





# Summarization Algorithms for News: A Study of the Coronavirus Theme and Its Impact on the News Extracting Algorithm

Lyudmila Gadasina<sup>1</sup>  , Vladislav Veklenko<sup>2</sup>, and Pasi Luukka<sup>3</sup> 

<sup>1</sup> St Petersburg University, 7/9 Universitetskaya nab., Saint Petersburg 199034, Russian Federation

[l.gadasina@spbu.ru](mailto:l.gadasina@spbu.ru)

<sup>2</sup> The Moscow Institute of Physics and Technology, 9 Institutskiy per., Moscow Region 141701 Dolgoprudny, Russian Federation

[veklenko.vs@phystech.edu](mailto:veklenko.vs@phystech.edu)

<sup>3</sup> LUT University, Skinnarilankatu 34, 53850 Lappeenranta, Finland

[pasi.luukka@lut.fi](mailto:pasi.luukka@lut.fi)

**Abstract.** Extract summarization algorithms help identify significant information from the news by extracting meaningful sentences from the original text. The information background existing at the time of the news release often significantly affects its content. Such background can distort the text summarization algorithm working results. The study was conducted with the example of the theme “coronavirus” (COVID-19), which at the time of the study was one of the main topics in news feeds. Experiments were carried out on sports news articles, concerned football. This news area was selected because it is not related to medical topics. The TextRank algorithm for sport news extraction was applied in two ways. First, the key information from the source text of news was extracted. Then, a list of the COVID related words was created and the key information from news without considering words from this list was extracted. Our approach showed that mentioning a popular theme such as COVID that is not related to sports can have a negative impact on the text summarization algorithm. We suggest that to obtain accurate results of the algorithm operation, it is necessary to first compile a dictionary of terms related to the coronavirus theme and then exclude them when identifying the main content of news texts.

**Keywords:** Summarization algorithm · News · Coronavirus · Text · Extracting

## 1 Introduction

The problem of texts summarizing is important because of the ever-growing flow of information. Text summarization systems extract brief information from it. Using the resulting summary, the users can select the news related to their needs. Note that, influencing factors for content of news feeds are e.g. specifics of the area and information background relevant at the time of news are published. In usage of automatic text summarizing one needs to realize that temporary information background is influencing the

results and also that results will be affected if it is applied to texts that do not take into account words related to such background.

At the present time coronavirus (COVID-19) is a major theme in news. Effect of coronavirus to media space has been interest of several researchers (see e.g. [1–4]. This has influenced e.g. the area of fakes news [1] and it has had an impact on financial markets [2, 3].

In this paper we are interested in examining: 1) Has coronavirus theme a significant semantic impact on the content of news; 2) Does coronavirus theme affect the results of summarizing news text algorithm?

## 2 Research Method

Research on Automatic Text Classification (ATC) can be traced back to seminal work of Luhn [5]. First studies were mainly centered on text categorization [6, 7]. Theoretical approaches for automatic indexing were developed in the 60's by several authors. For example in [6] proposed probabilistic approach using clue-words derived using frequencies with subject categories. Factor analysis is used in [7] and discriminant analysis for automatic indexing is used in [8]. Due to increased computational power in the late 90's research in the area started to shift from rules based on expert knowledge to learning based methods and machine learning based approaches started to emerge for solving text classification tasks. These include e.g.  $k$ -nearest neighbors [9, 10], decision trees [11], Bayesian classifiers [9, 11] and boosting methods such as AdaBoost [12].

There are two major categories of text summarization – Extractive and Abstractive Summarization [13]. Extractive summarization methods are based on extracting several parts, such as phrases and sentences from a text and stack them together to create a summary. These methods select the most significant words or sentences from the text. The other type, Abstract summation, uses advanced NLP techniques based on recurrent neural networks to create a new resume.

