AARMS Vol. 18, No. 2 (2019) 101–116.

DOI: 10.32565/aarms.2019.2.4

# Sentiment Analysis of Customers[1]

Gábor SZŰCS[2]

*There is an economic and political need to learn and know more and more information about customers, and the social media has recently become the most powerful tool for interaction with people. Customers became users in the social media, who express their opinion and share it with not only companies but also other users. Since a huge amount of reviews, opinions and comments appeared, there is a necessity to extract, aggregate, and analyse them; these are the aims of sentiment analysis. In this paper the technological details of sentiment analysis are presented. The problem types and their solution with text mining are described. Text mining is based on data mining, but steps of text preprocessing with tokenisation, stemming, filtering is an additional important phase before the data mining procedure. The correctness of sentiment analysis solutions can be measured by different validation methods. At the end of the paper the final conclusion is presented.*

**Keywords:** *sentiment analysis, text preprocessing, social media, text mining*

## Introduction

There is an economic and political (or governmental) need to learn and know more and more information about customers (citizens), but in most of the cases, the available information is unstructured, such as text. The governance and economy-developed enterprises were obliged to store terabytes of data, and to handle lots of text documents, so the information sources are available. Companies have to maintain either a lot of employees tasked by maintaining this document mass, or a complex system providing assistance to the workers; not only the maintenance requires lots of resources, but also their analysis. State organisations and enterprises may be interested in reviews, opinions of customers; that is why the concept of sentiment analysis (also called opinion mining, review mining or attitude analysis) [1] as an important type in text document analysis and in social media analysis [2] was born.

Social media has recently become the most powerful tool for businesses to advertise, to communicate towards customers and the most important is to interact with their customers. With the appearance of web 2.0 one-sided communication became two-sided; and customers became users who have the social media tools to express their opinion and

[1] The work was created in commission of the National University of Public Service under the priority project PACSDOP-2.1.2-CCHOP-15-2016-00001 entitled "Public Service Development Establishing Good Governance" in the Workshop for Science of Public Governance 2017/162 BME-VIK "Smart City – Smart Government".

[2] Associate Professor, Budapest University of Technology and Economics; e-mail: szucs@tmit.bme.hu; ORCID: 0000-0002-5781-1088

to share it with not only companies but also other users. Since huge amount of reviews, opinions, comments and clicks appeared, there is a need to aggregate them to statistics and to analyse them. This enormous volume of data can be a source of a lot of useful information if managed well, which could lead to advantage against competitors or higher profit at the companies and this can give also a large advantage for state organisations, as well; so the automatic analysis of this user-generated feedback can have a large impact.

In this paper, we try to present the importance of written feedbacks and their emotional impression, which has influence on the decision of other customers and citizens, let it be positive, negative or even neutral. It can be stated that the research is highly intense in this area, so the methods and the developed algorithms or even the mode they are used within business decisions are getting more precise and accurate every day. Based on this, decision-makers realise the importance and power of this topic, thus it gets a higher attention.

After introducing this topic, there is an overview into the details of text mining (types of text analytics) and sentiment analysis (sources and steps of sentiment analysis); after that the different opinion types, possible solutions are presented. Cross-industry standard process for data mining (CRISP-DM) procedure as a possible way is described in details, and there is text specific preprocessing before this procedure. Finally modelling, validation and evaluation are presented before the conclusion of this paper.

## Types of Text Analytics

During the past years, the need of easy and flexible access to the constantly increasing amount of text documents in digital form is growing rapidly each day. Due to this demand, the content-based document management [3] tasks have gained an outstanding position in the information systems field. One of such tasks is text classification (as a subset of text mining), as the process of labelling natural language texts with thematic categories from a predefined set.

Text mining [4] refers to the process of identifying, deriving and analysing high-quality information from text. Data mining deals with data stored in databases in a structured form; contrary texts definitely are not structured. So for processing text documents, the files automatically need to be transferred into databases, and this has a transaction cost. Because of unstructured data, the source of the problem is that natural languages are for human communication and not for computer processing. Text mining aims at the combination of the human linguistic knowledge with the large processing capacity of computers. The goal of text mining is to gain useful information from unstructured data, to extract meaningful numeric metrics from the text, and to modify the information accessible to data mining algorithms.

