Using “Text Mining” Analysis for the Assessment of the Health Quality of Dietary Supplements

Kacper Wróbel¹, Anna Justyna Milewska²

¹ Doctoral studies at Faculty of Health Sciences, Medical University of Lodz, Poland

² Department of Statistics and Medical Informatics, Medical University of Białystok, Poland

Abstract. Techniques of text data analysis have been known for many years and commonly used in many areas of life. Text mining enables, among others, the acquisition of information from the text, its filtering, and studying of similarities and relationships. The aim of this paper is to design a method that would make it possible to assess the health quality of dietary supplements, on the basis of text mining techniques. A fictional plant-based product was used in the study, which was compared with other products containing at least one of the tested ingredients registered in the years 2007–2019 in the register of dietary supplements kept by the Chief Sanitary Inspectorate (GIS), which were given either the “consistent” or “to be clarified” status. The obtained results concern the frequency of occurrence of the individual ingredients (St John’s wort/*Hypericum*, melissa, rose root/*Rhodiola*) in other products, considering their status in the register. The data thus obtained was subjected to classical statistical analysis in order to find correlations between the presence of a given ingredient and the product status. In view of the obtained results, the text mining analysis may be considered as a helpful tool in the process of internal risk assessment performed by manufacturers of dietary supplements.

Introduction

In the legal sense, dietary supplements were defined for the first time in 2001 in the Act on sanitary conditions of food and nutrition. At that time, the group of food products was presented as concentrated sources of vitamins or minerals as well as other nutrients. A single ingredient or several ingredients in combination could be present, comprising composite products in the form of, e.g., capsules, tablets, or sugar-coated tablets. Their role was – and still is – supplementing nutrients consumed as part of normal diet. (The Act of 11 May 2001 on sanitary conditions of food and nutrition, 2001). Depending on the context, the place of dietary supplements

within a given area of science may vary. From the point of view of the production process, dietary supplements can surely be partly classified as part of the field of food and nutrition technology; however, according to the form and the ingredients used, they belong to the discipline of pharmaceutical sciences (Silano et al., 2011). On the other hand, considering the health security of consumers related not so much to the production process but rather to the character of the ingredients used in dietary supplements, the products in question remain within the scope of interest of health sciences and medical sciences (Restani et al., 2016).

Ever since the discussed group of health-oriented products appeared on the Polish market, they have been unwaveringly popular among consumers. The register of dietary supplements and health-oriented products consists of over 113,000 items, of which 18% were registered in the period from 1 January to 8 November 2020, which indicates a significant increase of new products in those recent months. Since 2014, the year-on-year growth has always exceeded 100% (Główny Inspektorat Sanitarny, 2020).

Considering the fact that the process of selection of ingredients for the production of supplements is characterized by a high level of flexibility, control over the products in circulation may prove difficult. The current legal status, apart from certain clearly defined exceptions, does not specify which substances may or may not be used, with the whole burden of ensuring product safety resting with the manufacturer or the importer (Petkova-Gueorguieva et al., 2019). Practice shows, however, that manufacturers' attitudes do not always comply with good practices or are at odds with social responsibility (Mathioudakis, 2005). Notably, according to data published by the State Sanitary Inspection in 2019, the 4,125 samples taken for laboratory tests for various purposes constituted less than 27% of the 15,412 new products notified in that year. This rate of market development, especially in the e-commerce sector, alongside the limited possibilities on the part of official food control bodies, necessitates the search for new and more efficient methods of market control and selection of products and entities for control, as well as a tool supporting the process of internal control used by the entrepreneurs themselves. This need was also recognized by the Supreme Audit Office, after a countrywide control in Poland was performed in 2017 (Najwyższa Izba Kontroli, 2017). A method that may potentially be used for analysis of such large databases is the "text mining" analysis.

Text data describing marketed products may be seen as an unordered set that requires the use of analyses other than the classical methods. After appropriate preparation and application of the dedicated tools, however,

the data can be effectively explored and subjected to further analysis with the use of standard methods (Weiss et al., 2010).

The aim of this paper is to test the possibility of using the “text mining” method as a tool for the initial assessment of dietary supplement formulas in terms of the contents of banned or potentially dangerous ingredients.