The aim of this article is to identify the influence of the current coronavirus news topic on the text summarization algorithm. Therefore, abstract summarization algorithms are irrelevant in this case, since they are based on memorizing certain sequences of words. This creates a serious obstacle for creating a model that does not take into account coronavirus theme. In our study we use the extractive summarization algorithm based on the graph approach. In [14, 15], extractive summarization is used to extract keywords. In our research, we need to extract the entire sentences to reveal the key idea embedded in the text. In [16], a comparison of PageRank algorithms and Shortest-path algorithms is made. Authors note that the PageRank algorithms do not depend on the language specifics, and the Shortest-path Algorithm generate summaries, which is not so “smooth” to read as a manually written summary.

In this study, a TextRank algorithm based on an extractive approach was constructed. This algorithm was originally proposed by Balcerzak et al. [17, 18]. It is based on the principles of the more famous PageRank algorithm [19], which is mainly used for ranking web pages in online search results. PageRank is a link analysis algorithm that assigns a numerical weighting to each element of a hyperlink document set, to measure its relative importance within the set. The algorithm can be applied to any collection of

objects with mutual quotes and links. In [20], experiments were carried out on extraction to determine the reliability of web content (for identifying web content credibility). Although the method did not show top quality, the authors note that it can help with manual evaluation.

TextRank algorithm works with the same principle, but uses sentences instead of web pages. The similarity level between any two sentences is equivalent to the probability of clicking on a web page. In our study cosine distance was used for similarity calculations. Each word was matched with a numeric vector of a fixed dimension so that the words close in meaning correspond to the vectors close in meaning. Then to get a general sentence vector, the vectors of all the words in the sentence were summed up and divided by the number of words. The coefficients of similarity for each pair of sentences is stored in a square matrix, which is used to build the graph for analyzed text.

Thus, the stages of the TextRank algorithm are the following:

1. Split the text into separate sentences.
2. Find a vector representation for each sentence.
3. Calculate the similarities between sentence vectors and put them to the matrix.
4. Transform the similarity matrix into a graph with sentences as vertices and similarity estimates as edges.
5. Calculate the rank of sentences using a graph.
6. Form the final summary with a certain number of top-level ranking sentences.

### 3 Data Sources and Modelling

In our study, we experimented with a collection of texts from sports news articles, concerned football. We collected the data from a Russian-language sport information source [21]. We wanted to explore how the coronavirus theme affects the news feed in a particular area that is least related to medicine sphere. The corpus of 30,000 news published in the period from 07.09.2019 to 25.04.2020 was collected to create a dictionary containing vector representations of words to form a matrix of similarity of sentences. For this we used the standard libraries of Python programming language (such as nltk, nltk.tokenize, genism.models, rusentokenize and so on) and developed our own functions for data collecting, processing, and algorithm constructing.

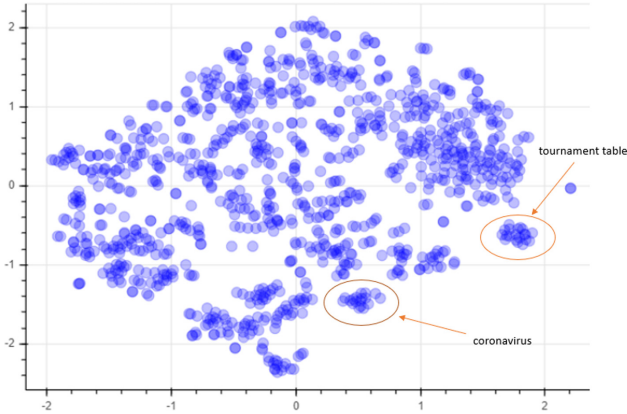
The next important step before building the model was to clean the texts. We created a function for clearing texts, which included the use of both the built-in stop-word library and regular expressions written for processing Russian-language texts, comprises:

- lowercase translation of texts,
- deleting all characters except Russian words,
- deleting stop words,
- deleting short words (less than 3 letters in length).

As a result, we compiled the dictionary with 91825 words mapped each word to a vector of dimension 100, using the Word2Vec library. In particular, the word coronavirus corresponds to the following top 10 words in the collected corpus, the degree of

compliance (probability of being in one sentence) is indicated in brackets<sup>1</sup>: test (0.890), positive (0.858), passed (0.848), suspicion (0.809), passed (plural, 0.797), discovered (0.784), condition (0.781), relative (0.777), informal (0.774), confirmed (0.771).

