Using sentiment analysis with Big Data tools to enrich knowledge on society in the city

Jacek MAŚLANKOWSKI
Departments of Business Informatics, Faculty of Management,
University of Gdańsk, Poland
jacek@ug.edu.pl

Abstract
Using Big Data in terms of providing valuable information for city authorities is usually related to the machine generated data, mostly coming from various sensors installed on different parts of the cities. One of the most common example is a road sensor. It can be used to plan the roads building in the city. However, the valuable data can also be provided from continuing analysis of human generated data, provided by people on different communication channels used by the city authorities. It includes social media portals and self-government websites in which people can create content, such as comments. The aim of this research is to show the value added for the city authorities by making sentiment analysis on various social media and comments on websites. These types of communication is very often a subject of analysis for enterprises to perform the market recognition of customers, but there is still lack of using these methods by city authorities. The goal of this paper is to show a case study of using different Web 2.0 and 3.0 communication forms to build a common view of city inhabitants related to different aspects of the city. For this case study, a proposal framework has been developed and illustrated, using different types of text mining methods to make sentiment analysis. The results from the study show that Big Data may have a big impact on supporting the development of the city. The proposal of the framework presented in this paper is ready to be applied in a production process and serve for the city. The threats and opportunities have been identified and future work has also been presented.

Keywords: big data, sentiment analysis, decision making, text mining, web mining

1. Introduction
In recent years, we can observe a wide range of possibilities of using Big Data tools for different purposes. The most common use of such tools is to make analysis on the content of web pages, named web scraping and web mining, to find valuable information. This type of data that can be processed on various websites is usually unstructured and well known as human sourced-information. In some cases, the data are semi-structured and shared in database files formats, such as XML or JSON. In this paper, Big Data is treated as a data source for sentiment analysis. This data source is unstructured and the valuable information can be provided by analysis of the context of data. A case study shown in this paper use a software that has been built to find public opinion about different initiatives.

The goal of this paper is to show the possibilities of using Big Data tools for sentiment analysis to provide information on public opinion on initiatives made by the city authorities. It was shown from two perspectives – when the number of observations is

low and the number of observations is rather high. It was presented in this paper as, respectively, use case 1 and use case 2.

The paper is divided into five parts. After introduction, in the second part of the paper, there is a theoretical background on the research done in this matter. In the third part, information on the case study was included. There is also an information on the framework used to achieve the goal of the paper. The fourth part presents the results of the analysis for the case study. It shows two aspects – in use case 1 with lower number of observations and in use case 2 with high number of observations. The analysis shows also issues that should be regarded when making sentiment analysis. In the last part, there is a summary and conclusions.

The hypothesis in this paper is as follows: although human-sourced information is not a high quality data source, using them as an input for sentiment analysis can provide valuable information.

2. Theoretical background of sentiment analysis with Big Data tools

A model of a smart city can have a various forms. Usually it is related to automation of transportation systems, energy grids or public services that can be monitored to provide a reliable information for decision makers and other data analysts (technologyreview.com, 2015). In this form, it can address environmental, social and economic problems (publishersweekly.com, 2013). In very complicated forms, Big Data can provide a framework for social security administration (Krishnamurthy, Desouza, 2014). However using sensors and other forms of city monitoring does not cover all the aspects of smart cities environment.

The one of the possible appliances for the smart city is to use Big Data as a tool for text mining and machine learning purposes. This type of using Big Data has a really high potential for city authorities. One of the possibility is to use a sentiment analysis with polarity detection to provide an information on either positive or negative category of selected text (Khan et al., 2015). In many cases, sentiment analysis can be easily used for deep analysis of short messages, such as Twitter posts. Tweets can be analyzed to provide the data on people arguments or opinions on different news or activities (Grosse et al., 2015). To provide good results, in many cases a training dataset must be developed to ensure that the algorithm is identifying the text correctly (Vilares et al., 2015). Very important issue regarding large datasets is to provide good results of analysis, even the information provided is insufficient (Domański, Jędrzejczak, 2015). For this matter, we should make an algorithm for proper identification of different entities that are going to be analyzed. It is especially a challenge when dealing with unstructured datasets. It can be done by linking data from de-identified dataset with the data from other sources (Angiuli, Blitzstein, Waldo, 2015).

It is a very common opinion that data from social media can be classified as dark data. This type of data that is stored on Facebook, YouTube or Twitter, is not used effectively by business (Kimble, Milolidakis, 2015). This is the effect of the fact that there are several limitations in using such data sources. If you plan to use such data source, you

must be aware of legal aspects related to store and process information that is available on such social media websites. Another data source that has a high potential for businesses is represented by websites. It concerns especially the nature of user interactions on websites (Henry, Venkatraman, 2015). Analyzing the content of the webpage and social media is also a topic of interest for business intelligence and analytics – BI&A 2.0 (Chen, Chiang, Storey, 2012).