Text Mining

Text mining is a method of processing of data sourced from unstructured databases. Although techniques of text analysis have a long history of use (Weiss et al., 2010), their importance grew when access to large data sets was becoming widespread (Aggarwal & Zhai, 2012). Unlike the classical methods of statistical data analysis, text analysis does not fall into a single defined pattern and may be performed in accordance with various schemes, depending on the software used and the original language of the source data (Gupta & Lehal, 2009). Among the most popular techniques of text analysis there are: (1) information retrieval, (2) text categorization, and (3) grouping (Dang & Ahmad, 2014). In the most general approach, each technique is based on the simple assumption that text data can be encoded in numerical form, and so converted, subjected to further tests. In its basic scope, text mining makes it possible to obtain information about the incidence of certain words or phrases in documents (Weiss et al., 2010); it also makes it possible to compare documents and search for similarities between them (Vijaymeena & Kavitha, 2016). Each single database record (case) can be called a document (Karl et al., 2015).

Practical implementation of the text mining method involves, among others: analysis of replies to questionnaires with open-ended questions (Karl et al., 2015), automatic classification of messages (a solution known from anti-SPAM systems) (Khan & Qamar, 2016), and detection of trends and correlations in documents (e.g. patients’ medical histories) (Zhou et al., 2006).

However, the character of data analyzed when the discussed techniques are used requires that the data should be prepared in a way that would ensure the highest effectiveness of analysis. The process of data preparation can be called tokenization. Tokenization involves, among others, determining the list of characters and words excluded from the analysis. Such characters are space and other punctuation marks, as well as conjunctions, whose presence in the text as a rule does not carry any substantial meaning. Although for a person fluent in the language of the analyzed text the process

is not complicated, the implementation of this step at the systemic level may in some cases lead to undesirable results (Weiss et al., 2010). To illustrate the point, a fictional example of the phenomenon in question may be given: processing the name of the chemical compound *3-monochloropropane-1,2-diol*. In the case of the analysis of a general text, the comma and the hyphens may be indicated as separating characters, not subjected to the analysis. However, taking this approach in the analysis of a text in the field of organic chemistry will result in artificial separation of the following elements: 1, 2, 3, monochloropropane, diol, which – presented in this form – will lose their meaning and scientific value. A logical conclusion of this issue is that the range of work required to be performed as part of tokenization – as well as the further subsequent steps – will largely depend on the original language of the source text.

An equally important action that constitutes part of text normalization is lemmatization, i.e. bringing various forms of the same word to its dictionary form. At this stage, the decision should be made as to how radical this procedure is to be, with at least two possible solutions (Plisson et al., 2004). In order to illustrate Plisson’s considerations, it can be shown that one of the variants is to create a type of dictionary of synonyms of the search words; for example, the analysis of a register of dietary supplements in search of information about the number of products containing wild rose requires at least two variants to be searched: *dzika róża* (Eng. wild rose; denominator), *dzikiej róży* (Eng. wild rose’s (fruit, extract); genitive). A more radical solution is to reduce both words to its roots. In practice, this would mean programming the system in such a way that, for the analyzed example, the algorithm would search for the following phrases: “dzik” and “róż”, which are common for all inflected forms of the searched words. Such a solution is surely less time consuming for the researcher, but it has obvious disadvantages; for instance, words with different meanings but the same root, e.g. *zielenice* (Eng. *chlorophyta*, i.e. the taxon that includes sea algae) and *ziele* (Eng. herb, i.e. part of a plant) – root “ziel” will be qualified identically.

Solutions based on elaborate dictionaries and algorithms exist which recognize text contexts and thus enable fast and reliable assessment of texts, without the need for a time consuming preparation prior to import to the program (Erhardt et al., 2006). However, there are limitations to the use of automated solutions of this kind (Leser & Hakenberg, 2005), which are nonetheless beyond the scope of this paper.

Despite all the limitations discussed above, text analysis is a tool that finds its use in many disciplines and areas. An excellent example of unstructured databases are without doubt social media, which are a practically

limitless source of information. Text mining methods were used in studies by Koh and Liew (2020) that concerned ways of speaking about loneliness during the COVID-19 pandemic.

