Bucharest, December 8-9, 2016

USING SENTIMENT ANALYSIS WITH BIG DATA TOOLS TO ENHANCE KNOWLEDGE ON SOCIETY

JACEK MAŚLANKOWSKI

DEPARTMENT OF BUSINESS INFORMATICS, FACULTY OF MANAGEMENT

UNIVERSITY OF GDAŃSK, POLAND

AGENDA

- Prerequisites
- Framework
- Results of analysis
- Conclusions

PREREQUISITES

CHALLENGES AND OVERVIEW

THE GOAL OF THE STUDY AND GENERAL CHARACTERISTICS

CHALLENGES AND OVERVIEW

- Lots of noise in data

Structured
Unstructured
Semi-structured

Type of the data source	Noise
Machine Generated Data	Low
Process Mediated Data	Medium
Human Sourced Information	High

- Data linkage problems

ID
Subject
Attributes

Problem
Entity identification
Different attributes
Different context

- The data can be duplicated

Same opinion expressed by the same person lots of times

Causes
More and less active users
Limited population
Anonymousness

OVERVIEW AND THE GOAL OF THE STUDY

The goal is to present **suggested framework** for retrieving and processing information on public opinions on specific events and public reaction to different inititatives and campaigns.

Identify **value added** for the city authorities by making sentiment analysis on various social media posts as well as on comments from websites – what people think about government or self-government inititatives (positive, negative, neutral).

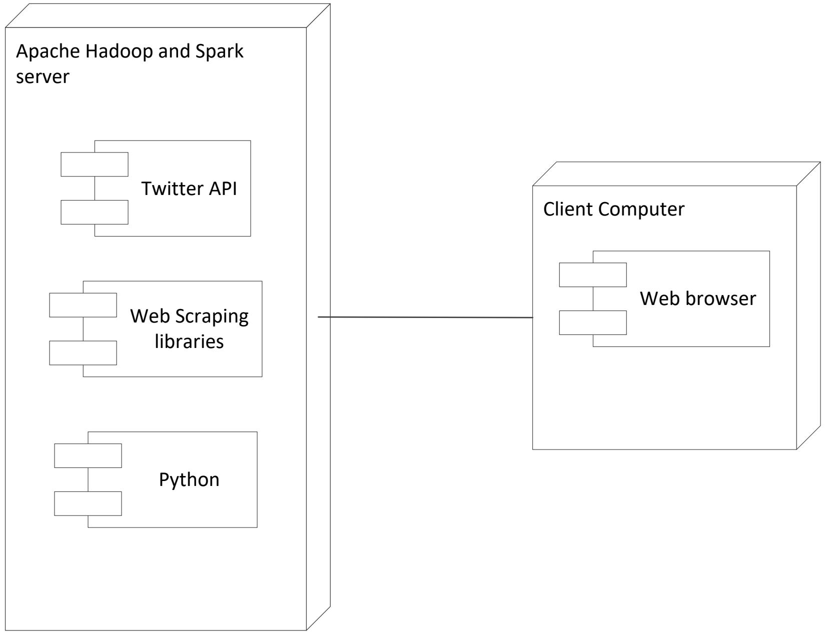
Although human-sourced information is not a **high quality data source**, using this source as an input for sentiment analysis can provide valuable information.

