Text Mining in Business Practice: Automatic Analysis of Customer Reviews for Business Support

Jiri Pribil, *Member, IAENG*

***Abstract* — Automatic or semi-automatic (human-assisted) analysis of text sources can be a crucial key element in the business process. Modern companies face a growing competitive environment in which they do not compete only with their market competitors, but in a certain sense even with their customers. It is the customers who mainly through their customer reviews shape a good or, vice versa, bad impression about the company, its products or services. Popularity of on-line trading, as well as the customer servers, is growing. Timely and accurate response to a customer rating, identification of market sentiment and the possibility of responding to these real-time events is, however, without the support of automated tools almost impossible or very costly. Software tools utilizing specific algorithmic procedures for the analysis of documents (text mining) can be both reliable and cheap solution.**

***Index Terms*—automated model, business support, text mining, n-grams, customer experience**

I. Introduction

CUSTOMER experience (CX) as a part of Customer Experience Management (CXM) is in recent years becoming a buzzword [1]–[3] around which business of many companies revolves.

Under this term companies understand especially broader concern for satisfaction of a customer, whom they are trying to make more loyal and willing to buy their products (services, products). They are trying to, among other things, find out – with a few exceptions which are more or less at random, and above all by hand – how customers respond to their products, advertising campaigns, in which context the references to the company itself, its products or services are occurring.

Precisely with the rise of CXM appear the actual conditions of the motto "The customer is always right" (Harry Gordon Selfridge, founder of London's Selfridges store, 1911). The question is whether this common concept (designated worker of the company monitors internet discussion and responds in particular to the criticism of products/services of a particular company and tries to improve the image) can succeed in the current age, whose motto is "big data" or globalization.

Manuscript received July 18, 2014; revised August 3, 2014. The research was partially supported by the Grant Agency of the Czech Republic under grant No. P403/12/2175 “Methods Supporting Managerial Desision-Making Based on the Analysis of Text Information Sources” and University of Economics in Prague under internal grant project No. F6/78/2014 “Optimization of Multilingual Texts Context Extraction”.

Jiri Pribil is an Assistant Professor and Vice Dean for External Affairs at the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (e-mail: pribil@fm.vse.cz).

A. Customer Feedback Management under the control of the company

For this kind of analysis there are tools on the market, frequently included under the Customer Feedback Management (CFM) services or, more generally, along with Customer relationship management (CRM) under the area of Customer data management (CDM).

Their use, however, is quite different from the goals we want to achieve. All of these tools require dedicated cooperation in the process of obtaining customer feedback on a particular product or service that the company provides. Many companies already use some of the procedures that allow you to get feedback from their customers; these include various surveys and polls provided by the company itself on its web site (for example, after the registration of the goods purchased or even on a random visit of a potential customer) or the so called idea management, when the company makes room for new ideas regarding their products or services. By the statistical evaluation of responses of these (potential) customers is then possible to determine their interests – but the question is how representative these tools can ever be. Very often they are implemented directly into the corporate CRM solution and allow the company to combine specific customer with a specific rating and thus go more "in depth", but they do not reveal much about the actual quality of the product or service.

B. Specialized Customer Feedback Sites

The company should in first place monitor customer reviews particularly on independent servers (review sites, ranking sites), such as the TripAdvisor, Yelp! or Cnet. These servers are primarily intended for sharing customer experiences and are increasingly powerful source of information for future purchases (or for maintaining customer loyalty with existing customers). Taking into consideration the classic 5-step marketing-oriented decision-making process describing the buying behavior – (a) problem recognition, (b) information search, (c) evaluation of alternatives, (d) purchase decision & (e) post purchase behavior – the huge role of ranking sites can be seen with all points except the first one. If the new customer encounters in phase (b) or (c) the negative rating, the company loses them; if the customer shares a negative feeling about the product/service along after the purchase –

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

in phase (e) – new customers won't purchase the product. Searching the customer reviews and reactions, and of course mainly the negative ones, will allow the company to evaluate the responses, putting them straight, or otherwise neutralize their impact. This task, which is often put aside of company's interest, should become part of the current crisis management.