The main purposes of text classification are: *spam detection (spam filtering),* [5] where a spam filter program is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox; *plagiarism detection,* [6] where the aim is to locate instances of plagiarism within a work or document (typically essays or reports, but, plagiarism can be found in any field, including novels, scientific papers, art designs and source code); *language identification* [7] (or language detection) as the problem of determining which natural language the given content is in; classification of news sets

(or advertisements); *age/gender identification;* [8] and *sentiment analysis*. [9] Sentiment analysis is the task of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in a textual input. [10]

Sentiment analysis [11] is also a field of study, which attempts to analyse people's opinions, attitudes, sentiments, and emotions on entities such as products, services and organisations. These organisations can be an enterprise, or state organisation, such as a council or public administration. The holders of opinions are clients (in the first case), or citizens (in the latter case), we can say—in a common word—they are customers. Sentiment analysis is an application of text analytics (text mining as one of the hot topics of data mining). Text analytics contains information extraction from unstructured data and the process of structuring the input text to derive patterns and evaluating or interpreting the output data. In text analytics, there is a process to convert unstructured text data into meaningful data analysis to measure customer opinions, product reviews, feedback, to provide search facility and sentimental analysis to support fact based decision-making. Besides sentiment analysis, text mining also involves categorisation, clustering, pattern recognition, information extraction, link analysis, visualisation, predictive analytics and relation modelling.

## Sources and Steps of Sentiment Analysis

Sentiment analysis is one of the text analysis types, where data is contained in a natural language text. This gives possibility for an organisation to derive potentially valuable business insights from text-based content such as emails and posts on social media streams like Facebook, Twitter and LinkedIn, and this can be applied in business decisions and strategy. Sentiment analysis helps to monitor attitudes and feelings across the Internet on various topics by building a system that collects and examines product or service reviews in blog posts, comments, reviews, or tweets. To collect these text type data, there is a need to use an application that automatically processes the contents of web pages by visiting a web site and begins to "crawl" the links that are found in the investigated pages. This way, the list of terms and documents available on the site can be derived automatically and can quickly determine the most important terms and features that pages and linked pages describe. So we can track products, brands and people, and determine their positive or negative opinions.

The typical output of the sentiment analysis task is a pair, i.e. the extracted opinion consists of two parts: a product (or target) "g", and a sentiment "s" on the target, where at the binary case the polarity of the sentiment can be positive (like good, great, fast, nice, and so on), or negative (bad, ugly, worse, etc.). Let us present an example: "I got the iPhone as a present for Christmas, and I like it very much." The extracted result of this sentence: (g = iPhone, s = positive).

An opinion can be described not only by a pair, but also by a more complex structure, for example by a quintuple (e; a; s; h; t), where "e" is entity, "a" is aspect, "s" is sentiment, "h" is holder, and "t" is time. The next paragraph presents an example review.

"I purchased the Kindle Fire hd7 for my wife's birthday, and she likes it. It was processed and shipped quickly by amazon. I think though the price of the hd7 is a bit high. Because the resolution of the camera is low." December 11, 2017, Joe.

In this example the extracted attributes and the quintuples are the following:

e: entity – the product, Kindle Fire HD7;

a: aspect of the product – price/camera of Kindle Fire HD7;

s: sentiment – positive or negative;

h: opinion holder – Joe, Joe's wife;

t: time of the opinion – December 11, 2017.

Quintuples:

1. sentence: (Kindle Fire HD7; product; positive; Joe's wife; December 11, 2017);
2. sentence: (amazon; positive; Joe; December 11, 2017);
3. sentence: (Kindle Fire HD7; price; negative; Joe; December 11, 2017);
4. sentence: (Kindle Fire HD7; resolution of the camera; negative; Joe; December 11, 2017).

The steps of this opinion extraction can be enumerated in five stages:

1. entity, aspect and opinion holder extraction and categorisation;
2. time extraction and standardisation;
3. aspect sentiment classification (determine whether the sentiment for a given aspect is positive or negative);
4. generation of the opinion quintuple;
5. summarisation of opinions.

## Opinion Types

### *Regular or Comparative Opinions*

In the sentiment analysis, opinions can be divided into different types: regular and comparative opinions. [12] A comparative opinion expresses a relation of similarities or differences between two entities or a preference of the opinion holder based on some shared aspects of the entities; for example: "Google search engine is better than Bing." A comparative opinion is usually expressed using the comparative or superlative form of an adjective or adverb, although not always (e.g. prefer). Regular opinions do not possess comparison. There are lots of typical sentences in this category, like: "Google search engine is good."

### *Explicit and Implicit Opinions*

Explicit opinion is a regular or comparison opinion that is expressed in a subjective statement. [13] Example: "The design of the Gmail Service is very nice." Implicit opinion is an objective statement, like: "Image search in Bing is two times faster than in Google."

