UDC 64.033:334.72

LULA Paweł, Doctor of Economics, Professor, Dean of Management Faculty of Cracow University of Economics

WÓJCIK Katarzyna, Assistant at Department of Computational Systems of Cracow University of Economics

OPINION MINING AND ITS IMPORTANCE IN THE ACTIVITY OF CONTEMPORARY ENTERPRISES

The concept of knowledge management has been presented. Economic and non-economic objectives of knowledge management have been determined. Typology (classification) of knowledge management strategies has been conducted. Author 's vision of a comprehensive approach to knowledge management has been offered.

Keywords: knowledge, knowledge management system, concept, organization (enterprise), potential, knowledge resources, knowledge management strategy

Лула П., Войчик К. Анализ мнений потребителей и его значение в деятельности современного предприятия*. Охарактеризован процес извлечения / собирания мнений потребителей. Предложен метод анализа мнений на конкретном примере для емпирической проверки его действенности. Обозначены отличия между «анализом настроений» и «извлечением мнений». Для комплексного решения задач такого анализа предложено использовать модель LDA (скрытого распределения Дирихле).*

Ключевые слова: мнения потребителей, настроения потребителей, модель скрытого распределения Дирихле, автоматизация анализа мнений.

Introduction. Through the years the ways of collecting customer opinions has changed. The development of mass media and especially the Internet has influenced the form and availability of customer opinions. Customers not only search for opinions of others before buying a product or using a service but also tend to put their own opinions in the Internet. Often they are encouraged to leave their reviews on producer, vendor or intermediary service web site.

© Lula P., Wójcik K., 2015

Increasing number of consumers' opinions created the need of its automatic analysis. This issue is gaining popularity for both – researchers and entrepreneurs, for whom consumers' reviews are important source of business information. The application of automated tools for opinions analysis opened new possibilities of their usage.

In the paper firstly opinion mining will be defined and characterized. Then we will move on to description of proposed solution for opinions analysis that uses domain knowledge. Next part of work will be devoted to empirical example of previously presented method. We will sum up with conclusions drawn from the conducted research.

Opinion mining definition and characteristics. In the literature phrases "sentiment analysis" and "opinion mininig" are often used as synonyms. However the difference in its definitions and in the aim of each of them can be determined.

Sentiment is defined by Pang et al. [1] as overall opinion towards the subjects mater. So the sentiment analysis can be defined as actions aiming only to determine the attitude of speaker or writer to the subject of opinion (often this attitude is called a polarity).

In the same time opinion mining is defined by Dave et al. [2] as analysis:

- processing a set of search results for a given item;
- generating a list of product attributes (quality, features, etc.);
- aggregating opinion about them (poor, mixed, good).

Within the opinion mining three types of analysis can be conducted [3]:

- sentiment recognition – analysis of the general attitude to a product or service (positive, negative, neutral), also called sentiment analysis;
- feature-based analysis – identification and evaluation of main features of a given product;
- comparative analysis which allows to compare a given product or a given feature to others.

According to analysis of definition and aims presented above sentiment analysis is part of opinion mining. In the same time opinion mining covers a wide range of actions aiming to perform complete analysis of opinions (not only its polarity determination).

Description above makes opinion mining similar to information retrieval. However while information retrieval is defined as identification and analysis objective pieces of data, opinion mining (as well as sentiment analysis) can be defined as identification and analysis of subjective opinions, emotions and feelings in the text. Often the first step that must be done is determination if certain text is objective or subjective [4]. Then proper methods can be applied [5].

Nowadays analytics often don't have to ask customers for opinions. E-shops and shopping services are doing the job for them. Consumers also put their opinions willingly on blogs or different kinds of forums [2].

The accessibility of consumers opinions made entrepreneurs more interested in their content. Managers can acquire information how company's products or services are evaluated by its users. They can also gain detailed knowledge about the reasons why customers have positive or negative attitude to certain products or services. Data obtained from opinions can be used as a basis for future changes not only in product development but also in sources and ways of its advertisement.