Quite specific are then comparison shopping servers that serve as a direct purchase channel for individual sellers of identical goods assortment. These servers even more clearly reinforce the phase (d) of the above mentioned decision-making process – e.g. Amazon.com, Barnes & Noble or many national-wide operating sites like Heureka.cz in the Czech Republic.

5-star user feedback increases the willingness to pay for the goods, according to some sources, up to 20 percent higher price than for goods of lower user ratings; below the average user rating of the store, service or product on the contrary and in the case of competitive products, means losing significant market shares.

C. Social Sites, On-line Discussions etc.

Large and essentially overlooked source of information, from which potential customers derive information for their future purchases, are social networks.

While in the case of Customer Feedback Sites the tracking of moods of (unsatisfied) customers is basically an easy task even for a human employee, the same task becomes not feasible to carry, if the company responds to the (negative) reactions about itself, its products or services in an open information environment.

There is a huge number of narrowly specialized web discussion groups (such as the fishermen, athletes, owners of watches or cars, etc.), within which stand out very well informed users of specific products or services [4], [5]. This segment of rather more luxurious goods creates a "parallel" world of user experiences, which are described usually more accurately than on the Customer Feedback Sites. People interested in these goods gather the information about these groups especially from a classic search (Google, Bing), they join the discussion in phase (b) and (c) of their decision-making process and the experience of other users either discourages them from the purchase (forces them to purchase competing alternatives) or transforms into the phase (d) of the decision-making process and after a certain time they join the community of users and share their own experiences, thus they get to the stage (e).

The same behavior as on specialized discussion servers then appears in the environment of the most important social networks [6], [7] – Twitter (here customers search for keywords and follow individual discussion) and especially Facebook. Here a company, product or service not only has its own site or group (within which it may in extreme cases even eliminate negative experience, although this procedure is very unproductive in the long term), but there are also dozens and sometimes hundreds of parallel "fan" groups.

The company should be able to monitor all of these resources, and respond, if necessary; all of this in real time, regardless of the language in which users share their experiences.

It is the rating in the unstructured information environment that we focus on in our research.

II. Language Dependence of the System

For the system to function properly it is necessary to distinguish between two possible approaches to the text analysis.

A. Language-dependent analysis of the documents

If it is necessary to track a customer rating in only one (previously known) language, it is possible to consider the dictionary approach. In this case we can use the specific tools such as dictionaries, stemmers reducing various forms of words to their basic form, the database of stop words (words unimportant in terms of their importance in the document, such as conjunctions or prepositions), and other specific language-dependent tools.

Especially in a linguistically rich languages (for example the vast majority of Slavic languages), in which words are made up much more irregularly than, for example, in English, the use of stemmer is one of the few ways to reduce the computational complexity of the process. It is the biggest (and based on our experiments, the only) advantage: a dictionary access is fast – there is only a limited set of words and terms that you need to work with in the system, the tools for working with the dictionary are optimized for the conditions of a particular language (e.g. stemmers are looking for a really basic word forms, not only their stems), the system can "understand" the words on the semantic level.

On the contrary, the dictionary-based system has major disadvantages:

1) Dictionary cannot quickly respond to the evolution of the users' language, the introduction of new words or terms – reaction time of dictionaries on new words is rather decades than years.
2) It mostly does not recognize ungrammatical language, vulgar, abbreviations or emoticons – which are the words that often have the strongest "emotional charge" and can describe the situation for the best.
3) Language tools are provided by third parties and their application entails additional licensing costs. Working with these tools is not unlimited (tools work as a "black box").
4) Language-dependent approach can be used only in situations where it is necessary to track customers in a monolingual environment. This is in practice only possible for small, local companies – in the case of an open international market, it is essential to monitor customer ratings in different languages. For example – the share of exports of the Czech Republic economy production in recent years is around 30 percent. Only a few companies can afford to ignore (talking in average and of course regarding various business sectors) opinions of a third of its customers.