Figure 1 created using the T-SNE dimension reduction method [22], shows groups of tokens (words) similar in meaning.



**Fig. 1.** Displaying word vectors with T-SNE algorithm. Examples of two groups of words are shown: words related to the “coronavirus” theme and words related to the “tournament table” theme.

Described procedure allows us to create a common vector for each sentence from analyzed texts. For these texts we built the graphs, using the TextRank algorithm discussed above to get a ranking of sentences.

Thus, our algorithm extracts the key information from the text where sentence vectors are based on all words from the created dictionary. To more thoroughly test the hypothesis about the impact of the coronavirus theme on the algorithm, we compiled a list of words that contained the major words associated with this virus: “coronavirus”, “virus”, “pandemic”, “infection”, “lethal”, “disease”, “patient”, “death”, “died”, “discovered”, “symptom”, “suspicion”, “quarantine”<sup>2</sup>. Then we implemented the same algorithm, but do not consider words from the created list in sentence vectors. We did not remove all sentences related to the virus theme from the news, but rank sentences according to the algorithm without taking into account the most popular words associated with coronavirus theme in them.

## 4 Results and Discussion

We tested each algorithm on original sports articles included a coronavirus theme. We considered the two most popular sentences of texts that should determine its content. To

<sup>1</sup> In the original corpus, these are Russian words.

<sup>2</sup> In the original corpus, these are Russian words in different forms (noun cases, verb conjugations etc.).

identify the most significant sentences from the text, we developed a web service that gives as an output results of both algorithms.

First, we illustrate our approach with the four following texts (analyzed texts in original language are presented in Appendix):

Text 1: {1} To complete the English championship, teams will have to play on neutral fields, the BBC reports. {2} The Season was interrupted because of the coronavirus pandemic. {3} It is reported that up to 10 stadiums will be needed to continue the tournament. {4} The Premier League Will also need up to 40,000 coronavirus tests for players and club employees. {5} Earlier, the clubs expressed their readiness to continue the championship when it is possible. {6} According to the source, the resuming football process tournaments in England will be lengthy. {7} Its timing has not yet been determined. {8} It is Planned that the matches will be held without spectators. {9} According to AR, the tournament can start on June 8. {10} According to Worldometer, the total number of people infected with coronavirus in the world has reached 3402886. {11} 239653 deaths were Recorded, and 1084606 people recovered.

Text 2: {1} Roma will suffer serious financial losses at the end of the 2019/20 season, Calciomercato reports. {2} According to the source, the Rome club will suffer losses of \$110 million. {3} It is reported that this is due to the financial crisis caused by the coronavirus pandemic. {4} “Roma did not manage to reduce losses even though the salaries of players and staff were reduced. {5} According to the latest Worldometers data, there are 203,591 cases of coronavirus infection have registered in Italy. {6} 27682 deaths were recorded, and 71252 people recovered.

Text 3: {1} Spain’s health Minister, Salvador Ilya, believes that football matches in the country will not be resumed until the summer. {2} Competitions in Spain were suspended due to the coronavirus pandemic on March 12. {3} – It is reckless to say that football will return before the summer – quotes AP El Salvador Ilya. {4} – We continue to monitor the evolution of the virus. {5} Recommendations will show how life can return in original areas of activity. {6} According to Worldometers, the world revealed 2977188 cases of infection with coronavirus. {7} 206 139 people died, 874587 recovered. {8} Spain is one of the most affected countries. {9} It recorded 226629 cases of infection, 23190 cases were fatal.

Text 4: {1} Borussia Monchengladbach released a special version of the club’s t-shirt. {2} All the money will be used to support medical personnel fighting the coronavirus. {3} Purchase is only possible through an online store. {4} Almost all sporting events were postponed or canceled because of the coronavirus epidemic. {5} According to Worldometer on the night of May 1 to 2, the coronavirus was confirmed in 3,389,933 people worldwide. {6} Dead – 239029, recovered – 1076487. {7} According to Rospotrebnadzor, in Russia on the day of May 1, a total 114431 cases of coronavirus were registered in 85 regions. {8} Over the entire period, 1169 deaths were recorded, and 13220 people recovered.