Opinion mining has many advantages and is a very useful tool in business. In the same time there are some difficulties that have to be taken into account. Those difficulties can have technical character like:

- necessity to collect data from different sources;
- processing of large amount of data;
- different forms of opinions available in the Internet;
- opinions in different languages concerning the same product or service.

In the same time there are difficulties with analysis of the content of opinions resulting from the need of computational processing of textual data like:

- hidden character of emotions and sentiment (very often they are not expressed directly);
- sarcasm (irony) present in opinions;
- mistakes in writing or intentional changes like omitting national diacritics;
- language of the Internet used in opinions – emoticons, abbreviations, mixtures of letters and numbers when read phonetically sounds like certain words;
- repetitions of letters, words and phrases;
- co-references – two or more expressions refer to the same person or thing (mobile, phone, headphone, etc, ...);
- negations (dislike, do not like);
- entity recognition problem – identification of names of persons, organizations, locations, monetary values;
- comparisons (... better then...);
- polysemy (the word book can mean *reserve* or *manuscript*).

Another problem that is worthy consideration is existence of fake opinions written by competitior, special PR companies or even by employees (upon instructions or independently) or ex-employees. They can darken the results of analysis. To overcome possible negative consequences of difficulties in opinion mining some opinions' objectivization actions can be performed like:

- voting – taking into account a large number of opinions; it can be treated as a form of opinion aggregation;
- opinion evaluation – by other customers;
- evaluation of author's authority – publishing some details about opinion's author and evaluation of his/her authority by others.

Opinions that can be found in the Internet can be divided into different groups. Two division criteria can be pointed out: opinions form and the scope of knowledge. Taking opinions' form into account following groups can be distinguished:

- binary opinions (yes/no, like/dislike, good/bad);
- nominal values (about mobile phone: heavy, expensive, modern, ...);
- ordered values (bad/typical/good/excellent; Likert scale).

Text:

- structured opinions;
- unstructured opinions.

When the scope of knowledge is the division criteria following types of opinions are possible:

- without additional domain knowledge (based only on opinions);
- with additional domain knowledge (based on opinions and on knowledge about products or services).

The form of opinion is frequently connected with its source. For example opinions on forums and blogs usually have the form of unstructured text while opinions on different online shops, auctions online, opinions services etc. commonly are structured and supported with kind of binary, nominal or ordered values.

There are many different methods that can be applied to opinion mining:

- approach based on frequency matrix;
- probabilistic approach (topic modelling, probabilistic LSA);
- rule-based methods (used regular expressions);
- approach based on domain knowledge (ontology-based approach, logic models);
- summarization and keywords identification methods;
- classification methods – for sentiment classification;
- visualization methods;
- aggregation methods – used for increasing the level of objectivism.

The choice of the method must be influenced by the aim of analysis and by the form of opinions that is subjected to analysis. Different methods can give better results when analysis is conducted on certain type of opinions. Most methods require analysed opinions to be in one form.

Latent Dirichlet Allocation (LDA) is a model which is used for description of documents' contents as a mixture of homogenous topics. It was proposed in [6].

Using this approach the process of analysis is composed of two main steps:

1. Topics' identification.
2. Topics based documents reconstruction.

Relations between these two processes are presented in *figure 1*.

