10. Text and opinion mining for policymaking

10.1 Focus of the use case

Text and opinion mining are two methods that may assist policymakers throughout **all stages of the policy cycle**. The methods are versatile, and can assist the process in many ways; from getting feedback on different policies, creating a map of a current crisis, or shedding light on places where citizen's feedback is needed. These methods are not confined to one policy area, rather, they are tools that can be used across the board on **many policy areas and topics**.

Text mining is a method that allows finding trends in a large canon of text. The method assists in highlighting topics that might have been unknown by creating numeric indices. It can create summaries of the frequency of a word, clusters of words, trends, and the like.

Opinion mining and sentiment analysis refers to analyzing **positive or negative valences** around topics.[77] While the term 'sentiment analysis' is more commonly used in the commercial world, the term 'opinion mining' is commonly used in academia.[78] When broad interpretations are applied, 'sentiment analysis' and 'opinion mining' denote the same field of study: the analysis of subjective states. Opinion mining allows for the categorization of content to either binary values of positive and negative or scales of values such as very good, good, satisfactory etc.[79] This is done through algorithms that classify documents and search key words that underline a sentiment.

Text and opinion mining are based on **natural language processing**, a computational process that retrieves high quality information from texts by detecting of patterns and trends in a corpus. It can be undertaken on many types of text from different types of media. Examples are social media, online and offline newsletters and study reports, letters, blogs and other documents by experts and citizens.

The tools also offer visualisation of the data that can help the policymaker to understand the complex data. In addition, many research projects, including EU funded projects, are operating to improve text and opinion mining for economic and research purposes.

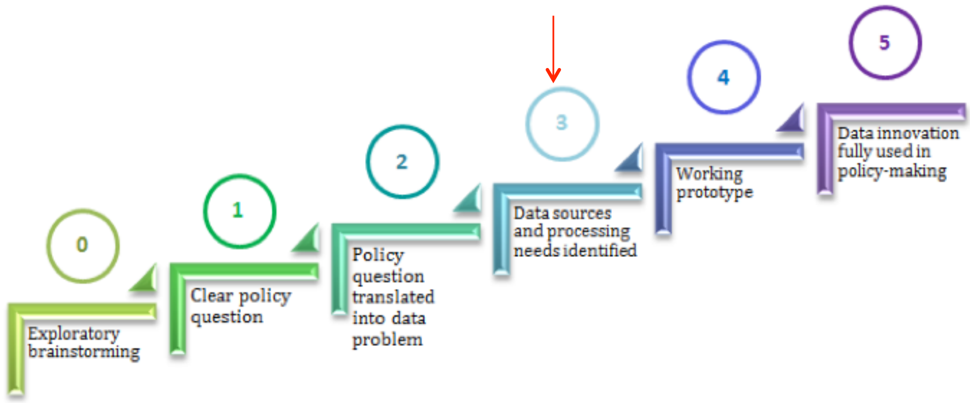
This case study is not focused on one policy area or based on one current experiments. Rather, it is a theoretical use case that explains the sources, analysis and outputs of text and opinion mining. Still, looking at the range of experiments, the maturity level can be assessed as 3 (see next page).

[77] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[78] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

[79] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combined approach." Journal of Informetrics 3, no. 2 (2009): 143-157.

Figure 18 Use case readiness level



10.2 The rationale

The analysis and results that are derived from text and opinion mining can shed light on **blind spots** throughout the policymaking cycle and, more in general, can complement and confront the results obtained by means of surveys, focus groups, etc. Text and opinion mining could help decrease 'false framing' that is not connected to the actual needs of the different stakeholders. Opinion mining might also assist in participatory policymaking. It can increase the input from the public and with the right settings, can engage citizen in the policy cycle. It may harness the 'wisdom of the crowd', rather of the wisdom of a single entity, potentially democratizing the policy process.

These are the ways in which these methods can assist the policymaking cycle:

- Collecting data for **framing** policy: by using text and opinion mining on social media networks, policymakers can gather information that can allow them to understand the stakeholders' needs and wants for a societal issues and an upcoming policy.[80]

- Creating a map of the current state of **opinion and satisfaction levels** from different groups of stakeholders: during the implementation stage of policies, opinion mining can help detect the satisfaction level from the policy interventions launched or the policies that have been adapted. This can inform policymakers in discussions about further improvements of policies.

- **Evaluating** the implementation of policies: halfway or at the final stage of the policy cycle, text and opinion mining can help summarise the feedback of stakeholders and to feed it again into the (re)design of policy interventions.

However, it is important to mention that text and opinion mining should not be used as the main source of evidence in the policy cycle. As this use case will show later, the sample of the data and data analysis abilities of both text and opinion mining can entail a bias towards specific populations or types of stakeholders. This should be acknowledged and policymakers should find ways to include other populations by using other, complementary methods.