The study was carried out on the basis of an analysis of 4,492 posts selected from Twitter according to specific criteria. The study included posts in English posted in the period from 1 May to 1 July 2020, which included the words “loneliness” and “COVID-19”. Taking into consideration the rules for text normalization, the researchers also searched for posts that included the word “lonely” as well as “COVID”, “COVID19”, “coronavirus” and “corona virus”.

As part of initial text preparation, all sentences were divided into separate words, whereas punctuation marks, conjunctions, and other components that do not bear information were deleted. Inflected forms of the same word were changed into their dictionary forms. Moreover, in order to minimize the so-called statistical noise, all the words that appeared in fewer than 50 database records were excluded.

In the first phase of the study proper, whose aim was to extract trends, grouping of words that exhibited a tendency to co-occur was performed. In order to narrow down the possible trends, 3 other variables were included, i.e. the continent from which the post originated, the date, and the number of followers of the post’s author. On the basis of an analysis of the variables, the algorithm indicated 3 trends, which the researchers titled themselves on the basis of keywords returned by the algorithm and example posts.

Analysis with the use of text mining methods made it possible to divide 4,492 posts into three basic thematic groups: (1) Community impact of loneliness during COVID-19, (2) Social distancing during COVID-19 and its effects on loneliness, (3) Mental health effect of loneliness during COVID-19. After an analysis which took into consideration the other variables included in the study, the authors came to several conclusions, including the main one that corroborates earlier studies, proving that loneliness is a multidimensional experience.

Materials

The study uses an excerpt from the register of dietary supplements in .xls format (Główny Inspektorat Sanitarny, 2019) as at 13.12.2019, kept according to the procedures and rules set in art. 30 pt. 5 of the Act of 25 August 2006 Act on sanitary conditions of food and nutrition (Ustawa z dnia 25 sierpnia 2006 r. o bezpieczeństwie żywności i żywienia, 2006). This

list constitutes the base spreadsheet in all the subsequent steps. The registry includes the following elements: no., name, date of notification, form, proposed qualification, qualitative composition, notifying entity, manufacturer/country, result of proceedings, type of special food product, and comments. All 87,223 items on the register that included the aforementioned information about dietary supplements and other health-oriented food products registered by entrepreneurs in the years 2007–2019 were qualified for the first stage of the study.

Statistica 13 (Dell Inc.) was used for further analyses, particularly the “text miner” function; chi-square tests were also performed to assess the obtained relationships. Statistically significant results were those at a level of $p < 0.05$.

The method’s concept involves performing initial qualification of a product from the group of dietary supplements as potentially consistent or potentially inconsistent with the food regulations in force. The assessment is performed on the basis of the qualitative composition of the tested product based on the analysis of the registry of health-oriented products registered in previous years and containing at least one of the ingredients in question.

A fictional dietary supplement envisioned as a mildly sedative preparation for adults was used as the tested product. The qualitative composition of the product was as follows: melissa (Pol. *melisa*) leaf extract, St John’s wort (Pol. *dziurawiec zwyczajny*) herb extract, *Rhodiola rosea* (Eng. rose root, Pol. *różeniec górski*) root extract.

Results

In order to meet the aim of the study, at the first stage a one-off adjustment of the base spreadsheet must be performed, consisting in adopting a customized method of presentation of information about product status (column: result of proceedings). An initial overview of instances of this variable, performed using the filtering option available in Excel 2010 (Microsoft), showed that information about the result of proceedings may take one of several tens of variants, with a large number of them in the form of descriptions and unstructured notes. Customization was performed in Statistica 13. In the first step, all cases of the variable in question were inscribed as a continuous record by deleting spaces using the “find and replace” method. This action was necessary for the next step of the analysis, i.e. retrieval. Tools from the Data Mining/Text Mining package were used for this purpose. The following parameters of the text retrieving function

were used: (1) text variables: column 9, (2) % of files where words occur: Min 1 Max 100. The other parameters were set to default. After indexing the algorithm returned two results: (1) *pwt-postępowaniewtoku* (i.e. proceedings pending), the status that occurred in 63,670 items of the register and (2) *s-suplementdiety* (i.e. dietary supplements) for 9,457 items. The statuses mean that (1) the body (a national food safety authority) raises legal concerns in relation to the product in question, or (2) there are no concerns in relation to the product. Hence, it was decided that only these statuses would be used for further analyses, thus reducing the number of cases in the spreadsheet to 73,127. An additional variable was created in the base spreadsheet presenting information about product status, according to the following scheme: 0 – no concerns, 1 – potential inconsistencies. Prepared in this manner, the file may be used for retrieving information and extracting patterns of relationship between the qualitative composition and the legal status of a product.