Results from the created algorithm to four texts above are presented in Table 1. In the Table 1 two most important sentences are reported from both cases, news items for text including and excluding the coronavirus theme.

As we can see applying algorithm to news content without words related to coronavirus allows indicate in texts 1–3 a significant information. For example, the sentence

**Table 1.** Results of algorithm application to four texts.

News	Content with all words		Content without words related to coronavirus	
	The most important sentence in the text	The second in importance	The most important sentence in the text	The second in importance
Text 1	{2}	{6}	{9}	{6}
Text 2	{3}	{5}	{4}	{3}
Text 3	{4}	{8}	{8}	{1}
Text 4	{5}	{7}	{5}	{3}

{9} in news 1 gives information which can be used for decision making, but sentence {2} gives only part of important information. Applying algorithm to original texts in gives noisy information: sentences {5} and {4} respectively. We can note that in news 4 (see Table 1), the sentence related to the coronavirus theme that is not important within the meaning of the news remained in the top. In many cases sentences related to coronavirus remain after deleting the created word list. Sometimes the algorithm pays attention to other words in such sentences that may also be significant for it and it focuses on words that go in combination with words associated with the virus. Examples of such words: “world”, “confirmed”, “crisis” and others. These words can’t be discarded from texts, because they can also appear in another context.

**Result for Full Dataset.** We analyzed all news for the period from 25.04.2020 to 02.05.2020: the total number of news is 921. Of these, 116 news contain the coronavirus theme. In 81 of the original news, the dominant topic is related to coronavirus theme. Of these, in 17 news are no longer the dominant topic is related to corona-virus theme after removing the most popular words associated with this theme.

Thus, 12.6% of news published during the analyzed period contains the virus theme. In almost 70% of cases, this theme is the key topic. The algorithm modification described above allows us to correct this result – the coronavirus theme is the key one in 55.1% of news containing it. Our result shows that in almost half of the cases, the virus theme is only the background context of current events feed of the topic football in Russia. This theme clutter up the news and does not add information.

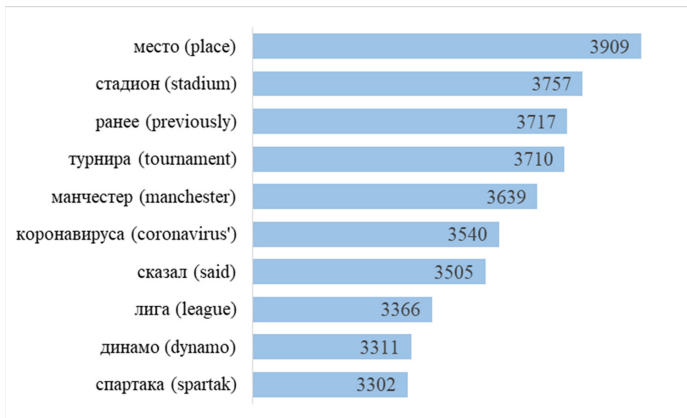
To test coronavirus theme impact on news we used the following approach. We considered the most popular sentence in the text that can be used as a news headline or can help with creating a headline. We compared the actual news headline that was written by experts and the headline predicted by algorithms using the BLEU (Bilingual Evaluation Understudy) [23] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [24] metrics. Table 2 shows the metrics values for original news and news without the most popular words associated with coronavirus theme. ROUGE metrics refers to the overlap of unigram (ROUGE-1), Bigrams (ROUGE-2) and longest common subsequence of texts were applied.

**Table 2.** Comparing metrics values for original news and news without words related to coronavirus.

Used metric	Metric value for content with all words	Metric value for content without words related to coronavirus
BLEU	0.216	0.220
ROUGE-1	0.128	0.132
ROUGE-2	0.047	0.049
ROUGE-L	0.120	0.124

Thus, we note an improvement in the quality of the algorithm after deleting words associated with the coronavirus. Our results show that the appearance of a new topic in the media that is not related to sports can negatively affect the quality of the extraction algorithm. In some cases, the words associated with the virus are information noise and do not define the main idea of the article.