It is obvious that, especially for companies from export-oriented countries, the language-dependent analysis is very difficult to utilize. It would be possible to consider parallel use of multiple monolingual tools (in the Czech Republic, aside from Czech language and in descending order of export volume also German, Slovak, Polish, French, English, ...), but this only strengthens the above-mentioned drawbacks and in our practical experiments it proves to be a difficult passage.

B. Language-independent analysis of the documents

Interesting and in the multilingual globalized world useful is the use of language-independent document analysis. It requires no language-specific tools and can therefore be deployed in the environment of almost any language that is using alphabetic writing system, in which each symbol represents just one phoneme – consonant or vowel – especially Latin, Cyrillic and Greek. Our team is currently focused on syllabic or even logographic languages, but at this point there have not been a sufficient number of experiments.

Because there are no dictionary or other auxiliary language tools available, any analysis of documents (in this case mainly user reviews and customer posts on discussion forums or social networks) is strictly based on statistical and subsequent algorithmic processing of the text.

Of course, this procedure is algorithmically more difficult than the use of language-dependent analysis, and its application must meet certain prerequisites for deployment, as mentioned in chapter III.

Our previous research shows that even without any knowledge of the language that the customer uses in the rating, it is possible to distinguish between positive and negative (which are in terms of the business process much more serious) customer reviews.

III. Terminology and Conditions

In the following text we will use (at the cost of slight inaccuracies) two terms – (a) quantified source of customer ratings *SCR(q)*, and (b) unstructured source of customer ratings *SCR(u)*.

There are strictly defined conditions for these two types of sources.

A. Quantified Source of Customer Ratings SCR(q)

For a properly functioning system it is necessary to define at least one source *SCR(q)*, ideally, however, to define more sources $SCR(q)_1, SCR(q)_2, \ldots, SCR(q)_n$.

Each source represents a particular Customer Feedback Site which contains both text and quantitative rating reflecting customer satisfaction with various companies/products/services (mostly in the form of stars, in a five-point scale from 1 to 5 or six-point scale 0-5).

To be able to pair "verbal rating – quantitative rating" from websites, we use our own universal software web robot, which is parametrically adjustable, so it can crawl through different *SCR(q)* (e.g. Amazon.com for English ratings, Heureka.cz or CSDF.cz for Czech ratings, TrovaPrezi.it for Italian ratings etc.) and store these ratings in the database.

For a description of the parameters of each *SCR(q)* we use in our application a simple XML format, which defines, among other things:

-- Crawling method (sequential, random, in depth, in breadth, number of reviews, …).

-- Crawl frequency.

-- The boundary HTML source code elements of the page with the rating(s) defining both the word form and a quantitative rating.

As a major improvement of our system, we mention that *the information about the language, in which the customer reviews are written, is not explicitly stored* – this is a very important for example for server Booking.com, which contains visitor ratings in different languages. For most other sites we could specify the language within the XML definition, it is, however, not necessary according to our experiments.

Our robot (crawler) is thus able to completely automatically crawl through different Quantified Sources of Customer Ratings *SCR(q)* and sequentially fill the database with quantified customer reviews.

B. Unstructured Source of Customer Ratings SCR(u)

Similarly, we define also the individual sources of customer reviews that contain only a verbal rating – as mentioned above, our experiments are carried out on selected specialized discussion sites, Twitter and Facebook.

In the experiments we perform the role of a company that tries to capture (especially critical) contributions for themselves, their products or services so that they could respond as soon as possible. This implies a need for a more specialized application, which contains potentially interesting resources in the form of URLs to specific pages and discussion forums/groups on Facebook or keywords (hashtag) on Twitter.

At this stage, it is necessary for the users to actively seek and define sources which they consider relevant with respect to their business, and also to define the keywords for which the system searches. These keywords can be the name of the company or the product or service that the company offers to customers – in all garbled or defamatory shapes or nicknames. It is obvious that this set of concepts is not rigid, but it will change over time (particularly in connection with the introduction of new products).

An interesting alternative is not only monitoring posts about products/services of our own company, but also monitoring our direct competitors on the market. It is appropriate to obtain our rival's issues warning (and respond to them as quickly as possible with marketing, such as targeted advertising campaigns), on the other hand, it would be advisable to follow also customer reviews, which are extremely positive for our competitor (e.g. as a reaction to competitor's new product or its variant, which also our company offers).