At the following stage of the method, all the cases in the register that contain at least one of the three main ingredients of the tested product must be found. This action was performed with the use of the “text mining” functionality in Statistica. The following parameters of the text retrieving analysis were used: (1) text variables: column 6, (2) % of files where words occurs: Min 0 Max 100, (3) inclusion words: *melisa* (Eng. melissa; denominator), *melisy* (Eng. melissa’s/melissa/of melissa; genitive), *Melissa* (alternative spelling), *dziurawiec* (Eng. St John’s wort; denominator), *dziurawca* (Eng. St John’s wort’s/St John’s wort/of St John’s wort; genitive), *Hypericum*, *różeńiec* (Eng. rose root; denominator), *różeńca* (Eng. rose root’s/rose root/of rose root; genitive), *rożeniec*, *rożeńca*, *różaniec* (common misspellings), *Rhodiola*. The “inclusion words” list included the denominator and genitive cases of the generic name of the main ingredients as well as the Latin name. The species name (species epithet) was omitted, as well as information concerning the part of the plant (*liść*, *ziele*, *korzeń* (Eng. leaf, herb, root)) and the form of ingredient (*ekstrakt/wyciąg*, Eng. extract). In addition, in the case of “różeńiec”, common misspelt variants of the generic name appearing in the register were included in the analysis, i.e. *rożeniec*, *rożeńca*, *różaniec*. The obtained results are presented in table 1.

The results were added, in the form of variables, to the base spreadsheet in binary form, i.e. 0 – ingredient not present, 1 – ingredient present in the product. At the same time, individual grammatical and lexical variants of the individual variables were consolidated to be presented as a single variable, e.g. the variable “melisa” includes the following variants: melisa,

Table 1. Frequency of occurrence of inclusion words

Word/Phrase	Count	Files
melisy	738	726
różeńca	346	343
rózeniec	295	292
melisa	288	286
rhodiola	137	136
dziurawca	94	94
dziurawiec	83	81
melissa	27	26
różaniec	26	26
hypericum	7	7
różeńca	4	4
rózeniec	2	2

melisy, melissa, etc. *per analogiam*. The described modifications resulted in obtaining the output data presented in tables 2–4. Statistically significant relationships between the product status and the presence of an ingredient were obtained.

Table 2. Relationship between the occurrence of the “dziurawiec” ingredient and its status

Status	DZIURAWIEC		TOTAL	chi-square test
	[0]	[1]		
no concerns	265 (18.1%)	7 (4.5%)	272 (16.8%)	$p < 0.001$
potential inconsistencies	1198 (81.9%)	147 (95.5%)	1345 (83.2%)	
TOTAL	1463	154	1617	Fi = 0.106

Table 3. Relationship between the occurrence of the “melisa” ingredient and its status

Status	MELISA		TOTAL	chi-square test
	[0]	[1]		
no concerns	38 (5.3%)	234 (26.0%)	272 (16.8%)	$p < 0.001$
potential inconsistencies	679 (94.7%)	666 (74.0%)	1345 (83.2%)	
TOTAL	717	900	1617	Fi = -0.275

Table 4. Relationship between the occurrence of the “różeniec” ingredient and its status

Status	RÓŻENIEC		TOTAL	chi-square test
	[0]	[1]		
no concerns	229 (24.3%)	43 (6.4%)	272 (16.8%)	$p < 0.001$
potential inconsistencies	713 (75.7%)	1345 (93.6%)	1345 (83.2%)	
TOTAL	942	675	1617	Fi = 0.237

Discussion

On the basis of the obtained results it can be concluded that the used research method was selected appropriately. Verification of input data with the use of text mining made it possible to exclude over 16% of cases from further analysis based on the value of the text variable informing about the status of the product. Interestingly, many authors draw attention to the importance of data preparation before commencing text mining analysis (Munková et al., 2013), whereas it turns out that the text mining technique itself may be used for initial preparation of text databases prior to their further analysis (Tunali & Bilgin, 2012).

