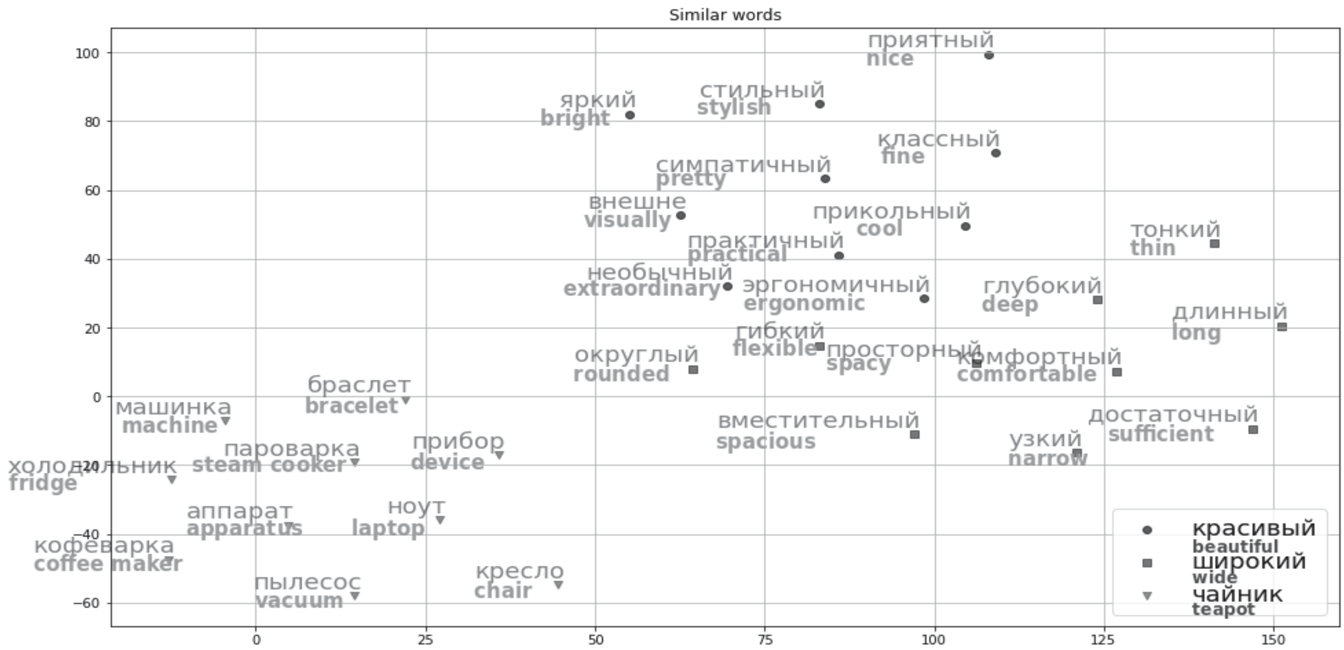

Fig. 1. The most similar words for words krasivyy (beautiful) shirokiy (wide), chaynik (teapot)

The most similar words for word krasivyy (beautiful) are shown in Table II.

TABLE II 10 CLOSEST WORDS TO THE WORD KRASIVYY (BEAUTIFUL)

Word	Translation	Similarity
'стильный'	*'stylish'*	*0.878*
'симпатичный'	*'pretty'*	*0.836*
'необычный'	*'extraordinary'*	*0.782*
'приятный'	*'nice'*	*0.766*
'яркий'	*'bright'*	*0.740*
'классный'	*'fine'*	*0.736*
'практичный'	*'practical'*	*0.725*
'прикольный'	*'cool'*	*0.719*
'внешне'	*'visually'*	*0.709*
'эргономичный'	*'ergonomic'*	*0.705*

The most similar words for word shirokiy (wide) are shown in Table III.

TABLE III 10 CLOSEST WORDS TO THE WORD SHIROKIY (WIDE)

Word	Translation	Similarity
'узкий'	*'narrow'*	*0.773*
'просторный'	*'spacy'*	*0.749*
'глубокий'	*'deep'*	*0.745*
'длинный'	*'long'*	*0.708*
'комфортный'	*'comfortable'*	*0.694*
'гибкий'	*'flexible'*	*0.684*
'достаточный'	*'sufficient'*	*0.665*
'тонкий'	*'thin'*	*0.650*
'вместительный'	*'spacious'*	*0.647*
'округлый'	*'rounded'*	*0.646*

The most similar words for word chaynik (teapot)are shown in Table IV.