Sources, that need to be monitored, are again within the XML file (Fig. 2), defining among others:

-- for specialized discussion sites: URL addresses and searched keywords (terms),

-- for sites/groups on Facebook: their addresses and searched keywords (terms),

-- for Twitter: the keywords (hashtags).

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. Analysis and evaluation of Customer Ratings

Analysis of user ratings, which the company must pay attention to, is carried out in several related steps; In doing so, we will pursue only a more complex case of Language-independent analysis, which does not use any additional language tools as described above:

1) The "learning" part of the system crawls repeatedly through Quantified Sources of Customer Ratings *SCR(q)*.
2) The analytical part of the system tracks changes in the defined Unstructured Sources of Customer Ratings *SCR(u)*.
   a) When finding a change in any source, a statistical analysis is carried out, which evaluates the likely mood (positive or negative) of a verbal user rating in the post.
   b) If the verbal rating is significantly positive or significantly negative, information for a responsible person in the company that handles the agenda is created.

The described system is dynamically designed — by adding more Quantified Sources of Customer Ratings we extend the database for more accurate analysis of verbal customer rating (the number of recorded pairs "verbal rating — quantitative rating" increases). Because our system does not recognize the language in which the individual comments are written, it is of course necessary to define at least one *SCR(q)* for each language, which we then analyze.

A. The learning part of the system

The role of the learning part of the system has been sufficiently described in Chapter III.A.

In a specified interval the robot crawls the Quantified Source of Customer Ratings *SCR(q)* and searches, based on the properties of the resource described in the relevant XML file, the pairs "verbal rating — quantitative rating".

Our previous researches [8] imply that even in morphologically rich languages, such as Czech, it does not seem to matter which business type ("domain") the Source of Customer Ratings covers. Even in the case of the training set of movie reviews, our simple SVM (Support Vector Machine) model is capable of achieving accuracy rate of the analysis on test set of customer goods (i.e. completely different domain) over 70%. In the case of one business domain (identical for testing and subsequently training set of comments) and in less morphologically rich languages, the accuracy rate rises above 85% (Italian, French) or even above 90% (English, German). This accuracy is comparable to how the specific verbal contribution is judged by quantitatively independent observer — human.

The ideal situation is, of course, if we are able to find a Quantified Source of Customer Ratings, which corresponds to a specific business domain of a particular company for each language that we want to examine in the analytic part.
An interesting question is the interval in which the resources should be crawled. It does not need to be particularly short; for a very popular *SCR(q)*, with a rapidly growing number of user ratings, the interval should be shorter than for the resources that are more "static". For most sources in our experiments, a reasonable interval seems to be about one month.

The question we want to develop further is whether, after some "reasonable" number of collected pairs "verbal rating — quantitative rating", this interval can be extended, for example, to a six-month cycle. In principle, by repeated addition of new pairs into the database, we are able to capture new language trends (phrases) that may be popular in the short term (advertising claims, for example) and which customers start to use massively in their ratings.

B. The analytical part of the system

The main task of the analytical part of the system, that has been shortly described in chapter III.A., is to search the specified resources (again, regardless of the language in which customers contribute their ratings) for keywords. In this respect, the system is basically no different from any internet search engine such as Google — in our case, however, it searches only the Unstructured Sources of Customer Ratings *SCR(u)*, which interest us in terms of supporting the business.

Logically it follows that the interval at which these *CSR(u)* are searched must be, compared with the *SCR(q)* interval, much shorter; our goal is to allow the company to response rapidly, therefore the resources (social networks, discussion sites) must be crawled basically without a break, or at least in a maximum interval of hours.

If in any specific Unstructured Source of Customer Ratings *SCR(u)* appears the word (term), which is, in terms of customer experience, interesting (name of company — your own or a your competitor's, product/service), the surrounding text will be analyzed. The scope of the text search segment is another parameter in the definition of *CSR(u)* — in practical experiments it seems to be appropriate to follow the text up to a distance of 50-100 words from such a searched term. It is of course possible to extend this limit and thus gain more "context" in which this term (keyword) is listed.