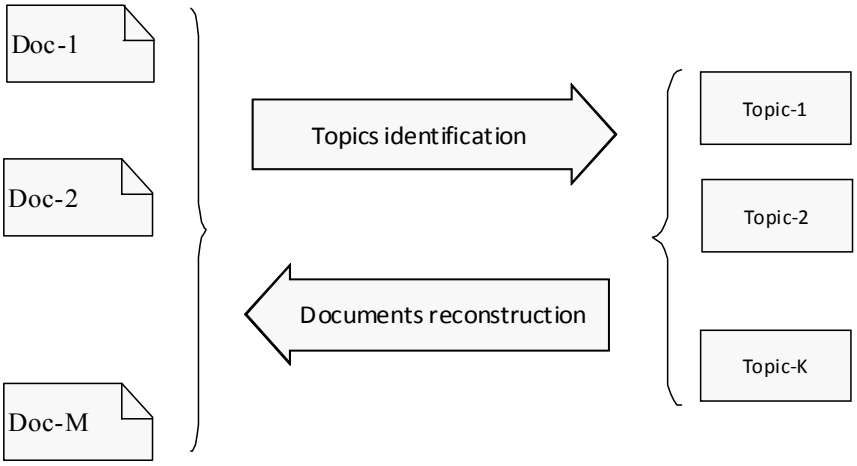

Figure 1. Main steps in Latent Dirichlet Allocation
(own elaboration)

Let's assume that WORDS is a set of all words which appear in the corpus (the collection of documents):

$$WORDS = \{word_1, word_2, ..., word_N\}.$$

Each document is represented by the vector in which the number of occurrence for every word is stored (bag of words representation). The information about occurrence can be also expressed in terms of probabilities of occurrence (dividing the number of word's occurrence by the number of all words).

The same approach can be used for topics definition. Each topic is defined as a vector of probabilities over words.

$$topic_i = [p_{i1} \quad p_{i2} \quad p_{i3} \quad \dots \quad p_{iN}],$$

where p_{ij} is a probability of occurrence in j-th word in the i-th topic.

The distribution of words within a given topic is defined by Dirichlet distribution. The main goal of the topics identification process can be stated as finding such topics (described by distribution over words) which can be used for reconstruction of original documents. The contribution of topics to the contents of a given document also is represented by the Dirichlet distribution. Unfortunately, topics discovered by the learning algorithm are hard to interpret. Pro process of topics identification and documents reconstruction is shown in the *figure*.

The proposal for solving a problem with interpretation of topics was presented by Ramage, Hall, Nallapati, Manning in 2009 and is called Labelled-LDA [7]. In this approach labels are manually assigned to every document. They are used for contents description and in further calculations

every label will be represented as one topic. It is possible to assign many labels for description of each document and the same label can be assign to many documents. Also in Labelled-LDA topics are defined by expressing the probability distribution over words. But the number of topics is equal to the number of unique labels and every topic represents one concept identified by one label. The schema of Labelled-LDA approach is shown in *figure 2*.

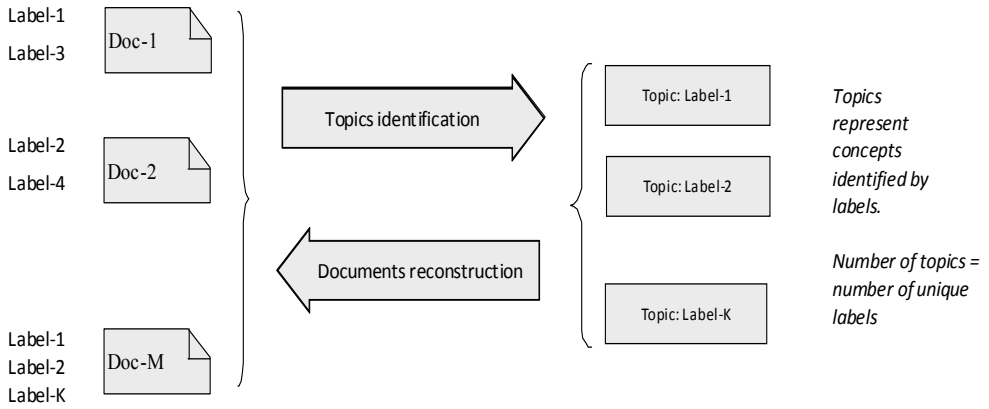

Figure 2. The idea of Labelled Latent Dirichlet Allocation
(own elaboration)

The possibility of topics' interpretation as concepts is the main advantage of Labelled-LDA method.