The topic of coronavirus affects the operation of the TextRank algorithm. However, we note that the characteristics of the data used in the study affected the result. In particular, coronavirus theme distorted the corpus on which we trained the algorithms. Words related to this theme are present in the texts for training many times. For example, the word “коронавируса” (one of the forms of the word “coronavirus”) was found in the texts 3540 times and ranked 51th in frequency of use (see Fig. 2). The word “коронавирусом” (another form of the word “coronavirus”) was found 1703 times and was placed in 145 places. Our total dictionary comprise 91 825 words. It means that a lot of attention is paid to the coronavirus pandemic despite the sports theme of the news.

**Fig. 2.** Words from 45 to 55 places by frequency of use.

## 5 Conclusion

In this paper, we explored whether the coronavirus theme is noisy information or has a significant semantic impact on the content of news. We applied a graph-based ranking algorithm TextRank for sport news processing. We identified the key information from news containing the coronavirus theme. Our results show a negative impact on news headline extraction algorithms when there also exists a popular unrelated general topic. To identify significant information during periods when there is a dominant theme more properly, it is necessary to cleanse the source texts from the most popular words associated with this theme. Otherwise, the summarization algorithms may produce distorted results that are not relevant to users' needs.

**Acknowledgments.** The authors are grateful to participants at the Centre for Econometrics and Business Analysis (CEBA, St Petersburg University) seminar series for helpful comments and suggestions.

## Appendix

Analyzed news examples in original (Russian) Language

Text 1: Для завершения чемпионата Англии командам придется играть на нейтральных полях, сообщает BBC. Сезон был прерван из-за пандемии коронавируса. Сообщается, что для продолжения турнира понадобится до 10 стадионов. Также премьер-лиге понадобится до 40 тысяч тестов на коронавирус для игроков и работников клубов. Ранее клубы выразили готовность продолжить чемпионат, когда это будет возможно. По данным источника, процесс возобновления футбольных турниров в Англии будет длительным. Его сроки пока не определены. Планируется, что матчи будут проходить без зрителей. По данным AP, турнир может начаться 8 июня. По данным Worldometer, общее число зараженных коронавирусом в мире достигло 3402 886 человек. Зафиксировано 239653 летальных исхода, 1084606 человек выздоровели.

Text 2: «Барселона» объявила, что клуб передает права на название стадиона «Камп Ноу» собственному фонду Barca Foundation. Фонд займется поиском спонсора, который получит право на имя арены на один сезон – 2020/21. Вырученные средства пойдут на борьбу с коронавирусом. Таким образом, стадион получит спонсорское название впервые в своей истории. Деньги от контракта пойдут на исследовательские проекты, связанные с борьбой с коронавирусом в Испании и во всем мире. Стадион «Камп Ноу» был открыт в 1957 году. Его вместимость составляет 99 354 зрителя. Он принимал матчи чемпионата Европы и мира, финал Лиги чемпионов, футбольный турнир Олимпиады-1992, а также концерты звезд мировой музыки.

Text 3: «Рома» понесет серьезные финансовые потери по итогам сезона-2019/20, сообщает Calciomercato. По информации источника, римский клуб понесет убытки в размере 110 миллионов долларов. Сообщается, что это связано с финансовым кризисом из-за пандемии коронавируса. «Роме» не



удалось сократить потери даже несмотря на сокращение зарплаты игроками и персоналу. По последним данным Worldometers, в Италии зарегистрирован 203 591 случай заражения коронавирусом. Зафиксировано 27 682 летальных исхода, выздоровели 71 252 человека.

Text 4: Министр здравоохранения Испании Сальвадор Илья считает, что футбольные матчи в стране не будут возобновлены до лета. Соревнования в Испании были приостановлены из-за пандемии коронавируса 12 марта.