The following step of the text segment analysis from Unstructured Sources of Customer Ratings is then only an algorithmic task that can be described as follows: *"Is there a significant number of positive/negative emotionally toned words?"*. In our experiments we detect this rate of positively or negatively toned words, again, either by using simple SVM model (which combines the analyzed section of text with the database of pairs "verbal rating — quantitative rating", which have been collected by the learning part of the system) or by relatively simple statistical analysis.

The crucial question is, how we know that a word from the text review of the company/product/service has a positive or negative connotation. The answer is the learning part of the system. If the word appears regularly in posts with low ratings (in English typically words such as *"bad"*, *"worst"*, *"fake"*, etc.) and in high-ranking posts its frequency is low, it is likely that it is a negative word. The combination of such words with a keyword we are searching for (company name/product/service) should be a signal problem for the company. On the other hand, words with higher frequency in high-ranking posts ("fine", "OK", "good" etc.) will surely mean customer satisfaction.

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

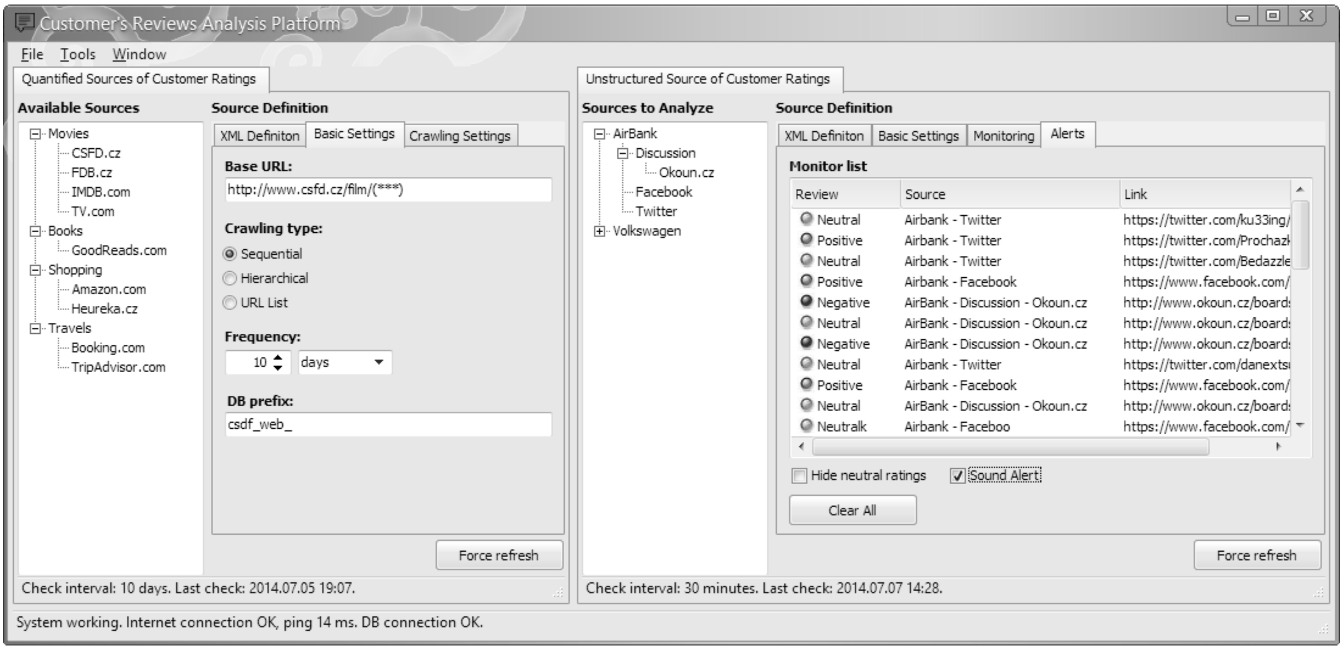

Fig. 1. GUI of the system, beta version.