It is a factual statement about measurable/comparable quantities, and it implies a positive sentiment towards Bing search and a negative towards Google search. Implicit opinion cannot be only a comparative one, but regular, as well; e.g. “a battery is very small”. The computers do not know that small is a good or bad attribute. A word that is considered to be positive in a situation may be considered negative in another situation. For example the word “long” is positive if a customer says about the battery life of a laptop. If the customer said that the start-up time of this laptop was long, then this is a negative opinion. It is a challenge to detect this kind of situations; additional challenges and difficulties are described at the end of this paper.

### *Direct and Indirect Opinions*

Opinions can be further categorised into direct and indirect opinions. [14] Direct opinion is a statement expressed directly on an entity or an entity aspect, like: “The battery life is good.” With indirect opinion, the opinion is hidden in the relation of two entities, for example in the next sentence: “After the upgrade of the ADSL modem, the Internet access became even slower.” In this opinion the entity is the ADSL modem, its aspect is the speed, and it gives indirectly negative sentiment. Explicit and direct opinions are easier to detect by computers than the indirect and implicit types.

Although sentiment words and phrases can be collected into a dictionary for positive and negative sentiments, however this will give only a poor accuracy in the results of the sentiment analysis; additionally, these are only the two poles of the polarity and sometimes the end user needs more sophisticated solutions between the two extremes. Instead of dictionary based methods, data mining algorithms give better solutions. In the next session CRISP-DM [15] is described, as the one of the most frequently used data mining methodology.

## CRISP-DM Methodology

During the solution of the sentiment analysis problem, we can use a standard process of data mining, the well-known CRISP-DM procedure (methodology) that describes commonly used approaches for data mining with the steps of the procedure. These steps are the following: business understanding, data understanding and preparation, modelling, evaluation and deployment.

### *Business Understanding*

The first phase is understanding the targets and requirements of the task, then converting this knowledge into a data mining problem, and designing a preliminary plan to achieve the objectives. In our case the goal is to predict reviews, which can be positive or negative (so-called sentiments).

### *Data Understanding and Preparation*

The second phase is collecting the initial dataset, getting familiar with them, and identifying data quality problems to avoid the loss of the later parts of the process. It is closely related to business understanding, therefore going back and forth between the two phases is frequent. Then the next step is the preparation of the final dataset by transformation, integration and cleaning from the initial raw data. Tasks include table, record, and attribute selection, as well.

### *Modelling*

In this step various modelling techniques can be selected and applied, and their parameters are adjusting to optimal values. Some techniques have specific requirements on the form of data, in these cases stepping back to the data preparation phase is required.

### *Evaluation*

The utility of the built model(s) can be evaluated, and the models can be tested whether they fulfil the initial requirements and give back proper values. The aim is to determine whether there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

### *Deployment*

At the end, the conclusion of the whole procedure will need to be organised and presented in a way that is useful to the customers. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring. It is important for the customer to understand the obtained knowledge and conclusion.

## *Data Mining Procedure for Sentiment Analysis*

Sentiment analysis is a problem in data mining, more precisely in text mining, which is concerned about classifying natural language sentences based on their negative or positive overtone. In order to predict the sentiments in an unknown text document, we can use CRISP-DM, so we need to learn from the text document, where are the known sentiments (this is the label). During this learning we build a model based on train data set (i.e. known text documents with label information). Two data subsets are required, one of them involves sentences labelled with positive sentiments and the other involves negative sentiments about reviews. The built machine learning model will be able to predict sentiments in test data sets (i.e. in unknown text documents). So the input data is a disjoint set of two different types of data:

- training data – a baseline data, which consists of sentences and the assigned sentiment labels 1 or 0, meaning a positive or negative sentiment of the sentence, and these labels come from humans' decisions;
- test data – a validation data, which contains sentences, but not the corresponding assigned labels. The goal is to classify these sentences, based on the insights and logical relations learnt from the training data set. Later, the predictions will be tested by humans.

In order to create a corpus for sentiment analysis, we need a large set of documents, but the collection of documents is sometimes not as trivial as it seems. Social websites are well protected from crawling because they are aware of the value of the data, and they do not give it to companies or to business owners in an easy or cheap way. Another problem is the constantly growing number of online reviews, posts. This phenomenon suggests the implementation of an automatized method, which grabs the new post and reviews without human interaction from time to time.

