Method of Cross-Language Aspect-Oriented Analysis of Statements Using Categorization Model of Machine Learning

Tetiana Kovaliuk[1'[0000-0002-1383-1589]'], Tamara Tielysheva[1'[0000-0001-5254-3371]'] and Nataliya Kobets[2'[0000-0003-4266-9741]']

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prospekt Peremohy, Kyiv 03056, Ukraine
[2]Borys Grinchenko Kyiv University, 18/2 Bulvarno-Kudriavska Str, Kyiv, 04053, Ukraine
tetyana.kovalyuk@gmail.com, telyshevatamara@gmail.com, nmkobets@gmail.com

Abstract. Product reviews are the foremost source of information for customers and manufacturers to help them make appropriate purchasing and production decisions. Today, the Internet has become the largest source of consumer thought. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. In this paper, we present a study of aspect-based opinion mining using a lexicon-based approach and their adaptation to the processing of responses written in Ukrainian and English. This information helps to build systems to understand customer's feedback and plan business strategies accordingly. This also helps in predicting the chances of product failure. In this paper, it is explained how machine learning can be used for opinion mining. The research methods used in the work are based on data mining methods, Web mining, machine learning, and information retrieval. The stages of the method of cross-language aspect-oriented analysis of statements are presented. The cross-language categorization of characteristics of goods is considered. The algorithm describes the model learning in cross-language virtual contextual documents.

Keywords: analysis of opinion, review, aspect, opinion orientation, sentiment analysis, categorization, machine learning

1 Introduction

The intellectual analysis of statements (opinion mining), which is in the extraction of subjective information (opinions, evaluative judgments, emotions, feelings, etc.) from text information becomes very important due to the development of information technologies and its implementation in all spheres of life. Identifying and evaluating the positivity or negativity of expressions regarding a particular research object can be applied to a variety of industries, including industry, marketing, education, etc. The

practical application of aspect-oriented analysis of statements is possible in content analysis as a formalized method of text analysis. Analysis of the tonality of the text allows you to evaluate the success of the advertising campaign, political and economic reforms; to determine the attitude of the press and the media to a particular person or event; to determine consumer attitude to certain products or services. Market research shows that online reviews have a significant impact on the behavior of the level of products sales [1]. However, their growing volume leads to the fact that it becomes impossible for consumers to get acquainted with each one. On the other hand, online reviews provide manufacturers with information on whether consumers are satisfied with their products. The manufacturer collects various attributes such as comments, wall post as raw data and use advanced data mining approaches for dispersal of intellectual knowledge. He also analyzes the data collected for decision making and product promoting [2].

Assessments of educational services users regarding the prestige of universities or the elitism of education in the applicant competition are becoming relevant in the field of higher education. Such estimations express the emotional perception of a product based on semantic parsing statements. Sentiment analysis and opinion mining is important for business and society due to the growth of social media such as reviews, forum, discussions, blogs, micro-blogs and social networks [3] [4]. Consequently, the task of determining the content and emotional color of consumer-related statements concerning aspects of goods (aspect-based opinion mining) in the evaluation system adapted to the Ukrainian market is relevant and important.

For analyzing user feedback, it is necessary to handle complex syntactical constructs of expressions, phrases that were used in a figurative sense, identify spam, noise, sarcasm etc. Therefore, the development of the latest information technologies in the area of opinion mining reduces to the following tasks:

- finding positive and negative statements in textual data;
- assigning of a certain numeric equivalent for positive or negative statements;
- summarizing positive and negative statements to a certain integral indicator in order to compare research objects.

There are aspect-based opinion mining methods that based on frequency-based analysis and use simple filters on noun constructs to extract aspects. Methods based on the syntactic structure of the text use natural language processing to find relationships between aspects and their related feelings. Hybrid methods use the natural language relation for filtering frequently encountered aspects. Accuracy of hybrid methods is much higher than the previous two. However, such as in the previous two cases, hybrid methods require manual adjustment of various parameters. To avoid the need of manually adjusting the parameters, they use educational methods with a teacher who automatically studies the parameters of the data model. Methods of education without a teacher, as well as probabilistic models, allow us to determine what is said in the text, the semantics of the text.