[80] See example of Regional Energy plan: components and data sources for opinion mining. http://cordis.europa.eu/docs/projects/cnect/7/288147/080/deliverables/001-D91.pdf

10.3 The policy context

The methods in this use case are not bound to one policy area and can be used in many policy discussions. Moreover, text and opinion mining can be positioned in the context of the EU's Better Regulation program[81] that aims to improve EU policies and laws so they will achieve their objectives at minimum costs. It ensures that policy is prepared, implemented and reviewed in an open, transparent manner, informed by the best available evidence and backed up by involving stakeholders.

Several prototypes of text and opinion mining are developed in the Better regulation program. Text mining is used as part of public consultations, such as the Digital4EU project.[82] Policymakers can decide when to use these methods to get feedback from different stakeholders, either by asking for active feedback and ordering a survey, which will then look for different sentiments using opinion mining or look at passive feedback coming from social media and other internet sources. The European Commission already monitors the media in the EU news explorer and runs consultations over the internet.[83]

The Commission is the entity that runs these analyses and incorporates the feedback to its policy cycle. In addition to benefits such as increased information and feedback, the Commission acknowledges the intrinsic value of stakeholder engagement and constructive dialogues between citizens, firms, policymakers and other actors.

10.4 The data process: from data collection to analysis and visualisation

Data sources

The data sources for text and opinion mining are usually public web sources such as media outlets, blogs, forums and social media. The data are being scraped (automatic scripts that download data) or accessed from an API (Application Programming Interface) and then analysed. This makes the data available and accessible easily. Text and opinion mining can be done on any corpus, and the data can be any textual source, from books and magazines to surveys and social media.

The selection of data sources should take into account the stakeholders of a specific policy issues or policy intervention. Analysis of social media for example, can exclude some populations, like the elderly or lower class, from participating in the process, since they are not using social media networks (or using them to a lesser extent or for different purposes). Therefore policymakers should be clear about the population they want to include or want to hear from. Policymakers and data scientists should determine which sources and methods to use; for example traditional surveys, online surveys or passive methods such as web scraping. While surveys can help in actively seeking feedback from segments of the population that are not publishing their views online (e.g. in Tweets), the survey approach also has its limitations. There can be gaps and inconsistencies in the data since people tend not to answer full length answers in surveys or have incomplete answers that can burden the algorithms.

Also note that specific social media networks have specific user populations. Twitter for example, in most European countries represents only 20% of the population and can be considered to be a tool of the elite.

Policymakers should make sure that the web sources that are targeted for use are suitable for the process legally. Many sources on the web are prohibited from re-use of the data and some of them might contain private information. Therefore, there is a need to check that the sources can be used for these types of analyses. In many cases, the raw, existing data that is mined continues to be owned by the original authors and

[81] http://ec.europa.eu/smart-regulation/index_en.htm
[82] https://ec.europa.eu/futurium/en
[83] http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html
http://ec.europa.eu/yourvoice/consultations/index_en.htm

platforms, while the results of the text and opinion mining analysis can be owned by the policymaker that funds and oversees this process. In publishing the results, anonymisation and aggregation of data are among the points of attention.

In addition, since sentiment analysis is dealing with opinions, it is important to acknowledge the legal and cultural aspects of freedom of speech. The data analysis or the interpretation of data could filter out comments or opinions that are very insensitive, that could be seen as hate speech, or that could even lead to physical abuse. Part of the filtering process could be a number of interventions to address and stop these types of comments.

The data collection process

Data can be collected in two ways: seeking active input/feedback and using the passive/listener mode. This second mode means collection of data that is user generated and is published freely and openly on the web, such as social media and blogs. In the case of active input/feedback, there are additional steps such as composing questionnaires, distribution of the survey and collection and prompting users such as moderating online discussions.

Policymakers and data scientists should collaborate in designing the set of questions for actively seeking active or the set of questions for the passive mode ('what to look for in the data'). Deciding on the set of questions is closely linked to deciding on the data sources and checking whether the main types of stakeholders can be engaged by the text and opinion mining approach that is foreseen.

There are many start-ups that offer services for public consultations and opinion mining. One of them is Insights. The Insights service allows people to submit their opinion. It also informs persons about the next steps in the consultation process and how the consultation affects the policy cycle. These updates are available on the online platform but also via email and mobile text notifications. Taking this approach, citizens and other contributors might be more engaged and hopefully will help to make other citizens engage with the consultation process (and future consultations).

Data analytics and visualisation

Text and opinion mining use high-level computer algorithms. Mining is sometimes considered part of Artificial Intelligence or of machine learning. It is a growing field that is still evolving and growing rapidly. Natural language processing (NLP), the main field behind the analyses, is very strong in the main European languages - English, German, Portuguese and French. However, the EU has 24 official and 60 indigenous languages. For some non-common languages, the capabilities of NLP are still not strong and need further work. A boost to NLP is needed in small markets where NLP is not economically viable for private companies, and where academia does not always have the capacity to develop it.