## *Software Possibilities*

In order to be able to solve the task presented above, there is a need for an analysis application, capable of predicting and assigning labels to sentences. There are several data analysis software on the market, like RapidMiner, [16] SAS [17] and of course also programming languages, which are designed for statistical analysis, or feature external modules, which support these algorithms. A typical example for the first case is language R, while for the latter Python is a good example. These platforms and programming languages have their pros and cons. Platforms provide an easy-to-use, user-friendly graphical interface to carry out complex analysis processes. It is an easy solution for those, who are beginners in this field of data science; it enables these users to get familiar with these methods. However, the usage is quite rigid, the flexibility is limited. The user can only use the previously implemented processes, which are available in a predefined manner. On the other hand, programming languages offer a wide range of possible solutions, one can be creative during the implementation, and there is no limitation included by a predesigned system. However, it must be mentioned, that the greater freedom comes with a cost, the implementation is harder, error-prone and requires deep knowledge of the used algorithms alongside with programming skills. The needed external data analysis python modules within the confines of opinion mining task are Natural Language Toolkit (NLTK) [18] and Scikit-learn. NLTK is a leading platform for building Python programs, which can process and work with human language data. It features several corpora alongside with different text processing methods such as stemming, tokenisation, classification, etc. Scikit-learn is a Python toolkit (besides the well-known python modules Numpy and Scipy) for efficient data mining and analysis, containing several data analysis models, methods, such as classifiers, regression classes, as well as the needed helper procedures for preprocessing and data cleaning.

For the machine learning models pre-processing is needed to get structured information from the continuous (unstructured) text. This pre-processing consists of many processes, as transforming letters to lower case, tokenising the words, stemming the words into

root words, filtering the stop-words, which are described below. The tokenisation phase is responsible for splitting the continuous text documents into character series, so called tokens. The stemming phase is liable for reducing a token to its stem or root. The usage of the stop-word filtering can be useful, because it removes those words that do not affect the classification task. The details are described later.

### *Levels of Sentiment Analysis*

Most of the documents consist of several hierarchical elements, which have significant contribution to the reader to facilitate interpretation. For example, in a book, the volumes, chapters, paragraphs, sentences can be the hierarchy levels. It is worth mentioning, of course, that individual documents may have different structures within the same types. In general, the sentences can be a proper hierarchy level; however, dividing the continuous text into sentences is not a trivial task. The idea is to segment the text based on sentence terminals, as these indicate the end of the sentences; nevertheless, these punctuation marks are widely used and have many other meanings, so they not necessarily sign the end of the sentences.

Not only documents consist of several hierarchical elements, but sentiment analysis task can be executed in different levels, as well, similarly to general hierarchical levels. Since this task is a classification of the polarity of a given text, this investigated text will give the appropriate level: this can be a document, or paragraph, or sentence, or another aspect level. The polarity can be binary (positive or negative), or can have three classes (positive, negative, or neutral). At advanced sentiment analysis, the classification of sentiments can be at emotional states such as nervous, sad and happy, etc.

If a review has both positive and negative sentences, then this will be a mixed expressed opinion at the level of the review. But we can drill down to sentence level: if each sentence in the review possesses a clear polarity, then this review can be summarised by the positive and negative sentences. However, some people combine different opinions in the same sentence, which is hard to understand for computers, because of ambiguity, for example “it was great even though I dislike this genre of movies” or “the movie was good, but some parts were quite boring” as mixed expressed opinions.

## Text Preprocessing

During preprocessing the unstructured texts are brought to numerical objects focusing on the aim of the task and bringing data to a format that is suitable for storage according to the nature of the text. The main role of preprocessing [20] is to analyse more effectively by unification, formatting and normalisation of data. This involves tokenisation, the stemming of words, filtering of stopwords and so on.

## *Tokenisation*

There is a lower level in the hierarchy described above, and we can get this level by tokenisation, as the first phase of pre-processing in text mining. Strings (character series) with a self-contained meaning are called tokens, which can be separated from other character series in some form. One of the most appropriate solutions to this task is that the document should be broken off along all punctuation marks and spaces. However, there is still a problem with punctuation in some cases, where it is not necessary to separate two strings, because it is separated by a punctuation mark. This kind of drawback is especially true in the IT field (with URL, IP-addresses, etc.), and these marks are usually used after abbreviations, or when using serial numbers and dates. Additional problems can occur at other marks, e.g. in case of a hyphen mark, it is not clear that there are two interlinked, but distinct words (they are hyphened or separated). Similarly, it may be true for spaces, as well, that a space-separated word may be associated, but you can even talk about triples, as well (e.g. family names and first names). Considering these problems, it can be seen that tokenisation is not an easy and clear subtask in the whole process. [21]

## *Stemming*

The next step in the preparation is the word process, in which we try to reach the stem (like dictionary format) of the words. In general, all languages use some kind of tags like prefixes and suffixes in order to get modified words. The Hungarian language contains lots of such tags, it has a very rich morphology. Stemming [22] differs from word lemmatisation, because the output of the latter one is the dictionary format. But from an informatics point of view, we use stemming to get the root of the word, the so called stem; because it is quite enough for words of the same meaning to be in the same form, while different words are different. This task can also be solved by cutting off suffixes, called truncation; but truncation is only a simplified version of stemming, and we can get pure results with this.