The tasks of the vocabulary analysis are divided into tasks of opinion mining at the level of the document (document-level), at the level of a separate sentence (sentence-

level) and opinion mining at the level of a separate phrase (phrase-level), which involves the analysis of individual characteristics of the product.

Aspect-oriented analysis of statements is widely used for practical applications. However, many scientists are working on improving the methods of analysis in such directions like identification of aspects in reviews, expression of emotions in relation to aspect, extraction of implicit attitudes towards aspects, identification of attitudes in comparative sentences, identification of aspects in multilingual systems. [5], [6], [7]

2 Stages of the cross-language aspect-oriented method of analyzing the statements

Each statement can be represented as the next five-dimensional vector [8]:

$$\left(e_j, a_{jk}, so_{ijkl}, h_i, t_l\right) \quad (1)$$

where e_j is j - essence for which the analysis of statements is performed;

a_{jk} is k - aspect of the essence e_j;

h_i is i - author of the statement;

t_l – time when the author h_i left his statement;

so_{ijkl} – the emotional direction of the statement left by the author h_i in relation to the aspect a_{jk} of the essence e_j in time t_l. May be positive, negative or neutral, may express different levels of intensity, for instance, from 1 to 5.

A couple e_j and a_{jk} (essence and aspect of the essence) always expresses the purpose of the statement.

The presence of indices emphasizes the correspondence of the five components of the expression (1) to each other. Any discrepancy will lead to an error during the analysis of statements. Each of the five components in (1) is significant. The absence of any of them makes the analysis problematic. This definition covers most, but not all possible aspects of semantic analysis of statements, which can in fact be arbitrarily complex. In this regard, a five-dimensional vector can lead to loss of information. In this case, the five-dimensional vector is still used.

Definition (1) is the basis for transforming unstructured text into structured data. A five-dimensional vector can be the basis of a database schema according to which the extracted statements will be placed in its table. Then qualitative, quantitative analysis and analysis of the expression's trends can be made using the capabilities of database management systems and OLAP tools.

Definition of the notion of utterance given in this paper is sharpened by more than regular expressions. Another type is a comparative statement that requires a different definition. As an input, a collection of user reviews written in English and Ukrainian, and a cross-language categorization model Φ [9]. The point of the method's stages is:

1. To categorize all aspects of the product that are found in the reviews in English and Ukrainian (referred to as "Multilingual") in semantic aspects.
2. To extract pairs of "aspect-expression" from multi-language reviews and grouping into aspect-oriented sets of statements. The association of product aspects and expressions will be carried out according to their mutual position in the text of the review. Through linguistic analysis of text and specific rules words are defined that indicate the author's attitude and are closest (within certain limits) to the term, which refers aspect of the product. The extracted statement is associated with the term aspect. Then the polarity of the expression is determined and is associated with the semantic aspect, to which the current term aspect refers. Determination of the power of emotionality of expressions that relate to the aspect of a product is made by summing up all the extracted statements of this aspect.
3. To summarize the cross-language differences in expressions for various aspects, for instance, in the form of aspect ratings.

3 Cross-language latent semantic association

Each aspect of a product is usually indicated by a set of terms. Cross-language categorization of product aspects focuses on their categorization into a single semantic categorical structure.

Let X be a space of characteristics for representing instances of multi-language product characteristics, and Y is a set of labels for semantic categories. Let $p_s(x, y)$ be predicted semantic distribution of categories and $p_t(x, y)$ be genuine semantic distribution of categories, according to which the pair (x, y) determines the relation of object X to category Y. It is expected that $p_s(x, y)$ will approximate $p_t(x, y)$ better without using any labeled data.