— Безрассудно говорить, что футбол вернется до лета, — цитирует AP Сальвадора Илью.

— Мы продолжаем следить за эволюцией вируса. Рекомендации покажут, как жизнь сможет вернуться в разных сферах деятельности. По данным Worldometers, в мире выявлено 2 977 188 случаев заражения коронавирусом. 206 139 человек умерли, 874 587 выздоровели. Испания — одна из наиболее пострадавших стран. В ней зафиксировано 226 629 случаев заражения, 23 190 случаев стали летальными.

## References

1. Groza, A.: Detecting fake news for the new coronavirus by reasoning on the COVID-19 ontology. arXiv preprint [arXiv:2004.12330](https://arxiv.org/abs/2004.12330) (2020)
2. Lopatta, K., Alexander, E.-K., Gastone, L., Tammen, T.: To Report or not to report about coronavirus? The Role of Periodic Reporting in Explaining Capital Market Reactions during the Global COVID-19 Pandemic (2020). <https://ssrn.com/abstract=3567778>
3. Mamaysky, H.: Financial markets and news about the coronavirus 27 March 2020. <https://ssrn.com/abstract=3565597>. <https://doi.org/10.2139/ssrn.3565597>
4. Mejova, Y., Kalimeri, K.: Advertisers jump on coronavirus bandwagon: politics, news, and business. arXiv preprint [arXiv:2003.00923](https://arxiv.org/abs/2003.00923) (2020)
5. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. 159–165 (1958). <https://doi.org/10.1147/rd.22.0159>
6. Maron, M.E.: Automatic indexing: an experimental inquiry. J. ACM **8**(3), 404–417 (1961)
7. Borko, H., Bernick, M.: Automatic document classification. J. ACM **10**(2), 151–162 (1963)
8. Williams, J.: Discriminant analysis for content classification, IBM, Technical report no. RADC-TR-66-6 (1966)
9. Larkey, L.: Automatic essay grading using text categorization techniques. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 90–95 (1998)
10. Larkey, L.: A patent search and classification system, DL 99 (1999)
11. Lewis, D.D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, vol. 33, pp. 81–93 (1994)
12. Wilbur, W.J., Kim, W.: The dimensions of indexing. In: AMIA Annual Symposium Proceedings, pp. 714–718 (2003)
13. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization, vol. 293. MIT Press (1999)
14. Litvak, M., Last, M.: Graph-based keyword extraction for single-document summarization. In: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17–24. Association for Computational Linguistics (2008)

15. Naidu, R., Bharti, S.K., Babu, K.S., Mohapatra, R.K.: Text summarization with automatic keyword extraction in Telugu e-Newspapers. In: Satapathy, S.C., Bhateja, V., Das, S. (eds.) *Smart Computing and Informatics. SIST*, vol. 77, pp. 555–564. Springer, Singapore (2018). [https://doi.org/10.1007/978-981-10-5544-7\\_54](https://doi.org/10.1007/978-981-10-5544-7_54)
16. Thakkar, K.S., Dharaskar, R.V., Chandak, M.B.: Graph-based algorithms for text summarization. In: 2010 3rd International Conference on Emerging Trends in Engineering and Technology, pp. 516–519. IEEE (2010). <https://doi.org/10.1109/ICETET.2010.104>
17. Balcerzak, B., Jaworski, W., Wierzbicki, A.: Application of TextRank algorithm for credibility assessment. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, pp. 451–454 (2014). <https://doi.org/10.1109/WI-IAT.2014.70>
18. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Geneva, Switzerland, pp. 404–411 (2004)
19. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* **30**, 1–7 (1998). [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
20. Li, W., Zhao, J.: TextRank algorithm by exploiting Wikipedia for short text keywords extraction. In: 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), pp. 683–686. IEEE (2016). <https://doi.org/10.1109/ICISCE.2016.151>
21. “Sport-ekspress” – sportivnyj portal. (“Sport-Express” — sports portal.). <https://www.sport-express.ru/>. Accessed 25 Apr 2020
22. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
24. Lin, C. Y.: Rouge: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*, pp. 74–81 (2004)