## *Stop-Word Filtering*

Token is an occurrence of the character series; we can define term (as a token type), because the same words (tokens) are repeated in a document many times, so we collect them into one type, so called term; and these terms constitute the raw dictionary. The raw dictionary contains representative words and less representative ones; the latter ones occur in every document (or even multiple times), they are considered superfluous, since they will not change the final result in the classification and in the analysis. These words are called stop-words. [23] The next step in the process is to filter out these words from the whole set. Typical examples are articles, prepositions, expletives, e.g. *the, a, an, whether, in,* etc. If we leave these words, we will not change the final result, but we can significantly reduce the number of tokens, which will change the result with a simpler and faster computing task.

### *Document Vectors*

At the end of the preprocessing, vectors are created for each document. Different types of document vectors can be formed with different values for term weights; tf-idf, tf, term occurrences or binary term occurrences are possibilities. The tf-idf weight system [24] is short for the term frequency—inversed document frequency, where term frequency stands for the ratio of a term in a document (i.e. the number of times that a term occurs in a document divided by all the terms in the investigated document). Inversed document frequency stands for the logarithmically scaled inverse fraction of the documents that contain a given word. Tf-idf multiplies these values to obtain term weights. The tf weight system stands for term frequency, the number of times that a term occurs in a document divided by all the terms in the investigated document. This is simply used to obtain term weights. As the name implies, the term occurrences method simply uses the amount of times a term occurs in a document as weight. Binary term occurrences: this is highly similar to the term occurrences method except that it uses binary values, so there is a 1 if the document contains the token and a 0 if it does not. These vectors can be used for machine learning models. [25]

## Modelling, Validation and Evaluation

There was a sentiment analysis problem that was solved by methods described in this paper. The steps of own solution are the following.

- Initialisation.
- Input data file processing.
- Transforming the input data to the desired format (each model requires a specific input data format, usually a list of tuples, with two members containing the tokenised, stemmed words and the assigned label for this sentence).
- Splitting the input data set to train set and validation set.
- Model training.
- Model validating, performance measurement.
- Prediction based on the previously trained model.
- Output data file writing, this will be the original test file enrichment with the predicted labels.

The step of splitting the input data set to train set and validation set is required to be able to approximate the results of the classifier. As during the development, the test data labels are not available, the only way to validate the accuracy of the training is to use only a subset of the training data to actually train a data analysis model and use a smaller subset of the training data to validate the results. Note that with this approach, the training data set will be stripped partially, with the cost of providing a smaller set for the model to learn on, however being able to measure the performance of the model. Without this approach, there would be no logical basis to select the best performing model. During the implementation 75% of the training data was actually used for training and 25% for validating. This parameter can be

fine-tuned. Possible validation methods discussed below can be used for splitting the data set in order to train and validate classifiers.

- *Bootstrapping validation*
  Bootstrapping validation [26] is sampling with replacement. In sampling with replacement, at every step all examples have an equal probability of being selected. Once an example has been selected for the sample, it remains candidate for selection and it can be selected again in any other next step. Thus a sample with replacement can have the same example multiple number of times. (The remain set, i.e. the non-selected examples can constitute the test set.) More importantly, a sample with replacement can be used to generate a sample that is greater in size than the original sample set.
- *Split validation*
  In split validation, the input sample set is partitioned into two subsets. One subset is used as the training set and the other one is used as the test set. The model is learned on the training set and is then applied on the test set. This is done in a single iteration, which can cause large variance in the results; therefore, it might not be very suitable in this case.
- *Cross-validation (X-validation)*
  Cross-validation [27] is highly similar to split validation, except that it is repeated for several iterations. The input sample set is partitioned into k subsets of equal size. A single subset of the k subsets is retained as the testing data set and the remaining k – 1 subsets are used as training data. The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the testing data. The k results from the k iterations then can be averaged, so variance in the results is smaller than in split validation; therefore, it is likely to use the most suitable on the relatively small training dataset provided, moreover on other general training datasets, as well.