FRAMEWORK

SUGGESTED FRAMEWORK

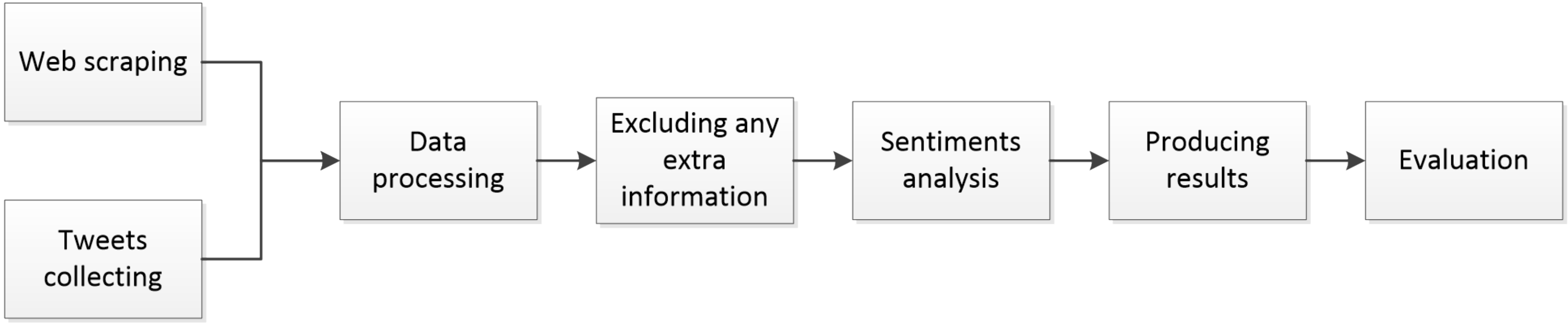
STEPS OF ANALYSIS

COMPONENTS USED



- Real time analysis
- Machine Learning tools
- Text Mining methods

STEPS OF ANALYSIS



RESULTS OF ANALYSIS

CASE STUDY 1

CASE STUDY 2

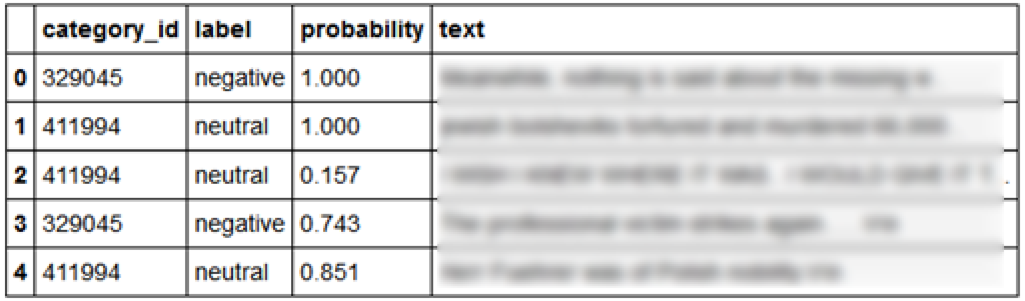
USE CASES

1st USE CASE

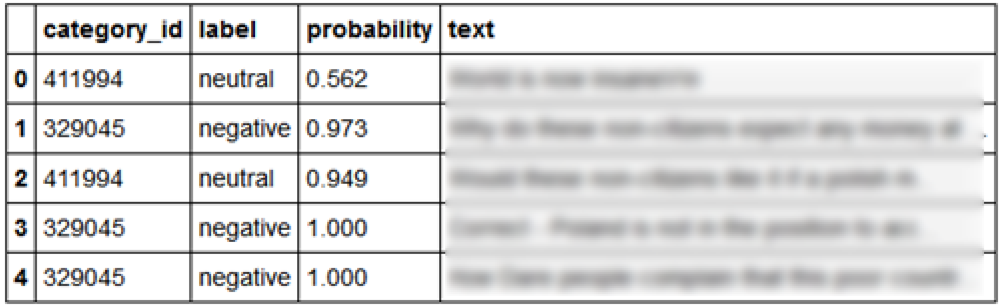
- 19 observations
- Rather neutral

2nd USE CASE

- 514 observations
- Rather controversial