Cross-language categorization of product characteristics, which is based on lexical comparison, is not capable of determining the basic semantic distribution of various multi-language characteristics [10]. Many terms that means the same aspects are not similar on the lexical level. Such hidden semantic associations between words provide an opportunity to determine the basic semantic distribution in the domain.

Therefore, for further research, the model Φ is used to define cross-language latent semantic associations between multilingual terms that means aspects of the product. This model learns on unlabeled text of user statements. In the learning process, a multivariate key vector characterizes each aspect of the product.

Characteristics of semantic associations in the model are hidden random variables derived from the data. Obviously, the model Φ can better define cross-language latent semantic associations between aspects of goods. It is possible to better approximate the actual distributions of semantic categories $p_t(y|x; M)$ using the model without the need of using labeled data.

4 Model training on cross-language contextual virtual documents

4.1 Cross-language contextual virtual document

In order to determine the hidden relationships between multilingual terms, each term of the aspect of a product is characterized by a cross-language contextual virtual document.

The term of the product aspect pf is given, cvd_{pf} is cross-language contextual virtual document, which consists of such multidimensional hidden semantic keys:

- the current term pf ;
- the term pf^T which is an automatic translation of term pf ;
- the set of components pf and pf^T, which are labeled as W_{pf} and W_{pf^T} ;
- hidden semantic themes of components pf and pf^T, which are labeled as S_{pf} and S_{pf^T} at the word-level;
- monolingual latent semantics pf of product aspects, which are labeled as MFS_{pf} .

Therefore, contextual virtual document is a set:

$$cvd_{pf} = \left\{ pf, pf^T, W_{pf}, W_{pf^T}, S_{pf}, S_{pf^T}, MFS_{pf} \right\} \quad (2)$$

For example term pf = «screen resolution». Table 1 provides a cross-language context-sensitive virtual document cvd_{pf} («screen resolution»), extracted from English and Ukrainian review texts.

Table 1. Components of a cross-language contextual virtual document cvd_{pf} = «screen resolution»

Keys	**Contextual virtual document cvd_{pf} («screen resolution»)**
pf	screen resolution (English)
pf^T	роздільна здатність екрану (Ukrainian)
W_{pf}	«screen», «resolution»
W_{pf^T}	«роздільна», «здатність», «екрану» (Ukrainian)
S_{pf}	S(«screen»), S(«resolution»)
S_{pf^T}	S(«роздільна»), S(«здатність»), S(«екрану»)
MFS_{pf}	MFS («screen resolution»)

In the construction of a cross-language virtual contextual document, they generate monolingual hidden semantic themes on equal aspects of the product and words, using the algorithms presented in [11].

Component words are grouped in the set of hidden themes, according to their context in a monolingual collection (corpus). A monolingual hidden semantic theme at the level of product aspects is created in accordance with their hidden semantic structure and contextual passages in the corresponding collection. A complete machine translation document is usually used to define semantic associations between aspects written in different languages. In order to reduce the noise that occurs in machine translation, the cross-language virtual context document only uses the translation of the individual term of the product aspect instead of the translation of the full text of the review.

Contextual virtual document cvd_{pf} usually describes the multidimensional cross-language hidden semantic aspects pf in the reviews. A vector is constructed for pf with all reviewed features from cvd_{pf} in the model:

$$Vector(cvd_{pf}) = \{x_1, \ldots, x_j, \ldots, x_m\} \quad (3)$$

where x_j - describes j context related feature associated with pf, m – total number of features in cvd_{pf}.

Weight of each contextual feature x_j in cvd_{pf} is calculated by PMI index (pointwise mutual information) between x_j and pf [4]:

$$PMI(x_j, pf) = \log_2 \frac{P(x_j, pf)}{P(x_j) \cdot P(pf)} \quad (4)$$

where $P(x_j, pf)$ – the probability that pf and x_j will be met in the text next to each other;

$P(x_j)$ – the probability that x_j will appear in the text;

$P(pf)$ – the probability that pf will appear in the text.