As written above, the ratio of the splitting was 75%–25% for validating, but instead of split validation the cross-validation was used with k = 4 parameter.

## *Results*

In order to test the developed solution described above two data sets were selected. These contain sentences labelled with positive or negative sentiment (neutral sentences were filtered out), extracted from reviews of restaurants (Yelp dataset with 1,000 opinions) and general products (Amazon dataset with 1,000 opinions). The accuracy, as a result of the binary classification of opinions can be seen in the next table.

Table 1. *The accuracy, as a result of the binary classification of opinions.*
[Edited by the author.]

| **$k^{th}$ subset in the cross-validation** | **Yelp** | **Amazon** |
|---|---|---|
| $1^{st}$ subset | 78.2% | 90.2% |
| $2^{nd}$ subset | 72.4% | 89.8% |
| $3^{rd}$ subset | 74.9% | 90.0% |

| | | |
|---|---|---|
| 4th subset | 68.6% | 90.9% |
| **Average** | **73.5%** | **90.2%** |

It was more difficult to learn the sentiments in the Yelp dataset as can be seen in the table (the professionalism of Amazon is better); the variance of the results was larger in the Yelp dataset. As can be seen, the results depend on the characteristics of the training set used.

## Application Possibilities and Difficulties

Opinion mining helps in achieving various goals like observing public mood regarding political movements, market intelligence, the measurement of customer satisfaction, movie sales prediction and many more. Such application possibilities are collected and described below.

At high traffic customer service centres automatic processing of user feedback, like messages, emails, opinions is crucial. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed automatically to the most appropriate department or agency; for instance email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments; at the same time the emails are screened for inappropriate or obscene messages, which are automatically returned to the sender with a request to remove the offending words or content, and also filed into a predefined folder on the appropriate server. The solution described in this paper is able to filter out automatically the most undesirable "spam emails" [28] based on certain terms or words that are not likely to appear in legitimate messages. This way such messages can automatically be discarded. Some subtasks can also be simplified in order to save time for the management of personal or work-related e-mails.

Similarly to sentiment analysis, text mining can also be used in order to fight violent (e.g. crime or terror) activities on the internet (instant messenger or internet relay chat). Information extractive and text analysing methods are applied on the examination of the huge quantity of the document. This is the most effective way to get names, location, and relations between them and the sentiments, and to detect crime or similar (violent) activities on the dark web forum. [29]

Customer feedback is an important measurement of success at enterprises that sell products or services. [30] It is not only important to develop but also to improve business strategy and of course to increase the income and the profit by getting more customers and make them more loyal. To reach this aim it is important to know what are their opinions about the products and services, and this information can be measured by analysing qualitative interviews, so customer preferences will be much clearer.

The opinion holders do not always express own opinions the same way. The traditional text mining relies on the fact that small differences between two pieces of text do not change the meaning very much. However in sentiment analysis this is not true, because there is a large difference between the semantics of a sentence "the service was great" and another sentence "the service was not great". This challenge can be partially solved by sentiment shifters [31] (sentiment shifters are words that change the meaning of the sentiment to its opposite: like *not good*). Another technique to solve this problem is

generating n-grams, [32] in sentiment analysis bigrams would be most suitable because we can take the preceding word into consideration this way. For instance, “good” might have a strong positive sentiment, but “not good” has a negative sentiment even though the sentence contains the word “good”. Bigram generation would give us the tokens “not, not_good, good”, where “not_good” will get a negative weight during classifier training. Where with unigrams just the sentiment of “good” would be slightly reduced.

Another aspect is that sentiment analysis can be hard to process in certain corpus because it may contain ambiguities and sometimes syntax and semantics are not too easily understandable. Slang, language specifics to age groups, irony, professional jargons and sarcasm belong to these kind of challenges. [33] Syntactic dependency (dependency tree) method can help with this problem.

Sometimes it is difficult to understand what the opinion holder thought based on a short piece of text because it lacks context. For example: “That service was as good as in the last year” is entirely dependent on what was the client’s opinion in the previous year. This kind of challenge can be solved by the revealing of cross-references among sentences.

## Conclusion

In public administration there are lots of documents (particularly text document), and a document management system should handle them by an intelligent process (the attribute “intelligent” is important for smart city, smart public administration). A text document can be of a wide variety of types, a document can be written in Word (Office), or as a blog post on the Internet or an e-mail. First, the task is to find a general representation in which each document can be described and the preprocessing of text mining is able to create this kind of general representation.