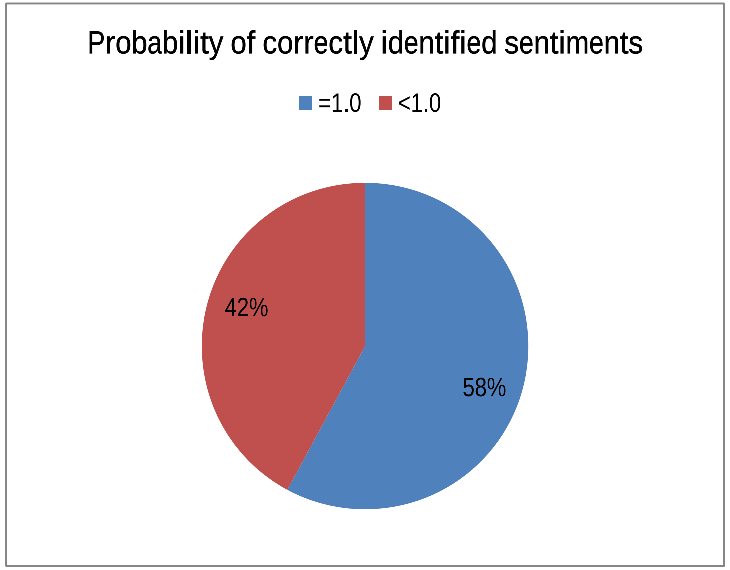


	category_id	label	probability	text
0	329045	negative	1.000	[redacted]
1	411994	neutral	1.000	[redacted]
2	411994	neutral	0.157	[redacted]
3	329045	negative	0.743	[redacted]
4	411994	neutral	0.851	[redacted]



	category_id	label	probability	text
0	411994	neutral	0.562	[redacted]
1	329045	negative	0.973	[redacted]
2	411994	neutral	0.949	[redacted]
3	329045	negative	1.000	[redacted]
4	329045	negative	1.000	[redacted]