The basic idea again is that this analysis, that takes into account only the statistical characteristics of the occurrence of each word in each Quantified Source of Customer Ratings, is language independent. E.g. the Czech word *"úžasný" ("awesome"*) will occur only in a satisfied customer reviews (4* or 5*) — and it is completely irrelevant, that the Czech pairs "verbal rating — quantitative rating" in the database form only a very small part alongside English, German or Russian. There is a very low probability that the same word can mean a completely different rating in two different languages; our experiments show that this problem can be almost neglected in a multilingual system.

Another way to improve the system is the use of n-grams to represent quantitative ratings in the learning and analytical phase. This means that the text will be represented by pairs (or triplets) of words and the frequency of such n-tuples is examined. This will allow us the capture even the linguistic elements such as negation. E.g. the pair *"not bad"* will be correctly classified as a rather positive rating (which will occur in a satisfied customer rating) — while in the case of the text representation of individual words, the word *"bad"* would appear both as positive *("This wine is not bad at all."*) and negative evaluation *("This wine is definitely bad."*). Our experiments then go further and in practice we use the concept of unordered {n}-grams, which do not depend on the order of words [9], [10].

For the human operator is then the result of the analysis displayed in the form of a control panel in the system; for every occurrence of a keyword in one of Unstructured Sources of Customer Ratings *SCR(u)*, for which the severe negativity/positivity is evaluated, a semaphore indicating the necessity of problem solving is shown (Fig. 1).

An important feature is that even if the user does not speak the language, they know that there is a good/bad situation and can pass it to the relevant national branch of the company.

V. Conclusion

The introduced system is currently in the phase of a functional prototype and we are negotiating on its pilot deployment in an enterprise environment. We believe that monitoring user evaluations of services and products is in many companies underestimated, especially because of its complexity. Our system can, however, strongly reduce this complexity and automate the tracking of customer ratings.

As a further extension of the system it is possible to search, alongside with discussion groups and social networks, essentially any internet source (blogs, news sites). But it faces the problem of having to use the services of an internet search engine and meet its license conditions (including the need to pay for the use of their API). In general, however, this feature is very easy to implement.

It turns out that even a very simple statistical analysis of word frequency (or n-grams or {n}-grams) can be a very powerful tool in business support.

References

[1] D. Grewal, M. Levy, and V. Kumar, “Customer experience management in retailing: An organizing framework,” J. Retail. 85(1), 2009, pp. 1–14.

[2] W. You, M. Xia, L. Liu, and D. Liu, “Customer knowledge discovery from online reviews”. Electronic Markets 22(3), 2012, pp. 131–142.

[3] T. Lee, “Constraint-based ontology induction from online customer reviews,” Group Decis. Negotiation 16(3), 2007, pp. 255–281.

[4] James H. McAlexander, John W. Schouten and H. F. Koenig, “Building Brand Community,” Journal of Marketing, Vol. 66, No. 1 (Jan., 2002), pp. 38–54.

[5] I. Szmigin, L. Canning, and A. E. Reppel, “Online community: Enhancing the relationship marketing concept through customer bonding”, International Journal of Service Industry Management 16(5), 2005, pp. 480–496.

[6] K. Alkilani, K. C. Ling, and A. A. Abzakh, “The impact of experiential marketing and customer satisfaction on customer commitment in the world of social networks”, Asian Social Science 9(1), 2013, pp. 262–270.

[7] D. Henschen, “Social networks meet customer service,” InformationWeek (1283), 2010, pp. 12–13.
[8] T. Kincl, M. Novák, and J. Přibil, “Getting Inside the Minds of the Customers: Automated Sentiment Analysis,” in *Proceedings of the European Conference on Management Leadership and Governance (ECMLG2013)*, Klagenfurt, Austria, 2013, pp. 122–129.
[9] J. Přibil, “On the Practical Issues of Document Context Extraction And Similarity Measuring,” in *Proceedings of the 16th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*. Mariánské Lázně, Czech Republic, 2013, pp. 253–261.
[10] J. Přibil, T. Kincl, V. Bína, and M. Novák, “Language Independent System for Document Context Extraction,” in *Proceedings of the World Congress on Engineering and Computer Science 2011*, San Francisco, CA, 2011, pp. 51–55.

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)