TABLE IV 10 CLOSEST WORDS TO THE WORD CHAYNIK (TEAPOT).

Word	Translation	Similarity
'холодильник'	*'fridge'*	*0.808*
'машинка'	*'machine'*	*0.737*
'пылесос'	*'vacuum'*	*0.731*
'пароварка'	*'steam cooker'*	*0.705*
'ноут'	*'laptop'*	*0.699*
'прибор'	*'device'*	*0.681*
'браслет'	*'bracelet'*	*0.671*
'кофеварка'	*'coffee maker'*	*0.650*
'аппарат'	*'apparatus'*	*0.640*
'кресло'	*'chair'*	*0.639*

5) At the next stage, we clustered the word vectors using the machine learning algorithm K-means.

The k-means algorithm splits the initial set of objects into k clusters so that the averages in the cluster differ as much as possible from each other.

To determine the optimal number of clusters, we used the elbow method. According to this method, we should choose as the value of k the point after which the inertia (i.e. the sum of squared distances of samples to their closest cluster centers) start decreasing in a linear fashion. Determination of the optimal number of clusters is presented on the Fig. 2.

As can be seen from the figure, the graph becomes linearly decreasing for K greater than 40. Therefore, we have decided to divide into 50 and 70 clusters and choose the best option when testing the model.

6) We have also used information about the frequency of the word. For this, we have compiled a frequency dictionary of our corps. This allows us to use the word itself as a feature.

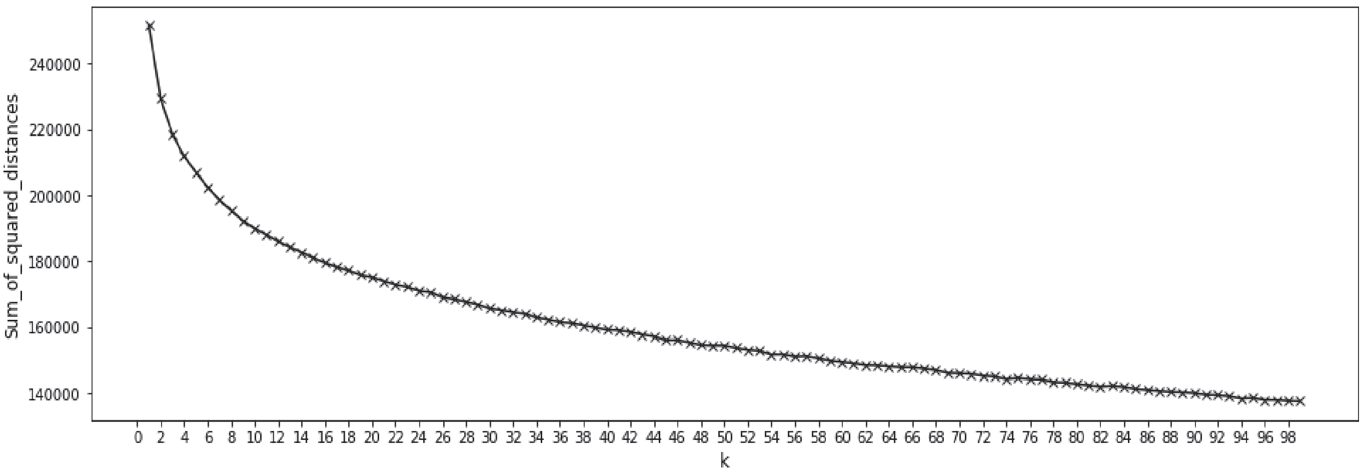

Fig. 2. The elbow method for optimal K

IV. Experiments and Results

For the task of automatic aspects extraction, we have chosen a well-known supervised machine learning algorithm – the Decision tree algorithm from the Scikit-learn library [20].

We have chosen this method because of its simplicity and also because it demonstrates good results in the task of classification. The classification task in this case is to determine for each word whether this word is an aspect or not.

Decision tree is a model that is a set of rules for adopting decisions. Graphically, it can be represented in the form of a tree structure, where the decision making moments correspond to the nodes. In the nodes, the process is branching, i.e. dividing it into branches depending on the choice made.

As features for the decision tree, we used the following features:

- Morphological tag (POS);
- Syntactic tag (DEP);
- Index of cluster;
- Frequency of the word.

We determined the depth of the tree using cross validation.

The best results were obtained with parameter:

- max_deph=20.
- max_features=0.95.