In addition, NLP still encounters difficulties in terms of understanding the cultural context of a sentence, detection of sarcasm, and typos. All of these are hard to comprehend as a human user, and difficult to detect by an algorithm. In order to overcome these challenges, the analysis of the results usually shows not only the main opinions and common phrases, but also the confidence of the algorithm in the results and any outliers that came up in the analysis.

Visualisation of text and opinion mining is mostly done by means of standard charts and graphs. Examples are counting of words and plotting positive and negative statements on a grid. In addition, word clouds are used. These have become one of the most common visualisation tools for text mining and, to a lesser extent, opinion mining. Word clouds present the frequency of words in a text. The bigger the font of the word, the bigger the frequency of the word. This type of visualisation can help to understand the main concepts, concerns, opportunities, etc. in an easy and simple way. Often, it is a starting point for a more detailed discussion of results.

An Example for a word cloud on main challenges for new EU projects - generated during the 2nd CESBA SPRINT Workshop

10.5 Reflections on challenges and next steps

Text and opinion mining can assist policymakers to understand the opinions of citizens and other stakeholders and to incorporate this feedback in the policy process. Text and opinion mining tools are becoming mature and accessible and can be implemented in a short amount of time. They are valuable tools that can be used during all stages of the policy cycle.

Since text and opinion mining often relies on social media and other web content that can have personal data, the data analyst should make sure that data is anonymised before it is being analysed. In addition, as mentioned before, different web platforms represent different populations. Policymakers and data scientists need to take into consideration the different biases that each social media platform has; mitigate this by combining data sources (online and/or offline); and present any bias as part of the report.

Some web content is under copyright license and does not permit the use of this data by third parties. The European Commission has had a working group that examined how to use and increase the use of data in the public domain for text and opinion analysis and other types of research. The emphasis of the discussion was on academic research.[84] This type of discussion should take place frequently to examine the use of text and opinion mining not only for academic research, but also for policymaking. One of the main challenges is to increase the number of data resources that can be used, in technical and legal terms.

Another topic that should be considered is language. While both text and sentiment analysis are strong in the main European languages, other languages are lagging behind. Since the European Commission needs to engage in its policy process stakeholders from many language groups, there is a need for improving text and opinion mining tools in all European languages. The Commission already invested in developing these tools, but should consider future research on the topic.[85]

Since the tools for text and opinion mining are widely available (and used) in the private sector, there are opportunities to adapt or scale these tools for use in a policymaking context. As such, these tools can be used by policymakers in different policy areas and at different policy levels, e.g. national and European and

---

[84] https://ec.europa.eu/licences-for-europe-dialogue/en/content/text-and-data-mining-working-group-wg4

[85] http://cordis.europa.eu/project/rcn/103841_en.html

international. Policymakers should consider employing personnel that can select and handle these tools, collect and analyse the data, and explain the results to policymakers and other stakeholders. Having staff with the right skills is only relevant for using the tools effectively, but also to contribute to improving the algorithms and for addressing that the tools (and data sources) succeed in covering the relevant stakeholders. This will further increase the inclusiveness of the policy process, at different stages of the policy cycle. In addition to training or hiring staff (e.g. experts that worked in research, marketing or social media) it is important to increase the awareness of text and opinion tools mining among policymakers at large. In order to increase awareness of the strengths and weaknesses of these tools, policymakers could consider internal workshops, webinars, online courses and participating in text and opinion mining projects of colleagues from other policy areas.

Further reading

Thelwall, Mike. “Society on the web.” The oxford handbook of internet studies, Dutton, William H., ed. Oxford University Press, 2013.

Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.

Aggarwal, Charu and Tarek Abdelzaher, *Social Sensing*, Chapter 9 in Managing and Mining Sensor Data, Springer 2013.

Ceron, Andrea, Luigi Curini, Stefano M. Iacus, and Giuseppe Porro. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France." New Media & Society 16, no. 2 (2014): 340-358.

Ceron, Andrea, and Fedra Negri. "The “social side” of public policies. Using sentiment analysis to support the action of policy-makers across the policy cycle." http://www.icpublicpolicy.org/conference/file/reponse/1435680811.pdf

A number of **commercial solutions** exist to provide big data solutions for information mining in support of better signal detection systems. A non-exhaustive list includes:

- Recorded Future – Web intelligence platform focusing on historical trends to help formulate hypothesis on future developments
- DataSift – Real-time and historic data analytics on social media, blogs and news items. Topic and trend discovery, classifiers and demographics analysis, opinion& sentiment analysis
- Brandwatch – Online dashboard based on Twitter data news, blogs, etc., trend and influencers discovery, sentiment analysis
- Semantria - Service, focusing on sentiment analysis, API version or Excel-based extraction
- Sentiment Viz – Sentiment analysis in real-time; open source; developed by North Carolina State University

In addition, a number of **current European research projects** are exploring the same domain, most notably, Social Sensor, Newsreader, X-LIKE, EUMSSI and PHEME