Big Data gave not only the possibilities of using various algorithms for data processing, but also various tools for data storing and make data analysis. It includes various Big Data compliant databases and data warehouses, such as Hive or Cassandra (Moorthy et al., 2015).

3. Case study

When deciding to develop the solution for sentiment analysis, we can choose between dedicated software that will be based on a set of keywords and text mining techniques. These techniques are developed for a specific purpose. In many cases a software can be created based on existing, well known open source libraries. In this case study, the decision was to choose a second one, which was to develop a solution based on existing libraries.

As mentioned in previous part, the case study included in this paper shows the potential of using human-sourced information. It includes posts on social media websites or comments on web portals regarding a specific news concerning self-government or city authorities activities and duties. To accomplish this task, the framework based on Apache Spark with Apache Hadoop was used. The source code was written in Python language. The framework used in this case study is presented in Figure 1.

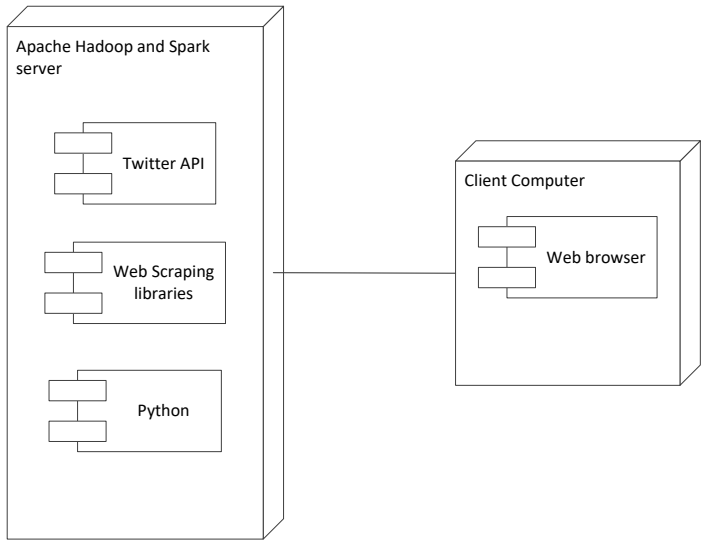

Fig. 1. The framework used in the case study
Source: own elaboration

As shown in UML deployment diagramin Figure 1, the server site is based on Apache Hadoop and Apache Spark. It consists of three extra packages that are Twitter API, Web scraping and extra packages for Python 3.5 language with libraries. The communication between the server and client computer is made via web browser.

First of all, it was necessary to indicate which websites are the most representative to conduct analysis on specific news or web articles. The first step was to find news articles on the initiatives performed by self-government institutions or cities authorities. It includes both regional web portals as well as any portals that publish such articles. In many cases, it is enough to receive the reliable results of the sentiment analysis of comments. The steps of analysis were presented in Figure 2.

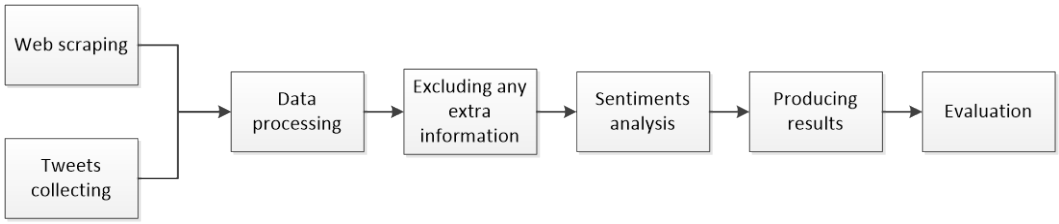

Fig. 2. Steps of analysis performed in the case study
Source: own elaboration

Steps used in the analysis are the same for each data source – Twitter or webpages. It does not matter how you gather the data – the processing, cleansing and analyzing phases are the same. Very important steps is to evaluate the model of machine learning that was applied for analysis. It is the last step of this framework.

For this use case, it was decided to use comments from one of the most popular web portal that publish lots of regional events. Then Twitter was also used to perform analysis on various events related to different actions in the city. It should be noted that web scraping is an important step when making analysis of comments from websites. It is not important when making analysis of Twitter posts, because it can be accessed via Application Programming Interface (API) that is described on Twitter webpage.

To provide good results of analysis, we need to tackle with lots of issues related to text mining and web mining. It includes especially elimination of duplicates as well as any stop words that are not important for the context of the post or comment. Although one data source will not result in many duplicates, using a dictionary of keywords (named stop words) that should be omitted when making text analysis is crucial. It ensures also that the information provided by the analysis will have higher quality.