The appropriate selection of parameters of text analysis concerning the frequency of occurrence of individual searched words (from 0% to 100%) made it possible to include all cases meeting the search criteria in the analysis, thus increasing the reliability of the obtained results.

Moreover, the results indicate a certain characteristic of the source data. In all the three tested cases, no less than 75% of products, regardless of the lack of the searched ingredients in their compositions (*dziurawiec*, *melisa*, *różeniec*), took “1” as the value of the “status” variable, i.e. they raised concerns as to the consistency of their compositions with regulations. This stems from the fact that on average a typical product is composed from more than one ingredient, with ingredients other than the searched ones possibly influencing the final status of the product. This phenomenon may have a negative impact on the substantive value of the final results, regardless of the used analytical tools.

However, regardless of the aforementioned issues, the text mining functionality retrieved text data effectively and illustrated their presence or lack of it in subsequent cases of a binary database. It is impossible to perform such an action on large databases, which is doubtlessly the type of database in the described case, using the popular spreadsheets in a time as short

as when text mining is used. Similar conclusions were achieved by Harmston et al. (2010) in their studies concerning the use of text mining in genomics, indicating the higher potential of the category of methods in question in the analysis of large data sets compared to manual methods. The advantage of the discussed technique was also recognized by Sahadevan et al. (Sahadevan et al., 2012) in their studies concerning genomics of farm animals.

The obtained results may be analyzed with the use of typical statistical tests. As Dang and Ahmad (2014) rightly indicate in their review, basic text mining techniques make it possible to obtain structured analytical data quickly and with low costs and effort. For the study described in this paper, the chi-square test of independence was used for further analysis, which highlighted the type of relationship between the status of a product and the presence of a given ingredient.

The results indicate the existence of a correlation between the tested ingredients and the product status. It can be concluded that the presence of St John's wort or rose root may have a negative influence on the health quality of a product, whereas in the case of melissa, using the ingredient carries a significantly lower risk of this kind.

Conclusions

In conclusion, it can be stated that the analysis of text data included in the register of dietary supplements with the use of text mining techniques returns valuable analytical data. A particularly important aspect of the designed method is the possibility to record information about the presence of a searched fragment or lack of it in the whole text in binary form. A considerable advantage of this functionality is its speed, also in relation to extensive databases.

The procedure described in this paper made it possible to estimate which of the tested ingredients are high-risk ones as far as the product's potential inconsistencies with food law regulations are concerned. Detecting these ingredients as early as at the stage of product design would enable entities in the food industry to make the necessary adjustments in product assumptions before it is placed on the market. These adjustments may include, among others, an in-depth analysis of scientific data in relation to a particular ingredient, adjusting its dosage, placing appropriate warnings on product labelling, or, in extreme cases, removing the ingredient from the formula. Undertaken during the conceptual design phase, such actions may prevent the entrepreneur from an investigation initiated by the Chief

Sanitary Inspector after the product has been offered for sale. This would entail temporary suspension of the product’s placement on the market, or – if the inconsistency is confirmed – even its withdrawal from the market.

The above leads to the conclusion that text mining may constitute a useful tool in the process of internal assessment of risk performed by manufacturers of dietary supplements, as well as those national administration bodies that supervise the area.