Sentiment analysis, or opinion mining, refers to the use of natural language processing, data mining and text analysis to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. In this paper the sources (like Facebook, Twitter and LinkedIn), steps and types of sentiment analysis were presented. In the sentiment analysis, the opinions can be divided into different types, regular and comparative opinions as explained above. Explicit opinion is a regular or comparison opinion that is expressed in a subjective statement; and the opposite of this, the implicit opinion is an objective statement. The opinions can be further classified into direct and indirect opinions; the direct opinion is a statement expressed directly on an entity, while the indirect one is hidden in the relation of entities.

In sentiment analysis, the hierarchical levels are also described from sentence level (as a low level) to document level (highest level). The polarity of opinions in each level can be binary (positive or negative), or can have three classes (positive, negative, or neutral). The preprocessing of opinion mining involves tokenisation, the stemming of words, filtering of stopwords and so on. At the end of the paper we can conclude that sentiment analysis is a difficult task, the larger complexity of opinion is the most difficult in this task. The difficulty of this problem also depends on the number of polarity classes and the number of aspects and products. Although there are some appropriate techniques, but policy

discussions, indirect, implicit expressions of opinions are still more difficult to reveal in the text.

## References


[1] WAWRE, S. V. – DESHMUKH, S. N.: Sentiment classification using machine learning techniques. *International Journal of Science and Research (IJSR),* 5 4 (2016), 819–821. DOI: https://doi.org/10.21275/v5i4.nov162724

[2] ANSTEAD, N. – O'LOUGHLIN, B.: Social Media Analysis and Public Opinion: The 2010 UK General Election. *Journal of Computer–Mediated Communication,* 20 2 (2015), 204–220. DOI: https://doi.org/10.1111/jcc4.12102

[3] ZANTOUT, H. – MARIR, F.: Document management systems from current capabilities towards intelligent information retrieval: an overview. *International Journal of Information Management,* 19 6 (1999), 471–484. DOI: https://doi.org/10.1016/s0268-4012(99)00043-2

[4] NASSIRTOUSSI, A. K. – AGHABOZORGI, S. – WAH, T. Y. – NGO, D. C. L.: Text mining for market prediction: A systematic review. *Expert Systems with Applications,* 41 16 (2014), 7653–7670. DOI: https://doi.org/10.1016/j.eswa.2014.06.009

[5] McCORD , M. – CHUAH, M.: Spam Detection on Twitter Using Traditional Classifiers. In *International Conference on Autonomic and Trusted Computing.* Berlin, Heidelberg, Springer, 2011. 175–186. DOI: https://doi.org/10.1007/978-3-642-23496-5_13

[6] LUKASHENKO, R. – GRAUDINA, V. – GRUNDSPENKIS, J.: Computer-based plagiarism detection methods and tools: an overview. In *Proceedings of the 2007 international conference on Computer systems and technologies.* Rousse, June 14–15, 2007. IIIA.18-1–IIIA.18-6. DOI: https://doi.org/10.1145/1330598.1330642

[7] LOPEZ-MORENO, I. – GONZALEZ-DOMINGUEZ, J. – PLCHOT, O. – MARTINEZ, D. – GONZALEZ-RODRIGUEZ, J. – MORENO, P.: Automatic language identification using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on Communications.* Sydney, 2014. 5337–5341. DOI: https://doi.org/10.1109/icassp.2014.6854622

[8] MARQUARDT, J. – FARNADI, G. – VASUDEVAN, G. – MOENS, M. F. – DAVALOS, S. – TEREDESAI, A. – COCK, M. De: Age and gender identification in social media. In *CLEF 2014 working notes.* Presented at the *5th Conference and Labs of the Evaluation Forum.* Sheffield, 15.09.2014. 1129–1136.

[9] CAMBRIA, E.: Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems,* 31 2 (2016), 102–107. DOI: https://doi.org/10.1109/mis.2016.31

[10] MONTOYO, A. – MARTÍNEZ-BARCO, P. – BALAHUR, A.: Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems,* 53 4 (2012), 675–679. DOI: https://doi.org/10.1016/j.dss.2012.05.022

[11] RAVI, K. – RAVI, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems,* 89, (2015), 14–46. DOI: https://doi.org/10.1016/j.knosys.2015.06.015

[12] VARATHAN, K. D. – GIACHANOU, A. – CRESTANI, F.: Comparative opinion mining: A review. *Journal of the Association for Information Science and Technology,* 68 4 (2017), 811–829. DOI: https://doi.org/10.1002/asi.23716

[13] JIN, W. – HO, H. H. – SRIHARI, R. K.: OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Paris, June 28 – July 1, 2009. 1195–1204. DOI: https://doi.org/10.1145/1557019.1557148

[14] LIU, B.: Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies,* 5 1 (2012), 1–167. DOI: https://doi.org/10.2200/s00416ed1v01y201204hlt016

[15] SHAFIQUE, U. – QAISER, H.: A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research,* 12 1 (2014), 217–222.