The weight is normalized as an integral part of the logarithmic function.

4.2 Model training

The Machine Learning provides a solution to the classification problem that involves two steps: learning the model from a corpus of training data, classifying the unseen data based on the trained model [13]. This model can in fact be considered as a general probabilistic topic model. It can be trained with non-tagged reviews using hidden thematic models, such as the latent placement of Dirichlet (Latent Dirichlet Allocation – LDA) [14] and probabilistic hidden semantic indexation (Probabilistic Latent Semantic Indexing – PLSI) [15]. Thematic models are models of text document collections that determine which topics each collection document refers to.

The LDA is a generative model that allows you to interpret the results of observations with implicit groups, which allows you to get an explanation of why some parts of the data are similar. The algorithm for constructing a thematic model receives a collection of text documents at the input. The output for each document is a numeric vector, which consists of assessing the degree of belonging of this document to each topic. The size of this vector is equal to the number of topics and can be set at the input of the model or determined by the model automatically.

Let us consider the algorithm of training given model.

Input data:

- R_{l_1} is collection of user reviews written in language l_1;
- R_{l_2} is collection of user reviews written in language l_2;
- $PF\,Set$ represents all titles of the aspects that are encountered in R_{l_1} and R_{l_2};
- monolingual latent thematic models $\theta_{wd}^{l_1}$ and $\theta_{wd}^{l_2}$ at the word level *wd* written in languages l_1 and l_2;
- monolingual latent thematic models $\theta_{wp}^{l_1}$ and $\theta_{wp}^{l_2}$ .at the aspect-level of products *wp*.

Output data: Cross-language aspect-categorization model Φ.

The scheme of the algorithm consists of such steps.

Initialization: Cross-language set of contextual documents $cvd\,Set = \varnothing$.

Step 1. For each term pf_i, which belongs to the set of all terms $PF\,Set$,. $pf_i \in PF\,Set$ do the following:

Step 1.1. Perform a machine translation of the term pf_i and determine pf_i^T: $pf_i^T = MT(pf_i)$.

Step 1.2. Define language l_s of the original aspect pf_i: $l_s \leftarrow Language(pf_i)$.

Step 1.3. Define language l_t of automatically translated aspect pf_i^T: $l_t \leftarrow Language\left(pf_i^T\right)$.

Step 1.4. Define the vector of component words for a term pf_i: $W_{pf_i} = GetComponentWords(pf_i)$.

Step 1.5. Define the vector of component words for the translated term pf_i^T: $W_{pf_i^T} = GetComponentWords\left(pf_i^T\right)$.

Step 1.6. For each component word w_j, which belongs to the vector W_{pf_i} ($w_j \in W_{pf_i}$) do the following:

Step 1.6.1. Generate latent theme S_{w_j} for the component word w_j using the model $\theta_{wd}^{l_s}$: $S_{w_j} = TP\left(w_j, \theta_{wd}^{l_s}\right)$.

Step 1.6.2. Add to the set S_{pf_j} of hidden semantic themes of components pf_j at the word-level latent theme S_{w_j} received in step 1.6.1: $AddTo\left(S_{w_j}, S_{pf_j}\right)$.

Step 1.7. For each component word w_k, which belongs to the vector $W_{pf_i^T}$ ($w_k \in W_{pf_t^T}$) do the following:

Step 1.7.1. Generate latent theme S_{w_k} for the component word w_k using the model $\theta_{wd}^{l_t}$: $S_{w_k} = TP\left(w_k, \theta_{wd}^{l_t}\right)$.

Step 1.7.2. Add to the set $S_{pf_j{}^T}$ of hidden semantic themes of components $pf_j{}^T$ at the word-level latent theme S_{w_k} received in step 1.7.1: $AddTo\left(S_{w_k}, S_{pf_j{}^T}\right)$.