It is not necessary to perform geolocation of comments when the news is published on a website, no matter if it is a regional web portal or even international. The goal is to identify a specific news and find a general opinion on the activity written in the article. In many cases, it is also impossible to identify a geographic location of the user that put a comment on a website. Twitter allows identifying a geographic location of the post that

was tweeted. It allows splitting population into regional and international. However, still when the tweet is on a specific action taken by local authorities, it is not necessary to identify the geographic location. Nevertheless, the framework used in this article can be enhanced by adding the geographic location attribute. Tests performed for the purpose of this paper show that it is still not necessary to include this attribute for analysis.

In this case study the results comes from two regional news on city activities regarding different events. The first news on activities in the city is rather neutral, while the second action taken in the city is very controversial. In the first article, there were just 19 comments under the news article. In the second case, 514 user reactions were recorded. In this case, study machine learning libraries for Python language were used to make sentiment analysis of user feelings on these two different types of activity related to the city. The number of comments can be extended by using a specific API to access Twitter comments related to specific news represented by a hashtag. However, Twitter is used by a selected group of people that may not be a representative group of users. Therefore, in part 4 of this paper, the presented results of analysis are based only on comments that were gathered from websites. Actually, there is no difference in machine learning algorithms for analysis of comments on webpage and tweets on Twitter. The goal is to find a reliable data source that will give promising results.

4. Results of analysis

4.1. First use case – 19 observations

As mentioned above, the first use case was to analyze a neutral news on action in the city. The results of the sentiment analysis are presented in Table 1.

Table 1. Results of sentiment analysis – use case 1

Type of comment	Number of comments
Neutral	12
Positive	5
Negative	2

Source: Own elaboration

Due to the fact that the first news article was on a neutral activity, we could expect the biggest number of positive comments related to other ones. From 19 comments, 12 were neutral, 5 positive and 2 negative. The machine learning tools indicate the probability of correctly identified comments of 1.0 for 11 comments. The detailed results are included in Table 2.

Table 2. Probability of correct identification of the comment – use case 1

Probability	Number of cases
1.000	11
0.980	1
0.936	1
0.902	1
0.851	1
0.834	1
0.787	1

0.743	1
0.157	1

Source: Own elaboration

Even most of the comments were identified with the probability of 1.0, manual analysis of comments shows that there were some mistakes in the results. Some of the explanations were included in Table 3.

Table 3. Possible mistakes in sentiment analysis – use case 1

Probability	Label	Part of comment
1.000	Negative	Meanwhile, nothing is said about the…
1.000	Neutral	…tortured and murdered…
0.157	Neutral	I wish I knew where it was. I would…
0.743	Negative	The professional strikes again…
0.851	Neutral	He was a nobility…

Source: Own elaboration

In Table 3 several different keywords were presented that could be wrongly classified by machine learning tool. The first case shows that the user is complaining on too small explanations in this specific news. It was classified as negative. In the second comment, there is a reference to the Second World War but the machine learning tool omitted this keywords and classified all the content of the comment as neutral, which in fact is a truth. The third comment has the smallest probability;therefore, it was classified as neutral. The fourth comment refers to a different case and in fact should not be classified as negative. The last one is neutral, however in many cases should be classified as negative, because this comment sounds like sarcasm.

Therefore, we can say that the first use case shows that too small amount of information (just 19 comments) includes in the analysis will not show the reliable results of the public opinion on a specific event or activity.

4.2. Second use case – 514 observations

In the second use case the number of comments that were analyzed was significant larger than in the first use case. First of all, the number of observation was 514. Selected data source did not have a geographical location provided with the comment, so it was not possible to split the dataset into regional attributes, as it is possible when working with Twitter posts. The results of analysis were presented in Table 4.

Table 4. Results of sentiment analysis – use case 2

Type of comment	Number of comments
Neutral	166
Positive	87
Negative	261

Source: Own elaboration

The results presented in the first table show that there is a high potential of using sentiment analysis for finding users opinion on specific events. Taking into account experiences from previous use case, the decision was to check for the probability of

comments correct identification. Table 5 shows the probability of identification of comments. Because of the size of the table, it is just a brief information and does not include all the probabilities identified in this use case.

Table 5. Probability of correct identification of the comment – use case 2

Probability	Number of cases
1.000	334
0.998	2
0.996	1
0.995	1
0.984	1
0.963	1
…	…
0.364	1
0.341	1

Source: Own elaboration

Based on the variety of probability, we can see that the machine learning tool was unable to identify everything correctly. As it was shown in the previous use case, the lowest probability means that it is not trustworthy. Therefore such comments, with lowest probability, should be classified as neutral. Nevertheless, some comments with low probability can be identified either as positive or negative. In Table 6, some examples of comments with its classification were shown.