[16] HOFMANN, M. –KLINKENBERG, R. eds.: *RapidMiner: Data mining use cases and business analytics applications.* New York, CRC Press, 2013.

[17] FERNANDEZ, G.: *Data mining using SAS applications.* New York, CRC Press, 2010. DOI: https://doi.org/10.1201/EBK1439810750

[18] PERKINS, J.: *Python 3 text processing with NLTK 3 cookbook.* Birmingham, Packt Publishing Ltd., 2014.

[19] PEDREGOSA, F. – VAROQUAUX, G. – GRAMFORT, A. – MICHEL, V. – THIRION, B. – GRISEL, O. – BLONDEL, M. – PRETTENHOFER, P. – WEISS, R. – DUBOURG, V. – PASSOS, A. – COURNAPEAU, D. – VANDERPLAS, J. – BRUCHER, M. – PERROT, M. – DUCHESNAY, É.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12 (2011), 2825–2830.

[20] UYSAL, A. K. – GUNAL, S.: The impact of preprocessing on text classification. *Information Processing & Management,* 50 1 (2014), 104–112. DOI: https://doi.org/10.1016/j.ipm.2013.08.006

[21] VERMA, T. – RENU, R. – GAUR, D. : Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems,* 7 2 (2014), 16–18. DOI: https://doi.org/10.5120/ijais14-451139

[22] WILLETT, P.: The Porter stemming algorithm: then and now. *Program,* 40 3 (2006), 219–223. DOI: https://doi.org/10.1108/00330330610681295

[23] YAO, Z. – ZE-WEN, C.: Research on the Construction and Filter Method of Stop-word List in Text Preprocessing. In *2011 Fourth International Conference on Intelligent Computation Technology and Automation (ICICTA 2011).* Vol. 1. Piscataway, Institute of Electrical and Electronics Engineers (IEEE), 2011. 217–221. DOI: https://doi.org/10.1109/icicta.2011.64

[24] AIZAWA, A.: An information-theoretic perspective of tf–idf measures. *Information Processing & Management,* 39 1 (2003), 45–65. DOI: https://doi.org/10.1016/s0306-4573(02)00021-3

[25] LAN, M. – TAN, C. L. – SU, J. – LU, Y.: Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions On Pattern Analysis and Machine Intelligence,* 31 4 (2009), 721–735. DOI: https://doi.org/10.1109/tpami.2008.110

[26] ISAKSSON, A. – WALLMAN, M. – GÖRANSSON, H. – GUSTAFSSON, M. G.: Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters,* 29 14 (2008), 1960–1965. DOI: https://doi.org/10.1016/j.patrec.2008.06.018

[27] ARLOT, S. – CELISSE, A.: A survey of cross-validation procedures for model selection. *Statistics Surveys,* 4 (2010), 40–79. DOI: https://doi.org/10.1214/09-ss054

[28] SHARMA, A. K. – SAHNI, S.: A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering*, 3 5 (2011), 1890–1895.

[29] ABBASI, A. – CHEN, H.: Affect intensity analysis of dark web forums. In *Intelligence and Security Informatics, 2007 IEEE.* Piscataway, Institute of Electrical and Electronics Engineers (IEEE), 2007. 282–288. DOI: https://doi.org/10.1109/isi.2007.379486

[30] GAMON, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics.* Geneva, 2004. DOI: https://doi.org/10.3115/1220355.1220476

[31] XIA, R. – XU, F. – YU, J., QI, Y. – CAMBRIA, E.: Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52 1 (2016), 36–45. DOI: https://doi.org/10.1016/j.ipm.2015.04.003

[32] GHIASSI, M. – SKINNER, J. – ZIMBRA, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40 16 (2013), 6266–6282. DOI: https://doi.org/10.1016/j.eswa.2013.05.057

[33] FARIAS, D. H. – ROSSO, P.: Irony, Sarcasm, and Sentiment Analysis. In POZZI, F. A. – FERSINI, E. – MESSINA, E. – LIU, B. eds.: *Sentiment Analysis in Social Networks.* Philadelphia, Elsevier, 2017. 113–128. DOI: https://doi.org/10.1016/B978-0-12-804412-4.00007-3