During the experiment, it turned out that the most important features are:

- frequency of the word,
- POS=NOUN,
- index of cluster (number of clusters=50),
- index of cluster (number of clusters=70),
- POS=PROPN,
- DEP=obl (the obl relation is used for a nominal (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct).

As for the number of clusters, the best results were achieved with the number of clusters equal to 70

Feature importances that are calculated as the (normalized) total reduction of the criterion (in our case, entropy) brought by that feature are shown in Table V.

A. Experiment 1

For an evaluation of the quality of our algorithm we have used the dataset of reviews on cars for Russian aspect-oriented sentiment analysis from SentiRuEval 2015 [21].

We have trained the decision tree on the training collection consisting of 200 reviews and tested the quality of work on the test collection containing another 200 reviews.

We have selected precision, recall and F1-measure as metrics for the evaluation extraction aspect from reviews. The results presented in Table VII.

TABLE V. Feature importances

Feature	Importance
frequency	0,328
pos_NOUN	0,281
index of cluster	0,229
pos_PROPN	0,042
dep_obl	0,028
dep_amod	0,014
dep_nsubj	0,011
dep_conj	0,009
dep_obj	0,007
dep_nmod	0,007
frequency	0,328
pos_NOUN	0,281
index of cluster	0,229
pos_PROPN	0,042
dep_obl	0,028
dep_amod	0,014

TABLE VI. Quality metrics in Experiment 1

Metric	Value
Precision	0,8087
Recall	0,6522
F1-measure	0,7221

We decided to compare our approach with approaches SentiRuEval 2015 and baseline (for the Task A – explicit aspects extraction). The results are presented in Table VII.

As seen above, our approach outperforms the baseline and the best overall (w.r.t. to F1-measure) approach from SentiRuEval 2015 in Precision and it is slightly inferior in Recall, while demonstrating adequate performance in F1-measure.

B. Experiment 2

Since we are interested in aspects for all review categories, we need to evaluate our algorithm in the cross-domain aspect extraction task. Such an evaluation requires a training collection from one category and test collection from another category of goods. To estimate the quality of the algorithm we compare its performance with two baseline approaches based on Naïve Bayes and Support Vector Machine (SVM).

As a source category we use cars review dataset from SentiRuEval 2015. For target category we choose reviews on teapots from our Yandex.Market dataset, because this category is very different from cars.

TABLE VII. RATING ASPECT EXTRACTION PERFORMANCE

№	Approach	Precision	Recall	F1-measure
Baseline		**0,7449**	**0,6724**	**0,7068**
1_3	Deep Recurrent Neural Networks	0,7917	0,7581	0,7581
2_1	The method based on sequential classification of tokens with SVM	0,8561	0,7422	0,7422
1_1	Deep Recurrent Neural Networks	0,7889	0,7287	0,7286
1_2	Deep Recurrent Neural Networks	0,7889	0,7287	0,7287
Our method		**0,8087**	**0,6522**	**0,7221**
4_1	Distributed representation-based approach	0,7417	0,7142	0,7142
7_2	Conditional random field algorithm [4]	0,7908	0,7126	0,7126
7_1	Conditional random field algorithm [4]	0,7970	0,6877	0,6877
8_1	Conditional random field algorithm	0,6609	0,6192	0,6192
5_1	unpublished	0,6879	0,5653	0,5653
5_2	unpublished	0,6879	0,5653	0,5653

As seen above, our approach outperforms the baseline and the best overall (w.r.t. to F1-measure) approach from SentiRuEval 2015 in Precision and it is slightly inferior in Recall, while demonstrating adequate performance in F1-measure.

C. *Experiment 2*

Since we are interested in aspects for all review categories, we need to evaluate our algorithm in the cross-domain aspect extraction task. Such an evaluation requires a training collection from one category and test collection from another category of goods. To estimate the quality of the algorithm we compare its performance with two baseline approaches based on Naïve Bayes and Support Vector Machine (SVM).

As a source category we use cars review dataset from SentiRuEval 2015. For target category we choose reviews on teapots from our Yandex.Market dataset, because this category is very different from cars.

We manually marked out 100 reviews on teapots. We marked only explicit aspects, i.e. words and phrases that clearly mention target aspects and do not contain their assessment. For example, in the teapot category such aspects are ‘volume’, ‘boiling speed’, ‘length of power cord’). The words expressing tonality, opinion or assessment are not considered as aspect terms.

We trained Decision tree with the same parameters, only slightly reducing the depth of the tree. We trained Naïve Bayes and SVM with frequency and POS tags features. The results are presented below in Table VIII.