REFERENCES

- Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.
- Dang, S., & Ahmad, P. H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations*, 1(4), 866–2348.
- Erhardt, R. A., Schneider, R., & Blaschke, C. (2006). Status of text-mining techniques applied to biomedical text. *Drug discovery today*, 11(7–8), 315–325.
- Główny Inspektorat Sanitarny. (2019). Rejestr produktów objętych powiadomieniem o pierwszym wprowadzeniu do obrotu na terytorium Rzeczypospolitej Polskiej. Retrived from: <http://rejestrzp.gis.gov.pl> (access: 2019-12-13).
- Główny Inspektorat Sanitarny. (2020). Rejestr produktów objętych powiadomieniem o pierwszym wprowadzeniu do obrotu. Retrived from: <http://powiadomienia.gis.gov.pl>
- Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1), 60–76.
- Harmston, N., Filsell, W., & Stumpf, M. P. (2010). What the papers say: Text mining for genomics and systems biology. *Human genomics*, 5(1), 17.
- Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(5), 326–340.
- Khan, Z., & Qamar, U. (2016). Text Mining Approach to Detect Spam in Emails. In *The International Conference on Innovations in Intelligent Systems and Computing Technologies (ICIISCT2016)* (p. 45).
- Koh, J. X., & Liew, T. M. (2020). How loneliness is talked about in social media during COVID-19 pandemic: text mining of 4,492 Twitter feeds. *Journal of Psychiatric Research*, Nov 07. doi: 10.1016/j.jpsychires.2020.11.015
- Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4), 357–369.
- Mathioudakis, B. (2005). The Emerging Regulatory Environment for Food Supplements in the EU. Food Supplements in Europe: Chllenges for the Future. BLL, EHPM, ERNA, IADSA Conference Report. Retrieved from: <https://www.lebensmittelverband.de/download/food-supplements-eu.pdf>.

- Munková, D., Munk, M., & Vozár, M. (2013). Data pre-processing evaluation for text mining: transaction/sequence model. *Procedia Computer Science*, 18, 1198–1207.
- Najwyższa Izba Kontroli. (2017). Informacja o wynikach kontroli dopuszczanie do obrotu suplementów diety. Retrieved from: <https://www.nik.gov.pl/plik/id,13031,vp,15443.pdf>
- Petkova-Gueorguieva, E. S., Getov, I. N., Ivanov, K. V., Ivanova, S. D., Gueorguiev, S. R., Getova, V. I., Mihaylova, A. A. et al. (2019). Regulatory requirements for food supplements in the European Union and Bulgaria. *Folia medica*, 61(1), 41–48.
- Plisson, J., Lavrac, N., & Mladenec, D. (2004). A rule based approach to word lemmatization. In *Proceedings of IS*, 3, pp. 83–86.
- Restani, P., Di Lorenzo, C., Garcia-Alvarez, A., Badea, M., Ceschi, A., Egan, B., Dima, L. et al. (2016). Adverse effects of plant food supplements self-reported by consumers in the PlantLIBRA survey involving six European countries. *PLoS One*, 11(2), e0150089.
- Sahadevan, S., Hofmann-Apitius, M., Schellander, K., Tesfaye, D., Fluck, J., & Friedrich, C. M. (2012). Text mining in livestock animal science: introducing the potential of text mining to animal sciences. *Journal of animal science*, 90(10), 3666–3676.
- Silano, V., Coppens, P., Larrañaga-Guetaria, A., Minghetti, P., & Roth-Ehrang, R. (2011). Regulations applicable to plant food supplements and related products in the European Union. *Food & function*, 2(12), 710–719.
- Tunali, V., & Bilgin, T. T. (2012). PRETO: A high-performance text mining tool for preprocessing turkish texts. In *Proceedings of the 13th International Conference on Computer Systems and Technologies* (pp. 134–140).
- Ustawa z dnia 11 maja 2001 r. o warunkach zdrowotnych żywności i żywienia. (2001). Dz. U. Nr 63, poz. 634. Retrieved from <https://sip.lex.pl/#/act/16901910/170388/warunki-zdrowotne-zywnosci-i-zywienia?cm=SREST>.
- Ustawa z dnia 25 sierpnia 2006 r. o bezpieczeństwie żywności i żywienia. (2006). Dz. U. z 2020 r. poz. 2021. Retrieved from <https://sip.lex.pl/#/act/17302608/2844276/bezpieczenstwo-zywnosci-i-zywienia?keyword=Bezpiecze%C5%84stwo%20%C5%BCywno%C5%9Bci%20i%20%C5%BCywienia&cm=SFIRST>.
- Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2), 19–28.
- Weiss, S. M., Indurkha, N., Zhang, T., & Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.
- Zhou, X., Han, H., Chankai, I., Prestrud, A., & Brooks, A. (2006). Approaches to text mining for clinical medical records. In *Proceedings of the 2006 ACM symposium on Applied computing* (pp. 235–239). <https://doi.org/10.1145/1141277.1141330>