Table 3. Possible mistakes in sentiment analysis – use case 1

Probability	Label	Part of comment
1.000	Negative	How people complain that this poor…
1.000	Negative	Correct –it is not in the position to…
0.973	Negative	Why do they expect any money at…
0.949	Neutral	Would they like it if a…
0.562	Neutral	World is now insane…

Source: Own elaboration

We can see, as it was in previous example, that some of the comments can be classified wrongly. Nevertheless, the results are better than for the dataset with smallest number of observations.

5. Summary and conclusions

In many cases analyzed comments may be negative just because of the fact that people that has arguments to the specific initiatives, prefer to express them anonymously via website. Although making analysis of comments on webpages is huge challenging, in many cases it can provide better results than analysis of tweets. It is just because posts on Twitter are very short and cannot exceed 140 characters. Such posts are usually shortened to some keywords and hashtags. In some cases, machine learning tools will not have a chance to identify correctly the specific phrase or sentence.

Still there is no possibility to identify sarcasms in the text. Moreover, machine learning tools are still not able to find a specific collocations of words regarding different

aphorisms. Therefore, very important and crucial step in developing a good solution for sentiment analysis is to provide a set of rules to tackle with problems related with identifying of aphorisms or sarcasms. On the other hand, there is still a need to control manually a set of techniques to work with lemmatization and stop words.

Small population that is an input for the sentiment analysis will not show the reliable results of analysis. It was identified in the first use case. Therefore, it is necessary to find such data sources in which the number of public opinions is rather high. All data sources that has little number of comments or posts should not be analyzed with machine learning tools. We have to bear in mind that most of regional websites for smaller cities do not have such extended possibilities for users to add comments than a typical, well-known Internet web portals.

The experience in working with Big Data unstructured repositories led to create a conclusion that each data source must be treated individually. It needs to create different lexicons and dictionaries for best identification of the context of specific text, such as website comments. For instance, specific web portals added some sort of typical keywords in the content of comments. These keywords were scrapped as a part of comments. It led to have too many words in the comment that in fact should be classified as stop words. Therefore, the strong conclusions is that stop words dictionary has not got an universal form. Although we can find stop words dictionaries for different languages on several different websites, they are quite different depending on the website, and may provide a wrong results if used for many different data sources.

To summarize the conclusions it can be said that the hypothesis from introduction was confirmed and we can say that using sentiment analysis for human sourced information can result in providing a valuable information for city authorities. However, there are lots of different obstacles that were described in this paper and should be considered when developing software to perform analysis for decision makers, including city authorities.

References

- *Smart Cities Will Take Many Forms*, MIT Technology Review, 118, 1, (2015), pp. 63-64, http://www.technologyreview.com, date: 02.11.2016.
- *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*, Publishers Weekly, 260, 33, pp. 59, (2013), http://www.publishersweekly.com, date: 02.11.2016.
- Khan, F., Qamar, U., Bashir, S.(2015), *Building Normalized SentiMI to enhance semi-supervised sentiment analysis*, Journal Of Intelligent & Fuzzy Systems, 29, 5, pp. 1805-1816.
- Grosse, K., González, M., Chesñevar, C., Maguitman, A.(2015), *Integrating argumentation and sentiment analysis for mining opinions from Twitter*, AI Communications, 28, 3, pp. 387-401.
- Vilares, D., Alonso, M., Gómez-Rodríguez, C.(2015), *On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages*, Journal Of The Association For Information Science & Technology, 66, 9, pp. 1799-1816.
- Domański, C., Jędrzejczak, A.(2015), *Statistical Computing in Information Society*, Folia Oeconomica Stetinensia, 15, 2, pp. 144-152.
- Angiuli, O., Blitzstein, J., Waldo, J.(2015), *How to De-Identify Your Data*, Communications Of The ACM, 58, 12, pp. 48-55.

- Kimble, C., Milolidakis, G.(2015), *Big Data and Business Intelligence: Debunking the Myths*, Global Business & Organizational Excellence, 35, 1, pp. 23-34.
- Henry, R., Venkatraman, S.(2015), *Big Data Analytics The Next Big Learning Opportunity*, Academy Of Information & Management Sciences Journal, 18, 2, pp. 17-29.
- Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., Ghosh, P.(2015), *Big Data: Prospects and Challenges*, Vikalpa: The Journal For Decision Makers, 40, 1, pp. 74-96.
- Krishnamurthy, R., & Desouza, K.(2014), *Big data analytics: The case of the social security administration*, Information Polity: The International Journal Of Government & Democracy In The Information Age, 19, 3/4, pp. 165-178.
- Chen, H., Chiang, R. L., Storey, V. C. (2012). *Business Intelligence And Analytics: From Big Data To Big Impact*, MIS Quarterly, 36(4), pp. 1165-1188.