Step 1.8. Generate monolingual latent semantics for pf_j at the aspect-level of the product using the model $\theta_{mp}^{l_s}$: $MFS_{pf_i} = TP(pf_i, \theta_{mp}^{l_s})$.

Step 1.9. Provide the values for components of the cross-language virtual context document: $cvd_{pf_i} = \left\{pf_i, pf_i{}^T, W_{pf_i}, W_{pf_i^T}, S_{pf_i}, S_{pf_i^T}, MFS_{pf_i}\right\}$.

Step 1.10. Add to a set of cross-language virtual contextual documents $CVD\,Set$ current document cvd_{pf_i}: $AddTo\left(cvd_{pf_i}, CVD\,Set\right)$.

Step 1.11. Generate model Φ with Dirichlet distribution on set $CVD\,Set$.

Consider a more detailed training process of model Φ type LDA on cross-language virtual semantic contextual documents.

Non-tagged collections of reviews R_{l_1} and R_{l_2}, which are written in languages l_1 and l_2 are given. Will consider the terms of goods aspects written in the language l. In the construction, cvd_{pf} latent topics from component words are generated using a monolingual word-level model θ_{wd}^{l}. The monolingual latent semantic MFS_{pf} of each product aspect is generated using a monolingual thematic model θ_{mp}^{l} of the term-aspects level. The weight of each element cvd_{pf} is calculated using the PMI index by the formula (1). Next, the studied model Φ with Dirichlet distribution generates a set of cross-language virtual context documents. In experiments conducted within the framework of this article coefficient $\alpha = 0.1$ and the number of iterations was 1000. The given modeling algorithm describes in detail the complete training process, where:

- the function $MT(pf_i)$ means the result of the automatic translation of the term pf_i;
- the function $TP(data,\theta)$ generates a latent theme for an argument $data$ using a latent thematic model θ;
- θ_{wd}^{l} describes a monolingual thematic model at the word-level for a given language l;
- θ_{mp}^{l} describes a monolingual thematic model of the product aspects for a given language l.

The investigated model studies the a posteriori probability of decomposing multilingual aspects of terms and their virtual contextual documents in the subject. It expands the traditional "bag of words" thematic models into a context-dependent, cross-language concept associative model.

5 Experimental studies

Input data is collected from user reviews of mobile phones and laptops in English and Ukrainian. Reviews are accumulated on popular websites designed to consolidate custom product reviews [16], [17]. All multilingual designations of product aspects are automatically removed from the data obtained using the statistical method [11]. For pre-processing data, Maximum Entropy part-of-speech (POS) tagger uses the maximum entropy for generating POS markup for data in English. For the data in Ukrainian, a hidden Markov model is used to generate POS-markup.

While carrying out experiments, the categorization of multilingual titles of the each aspect of the subject areas (mobile phones and laptops) in semantic aspects was performed and a cross-language aspect-oriented analysis of statements was made.

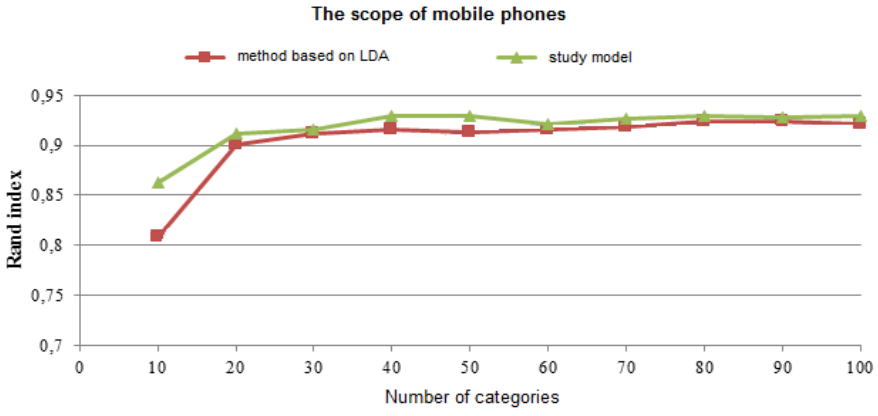

Fig. 1. Estimation of the cross-language categorization of aspects for mobile phones for different topics

Figure 1 shows the dependence of the Rand Index on the number of topics for two comparative methods: the investigated method and method based on the LDA. These methods effectively detect latent semantic associations in reviews.