USE CASE 1 – 19 OBSERVATIONS



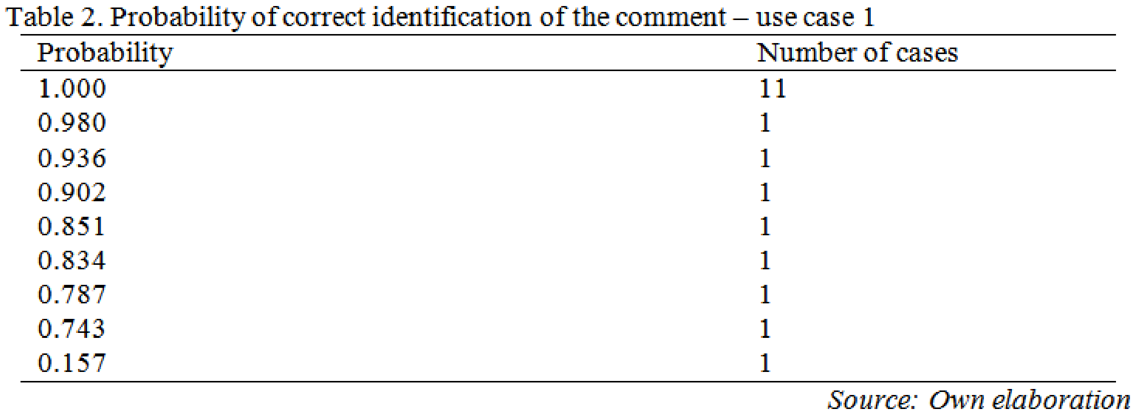


Table 2. Probability of correct identification of the comment – use case 1

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

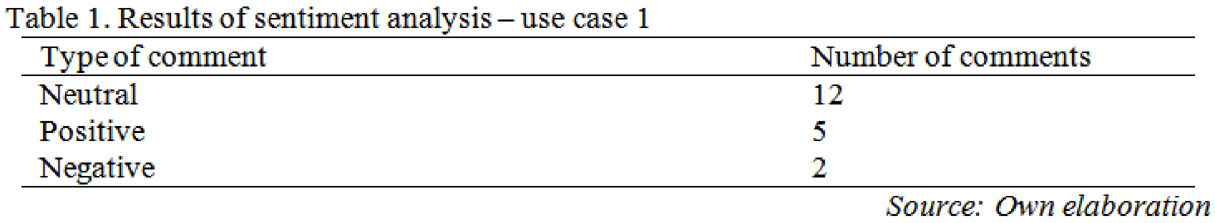


Table 1. Results of sentiment analysis – use case 1

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

Source: Own elaboration

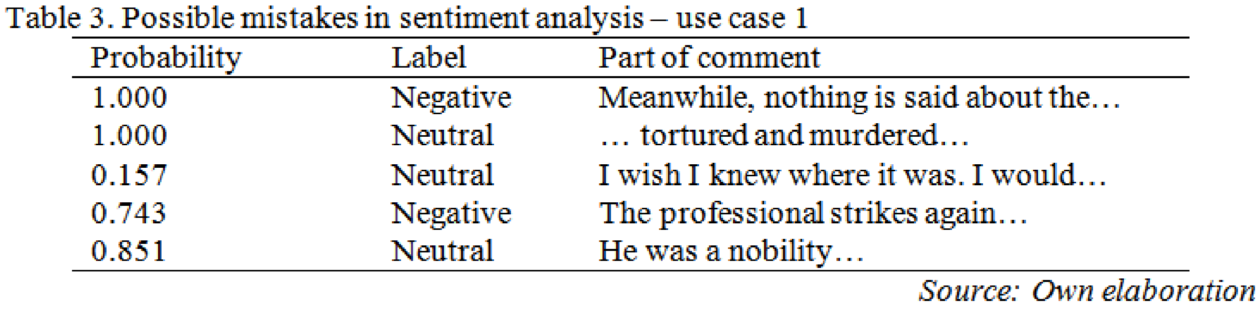
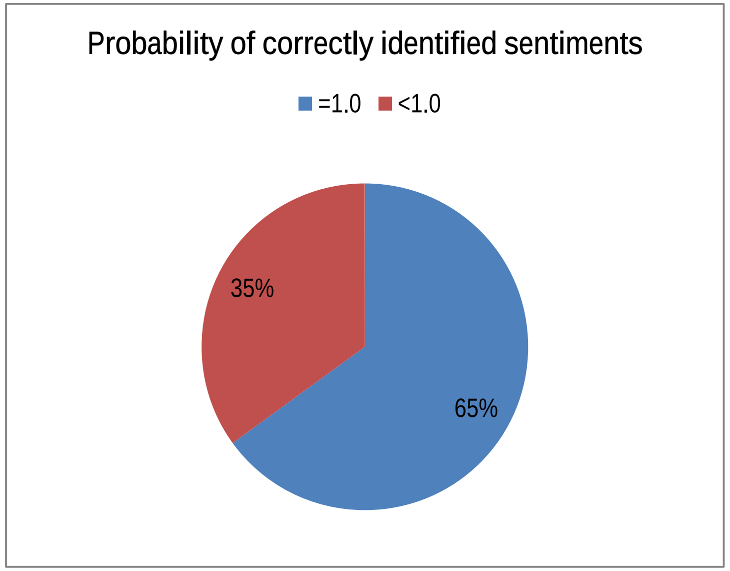