In the reconstruction phase the contribution of every topic is estimated by expressing the probability of topic's occurrence in the contents of a given document.

Empirical research. In the empirical research opinions about hotel rooms in London were used*. The feature-based analysis was performed. The scope of analysis was defined by a list of attributes related to the subject of research.

The list of attributes embraced several features of hotel rooms. For example: general evaluation, room size, staff evaluation, quality of bed, cleanliness, room equipment, room location, comfort, price, temperature and many others.

Opinions were manually labelled. For description of every room feature two labels were created which represented positive (feature name plus suffix *pos*) or negative (feature name plus suffix *neg*) evaluation of a given attribute. For example for staff evaluation two labels were used: *staff-pos* and *staff-neg*.

The set of 380 labelled opinions was used. The part of data set was presented in the *figure 3*.

* Source: http://kavita-ganesan.com/opinosis-opinion-dataset.

	A	B
1	LABELS	DESCRIPTIONS
2	staff-neg	We arrived at 23,30 hours and they could not recommend a restaurant so we decided to go to Tesco, with very limited choices but when you are hungry you do not careNext day they rang the bell at 8,00 hours to clean the room, not being very nice being waken up so earlyEvery day
3	size-pos	We had a room with two double beds which was surprisingly roomy, considering the small hotel rooms I have in previous trips to London .
4	staff-pos clean-pos bed-pos	The room was quiet, clean, the bed and pillows were comfortable, and the service was
5	readiness-pos	We arrived about 11 am, room was ready .
6	size-pos clean-pos	Room was good size for Europe , clean throughout .
7	staff-pos	The Concierge desk called our room to ask if we needed any information or assistance .
8	size-pos clean-pos bed-pos	Room was plenty big enough and clean and tidy, bed was comfordable .
9	equip-neg	First, we walked in and the restroom door was broken .
10	clean-pos	Our room was typical holiday inn the bathroom could have done with updating but was
11	readiness-neg	Our rooms were not ready, we were promised rooms at a later time, etc .
12	size-pos	My room was positively huge by European standards .

Figure 3. The dataset structure
(own elaboration)

The data set was divided into two parts: learning set (350 opinions) and testing set (30 opinions). For learning set the following stages of analysis were performed: stemming, usage of stop-list filter, building of Labelled-LDA model. For the last step the Stanford Topic Modeling Toolbox was used. All procedures were implemented in Scala language. Topics discovered during analysis represented main concepts which can be used for room evaluation. The final assessment can be expressed by calculating the probability of every topic's occurrence (concept) in the opinion.

For every concept (identified by a unique label) its importance was evaluated. The results of topic importance calculation was presented in the *figure 4*.