Experimental results show that the studied model effectively group multivolume titles of aspects into semantic categories.

6 Conclusions

Aspect-oriented analysis is the most detailed among the all levels of the analysis of statements and is necessary for most practical applications. In this article, the mathematical formulation of aspect-oriented expression problem and the cross-language latent semantic association are considered, the characteristic of the product aspect under the cross-language virtual contextual document and the model learning process is reviewed. Method of aspect-oriented analysis based on the categorization model and the LDA, is trained in virtual contextual documents. Experimental results show that the studied model effectively groups multivolume names of aspects into semantic categories.

References

1. Lipsman, A.: Online consumer-generated reviews have significant impact on offline purchase behavior. In: Technical report, Comscore Inc., (2007) http://www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior
2. Rahman, M. M.: Mining Social Data to Extract Intellectual Knowledge. In: International Journal of Intelligent Systems and Applications (IJISA), vol.4, no.10, pp.15 -24 (2012)
3. Kumar, A., Sebastian T.M.: Sentiment Analysis: A Perspective on its Past, Present and Future. In: International Journal of Intelligent Systems and Applications (IJISA), vol. 4, no. 10, pp.1-14 (2017)
4. Liu, B.: Sentiment analysis and subjectivity. Handbook of Natural Language Processing, second edition, New-York, Now Publishers Inc. (2010)
5. Wogenstein, F., Drescher, J., Reinel, D., Rill, S., Scheidt, J.: Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM New York, NY, USA, (2013)
6. Samha, A.K., Li, Y., Zhang, J.: Aspect-Based Opinion Mining from Product Reviews Using Conditional Random Fields. In: Proceedings of the 13-th Australasian Data Mining Conference (AusDM 2015), Sydney, Australia. pp. 118-128 (2015)
7. Dragoni, M., Da Costa Pereira, C., Tettamanzi, A.G.B., Villata, S.: Combining Argumentation and Aspect-Based Opinion Mining: The SMACk System. In: AI Communications, IOS Press, No 31 (1), pp. 75-95 (2018)
8. Moghaddam, S.: Aspect-based Opinion Mining in Online Product Reviews Burnaby, Simon Fraser University (2013)
9. Dodonov, A.G., Lande, D.V., Putyatin, V.G.: Computer networks and analytical studies, Kiev, IPRI NAS of Ukraine, p. 486 (2014)

10. Riloff, E., Patwardhan, S., Wiebe, J.: Feature Subsumption for Opinion Analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 440–448, Sydney (2006)
11. Wiebe, J., Bruce, R., O'Hara, T.: Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, pp. 246-253 (1999)
12. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424 (2002)
13. Narendra, B., Uday Sai, K., Rajesh, G., Hemanth, K., Chaitanya Teja, M. V., Deva Kumar, K.: Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies. In: International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.8, pp.66-70 (2016)
14. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. In: Journal of Machine Learning Research, No. 3:pp. 993–1022 (2003)
15. Kim, J., Li, J., Lee, J.: Evaluating multilanguage-comparability of subjectivity analysis systems. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 595–603 (2010)
16. Custom Reviews. Shopify website and collects customer feedback. [Electronic resource]. Access mode: https://apps.shopify.com/custom-reviews
17. Mobile phone reviews – GSMArena.com. [Electronic resource]. Access mode: https://www.gsmarena.com/reviews.php3