Table 3. Possible mistakes in sentiment analysis – use case 1

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

USE CASE 2 – 514 OBSERVATIONS



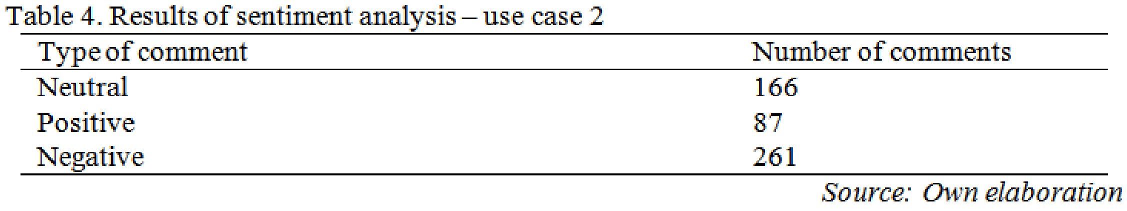


Table 4. Results of sentiment analysis – use case 2

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

Source: Own elaboration

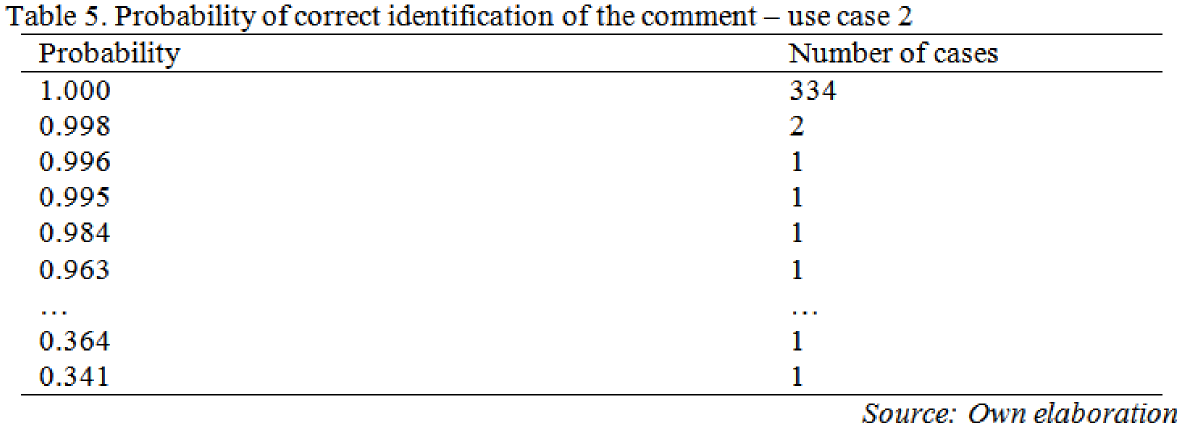


Table 5. Probability of correct identification of the comment – use case 2

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

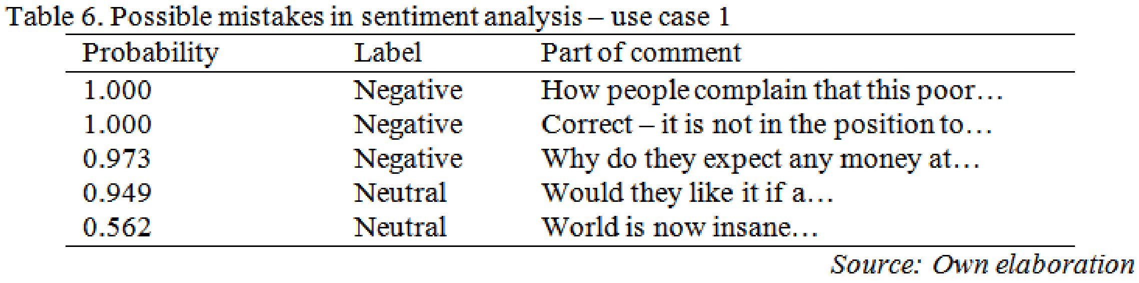


Table 6. Possible mistakes in sentiment analysis – use case 1

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

CONCLUSIONS

SUMMARY

FUTURE WORK

CONCLUSIONS (1/2)

Anonymously expressed opinion on websites

Posts on Twitter are not anonymous

Shorter text is better for sentiment analysis

CONCLUSIONS (2/2)

Aphorisms, sarcasms, lemmatization, stop words…

Small population will not show the reliable results of analysis

Each data source must be treated individually

FUTURE WORK

Testing the environment on larger datasets and in different areas.

Combining different datasets (structured, unstructured) into one repository.

Developing dictionaries on stop words and lemmatization.

THANK YOU!

JACEK MAŚLANKOWSKI

UNIVERSITY OF GDAŃSK

POLAND