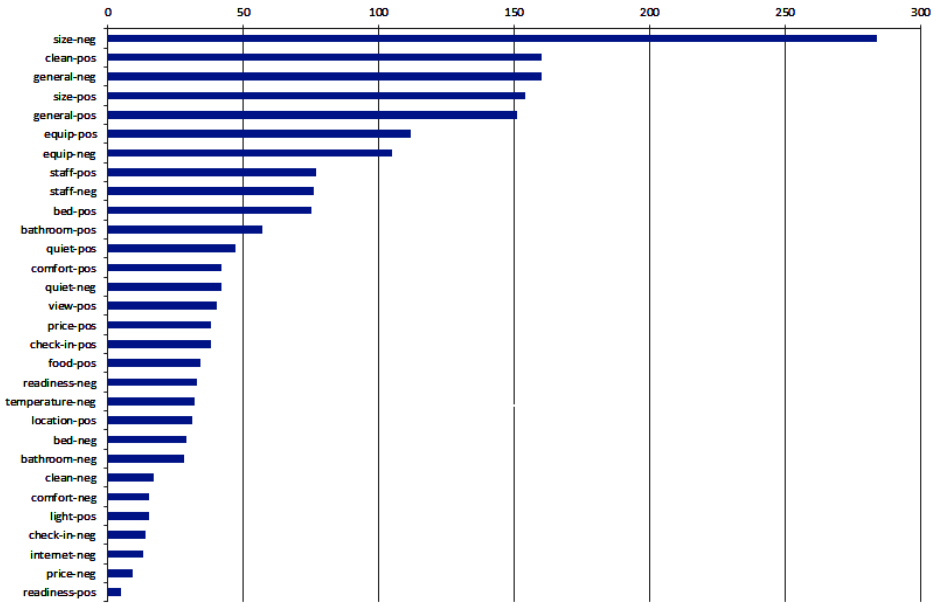

Figure 4. The importance of topics (concepts)
(own elaboration)

For every concept its description was estimated as a distribution over words. The *figure 5* shows an exemplary distribution for the concept *size-neg* (negative opinion about room's size).

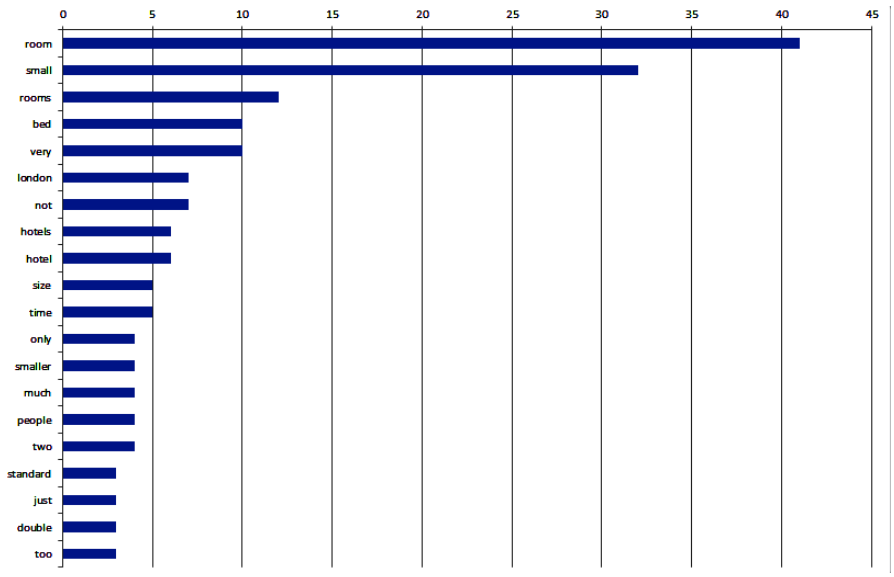

Figure 5. The distribution over words for *size-neg* topic
(own elaboration)

The model created as a result of the above procedure can be used for analysis of new, unseen before opinions.

Table

The results of usage of Labelled-LDA model for new opinions
(own elaboration)

Opinion	Labels assigned manually	Labels assigned by the model (with probabilities of occurrence in the document's contents)
The bathroom is a good size.	bathroom-pos	bathroom-pos (1.0)
The room was clean and, by London standards, decently sized.	clean-pos size-pos	size-pos (0.98), clean-pos (0.02)
When we tried to use a phone card from our room it would not work so I asked the front desk to help me and was told they couldn't really!	staff-neg	staff-neg (1.00)
The hotel room was very clean and the cleaning staff and breakfast staff were very attentive.	clean-pos staff-pos	staff-pos (0.7) clean-pos (0.3)

For evaluation of the quality of the built model two measures were calculated for testing set.

Precision (which is defined as the fraction of retrieved instances that are relevant) was equal to 0.94. And the recall (defined the fraction of relevant instances that are retrieved) was 0.98. These two measures indicate that the quality of the model is very high.

Conclusion. Thanks to the development of information technology consumers can express their opinions about products and services. Their views can be used as a valuable information source provided that the process of their analysis is automated. Unarguably the Latent Dirichlet Allocation seems to be a very useful tool for automatic opinion analysis, especially its supervised version called Labelled-LDA.