TABLE VIII. QUALITY METRICS IN EXPERIMENT 2

Approach	Precision	Recall	F1-measure
Naïve Bayes with frequency and POS tags features	0.5475	0.6808	0.6027
SVM with frequency and POS tags features	0.6514	0.2871	0.3986
Our method	**0,7604**	**0,5145**	**0,6137**

It can be seen that our method outperforms the baseline algorithms, but the precision and recall are decreased comparing to results in one category. But still they are high enough for automatic aspect extraction in various categories for a further goal of linguistic study of such aspects, even though we did not carry out any adaptation to the new subject area. It seems possible to further improve the result for the task of the cross-domain aspect extraction.

V CONCLUSION

Thus, Internet review is a special type of text containing an author’s assessment of a product or service, posted on the Internet with the aim of exchanging information between real and potential consumers and manufacturers.

Online reviews have a certain set of parameters. Aspects are among the most significant parameters. This article is the first step towards a profound examination of the aspects of product reviews.

We offer an approach to automatic aspects extraction using the Decision Tree algorithm. This method shows good accuracy and completeness for one subject area, as well as when transferring to another.

In the future, we plan to improve the algorithm for aspect extracting, as well as to analyze and systematize aspects in reviews.

REFERENCES

[1] Local Consumer Review Survey 2019. Web: https://www.brightlocal.com/research/local-consumer-review-survey
[2] Loukachevitch, N. et al.: “SentiRuEval: Testing object-oriented sentiment analysis systems in Russian”, in Proceedings of international conference Dialog., 2015, pp. 3–13
[3] Bol'shakova E. I., Vorontsov K. V., Efremova N. E., Klyshinskiy E. S., Lukashevich N. V., Sapin A. S. “Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh” [Automatic natural language text processing and data analysis]. Moscow, HSE Publishing House, 2017. 269 p
[4] Rubtsova Yu., Koshelnikov S.A. “Extraction of aspects of goods and services from consumer’s reviews using Conditional Fields model”, Russian Digital Libraries Journal, 2015, pp 203-221
[5] Senarath, Yasas, Nadheesh Jihan, and Surangika Ranathunga. "A Hybrid Approach for Aspect Extraction from Customer Reviews.", ICTer 12.1, 2019, pp
[6] Yang Y. et al. “Aspect extraction from product reviews using category hierarchy information”, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. , 2017, pp 675-680
[7] Marstawi A. et al. “Ontology-based aspect extraction for an improved sentiment analysis in summarization of product reviews”, in:

Proceedings of the 8th International Conference on Computer Modeling and Simulation, 2017, pp 100-104
[8] Jakob, Niklas, and Iryna Gurevych. "Extracting opinion targets in a single-and cross-domain setting with conditional random fields." Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2010, pp. 1035-1045
[9] Zhang, Kai, et al. "Interactive attention transfer network for cross-domain sentiment classification." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, 2019, pp. 5573-5780
[10] Li, Zheng, et al. "Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning." arXiv preprint arXiv:1910.14192, 2019
[11] Chauhan, Ganpat Singh, and Yogesh Kumar Meena. "DomSent: Domain-Specific Aspect Term Extraction in Aspect-Based Sentiment Analysis." Smart Systems and IoT: Innovations in Computing. Springer, Singapore, 2020, pp. 103-109.
[12] Straka, M., Hajic, J. and Straková, J.: "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing", in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 2016, pp. 4290-4297
[13] Boguslavsky, I.: "SynTagRus–a Deeply Annotated Corpus of Russian", Les émotions dans le discours-Emotions in Discourse, 2014, pp.367-380.
[14] CoNLL 2017 Shared Task.
Web: http://universaldependencies.org/conll17/results.html
[15] CoNLL 2018 Shared Task.
Web: http://universaldependencies.org/conll18/results.html
[16] CoNLL-U Format.
Web: https://universaldependencies.org/format.html
[17] Korobov, M.: "Morphological analyzer and generator for Russian and Ukrainian languages". In: International Conference on Analysis of Images, Social Networks and Texts, Yekaterinburg, 2015, pp. 320-332
[18] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781. 2013.
[19] Rehurek, R. and Sojka, P.: "Software framework for topic modelling with large corpora". in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.
[20] Decision Trees.
Web: https://scikit-learn.org/stable/modules/tree
[21] SentiRuEval-2015.
Web:https://drive.google.com/open?id=0B7y8Oyhu03y_fjNIeEo3UFZObTVDQXBrSkNxOVlPaVAxNTJPR1Rpd2U1WEktUVNkcjd3Wms