Main properties of Labelled-LDA prove that it is useful for feature-based opinion analysis. Unfortunately this approach requires manually tagged opinions for model building and therefore is very time-consuming.

REFERENCES

1. *Pang B.* Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, 10 / Pang B., Lee L. i Vaithyanathan S. — 2002. — P. 79–86. Stroudsburg, PA, USA: Association for Computational Linguistics. — Way of access : doi:10.3115/1118693.1118704.
2. *Dave K.* Mining the peanut gallery: Opinion extraction and semantic classification of product reviews / Dave K., Lawrence S. i Pennock D. M. : Proceedings of the 12th international conference on World Wide Web, 2003. — P. 519–528.
3. *Liu B.* Web DataMining. Exploring Hyperlinks, Contents, and Usage Data / Liu B. — Heidelberg : Springer-Verlag Berlin, 2007.
4. *Wibe J.* A corpus study of evaluative and speculative language / Wibe J., Bruce R., Bell M., Martin M. i Wilson T. : Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, 16, 2001. — P. 1–10.
5. *Pang B.* Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval / Pang B. i Lee L, 2008. — P. 1–135.
6. *Blei David M.* Latent Dirichlet allocation / *Blei* David M., Ng Andrew Y., Jordan Michael , Lafferty John, ed. // Journal of Machine Learning Research 3 (4–5). — P. 993–1022. — Way of access : doi:10.1162/jmlr.2003.3.4-5.993.
7. *Daniel Ramage.* Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora / Daniel Ramage, David Hall, Ramesh Nallapati, Christopher D. : Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing // Association for Computational Linguistics. — 2009. — Vol. 1. — P. 248–256.

Articles submitted to editors office of 10.11.2015.

Лула П., Войчик К. Аналіз думок споживачів та його значення у діяльності сучасного підприємства.

Постановка проблеми. *Протягом років шляхи накопичення даних споживачів щодо товарів (продуктів), які вони купують, змінилися. Зростання кількості думок споживачів про товари і послуги, що, зокрема, накопичуються в Інтернеті, обумовлюють необхідність автоматизації процесу їх аналізу.*

Мета і завдання дослідження, *представленого у статті, полягають у: визначенні та характеристиці процесу видобутку / збирання думок споживачів; пропозиції рішення щодо того, як проводити їх аналіз; презентації описаного методу аналізу думок на конкретному прикладі для емпіричної перевірки його дієвості.*

Результати дослідження. *Визначено відмінності між «аналізом настроїв» та «видобутком думок». Для всебічного розуміння думки споживача слід застосовувати три типи аналізу: настроїв, де визначається ставлення до товару (продукту); основних характеристик товару, у ході якого визначаються його особливості; порівняльний. Встановлено, що нині аналітики, як правило, не запитують думку споживача. Споживач сам викладає свої думки у блогах та на різних форумах. Оскільки збір думок є корисним інструментом для прийняття рішень щодо розвитку бізнесу, виникає необхідність у пошуку нових методів щодо накопичення і обробки думок споживачів в інший спосіб, ніж раніше. При цьому у сучасних реаліях (внаслідок розвитку електронної торгівлі і т. п.) під час видобутку думок виникають певні труднощі: необхідність збору даних з різних джерел та обробка великих масивів інформації; різні форми представлення думок в Інтернеті; наявність думок про один і той самий товар на різних мовах тощо.*

Висновки. *Існує багато різних методів для вивчення думок споживачів. При цьому вибір методу має залежати від мети аналізу і форми представлення думки, що аналізується. Більшість методів вимагають, щоб думки були представлені в одній формі. Для вирішення цієї задачі запропоновано використовувати модель LDA (прихованого розподілення Діріхле). Відповідно до цього аналіз думок потрібно проводити у два етапи: ідентифікація теми та реконструкція.*

Ключові слова: думки споживачів, настрої споживачів, модель прихованого розподілення Діріхле, автоматизація аналізу думок.