Companies in Multilingual Wikipedia: Articles Quality and Important Sources of Information

Włodzimierz Lewoniewski (✉)[0000−0002−0163−5492], Krzysztof Węcel[0000−0001−5641−3160], and Witold Abramowicz[0000−0001−5464−9698]

Department of Information Systems, Poznan University of Economics and Business
Al. Niepodleglosci 10, Poznan 61-875, Poland
Email: {wlodzimierz.lewoniewski, krzysztof.wecel, witold.abramowicz}@ue.poznan.pl

Abstract. [1] In this paper, we provide a method for the identification and assessment of reliable internet sources about companies. We first identified 516,586 Wikipedia articles related to companies in 310 language versions, and then extracted and analyzed references contained in them using three different models for article quality assessment. As a result, we compiled a ranking of reliable sources. We found that there are several universal sources shared by many languages, but usually each language has its own specific sources. Our ranking of sources can be useful for Wikipedia editors looking for source material for their articles. Companies themselves can leverage this ranking for public relations activities. Moreover, our method can be used to automatically maintain a list of reliable internet sources.

Keywords: Information quality · Credibility of information sources · Wikipedia · Wikidata · DBpedia

1 Introduction

Information presented in Wikipedia articles should be based on reliable sources [9]. The source can be understood as the work (book, paper, etc.), the author, or the publisher. Such sources must have a proper reputation and should present all majority and significant minority views on some piece of information. Following this rule ensures that the readers of the Wikipedia article can be assured that each specific statement provided is supported by a published and reliable source. Therefore, before adding any information to this online encyclopedia, Wikipedia editors (volunteer authors) should ensure that the facts presented in the article can be verified by other people who read Wikipedia [11].

Few developed language versions of Wikipedia contain a non-exhaustive list of sources whose reliability and use in Wikipedia are frequently discussed. Even

[1] This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this contribution is published in Lecture Notes in Business Information Processing, vol 471. Springer, Cham, and is available online at https://doi.org/10.1007/978-3-031-29570-6_3

English Wikipedia, the largest chapter, has such a general list with information on reliability for only 400 websites [10]. Sometimes we can find such lists for specific topics (e.g., video games and movies).

It could take significant human effort to produce a more complete list of trusted Internet sources - there are more than a billion websites available on the Internet [15,27] and many of them can be considered a source of information. Therefore, it can be a very challenging and time-consuming task for Wikipedia volunteers to assess the reliability of each source. Moreover, the reputation of each website can change with time; hence, such lists must be updated regularly. Each source may also have a different reliability score depending on the topic and language version of Wikipedia.

On one hand, we can state that such a list of reliable information sources would be helpful to editors. On the other hand, we have not identified such an approach in the literature. The lack of methods for maintaining a list of reliable sources is a significant research gap. This study presents a method for automating this process by analyzing existing and accepted content with sources from Wikipedia articles on companies in different languages. We use existing and new models to assess the reliability and popularity of websites. We found that, depending on the models, it is possible to find such important sources in selected Wikipedia languages. Additionally, the assessment of the same sources can vary, depending on the language of this encyclopedia.

The paper is structured as follows. Section 2 provides a literature review. In Section 3 we explain our research methodology, i.e. how articles related to companies were identified and what data was collected along with its characteristics. Section 4 extends the research methodology with regard to the extraction of references from previously identified articles. The research findings using three models for source assessment are presented in Section 5. A discussion of the results is carried out in Section 6. Conclusions and future work can be found in Section 7.

2 Related Work

Researching the quality of Wikipedia content is a fairly developed topic in scientific work. As one of the key factors influencing the quality of Wikipedia articles is the presence of references, some studies focused on researching information sources. Some works use the number of references to automatically assess the quality of the information on Wikipedia [34,3]. Such important measures are implemented in different approaches to automatic quality assessment of Wikipedia articles (for example, WikiRank [39]). References often contain external links (URL addresses) where cited information is placed. Such links can be assessed by indicating the degree to which they conform to their intended purpose [36]. Furthermore, these links can be used separately to assess the quality of Wikipedia articles [42,6].

Some of the studies focused on the metadata analysis of sources in Wikipedia references. One of the previous works used ISBN and DOI identifiers to unify references and find the similarity of sources between various Wikipedia language editions [21]. It is becoming more common practice to include scientific sources in references in Wikipedia articles [21,29,22,33]. At the same time, it should be noted that such references often link to open-access works [35] and recently published journal articles [16]. One of the studies devoted to scientific work related to COVID-19 cited in Wikipedia articles found that information comes from about 2% of the scientific works published at that time [5].

News websites are also one of the most popular sources of information in Wikipedia, and there is a method to automatically suggest new references to the selected piece of information [13]. Particularly popular are references about recent content or life events [30]. For example, for information related to the COVID-19 pandemic, Wikipedia editors tend to cite the latest scientific articles and insert more recent information into Wikipedia shortly after the publication of these works [5].

The previous publication [22], relevant to this article, proposed and implemented 10 models for the evaluation of sources in Wikipedia articles. The evaluation results are also implemented in the online tool "BestRef" [2]. Such approaches use features (or measures) that can be extracted from publicly available data (Wikimedia Downloads [38]) so that anyone can use those models for different purposes.

This work is a continuation of the previous study [23]. Compared to the previous article, this study significantly expanded the scope of the analyzed language versions (all language versions available during the period analyzed). We also used more recent data from Wikipedia and Wikidata in order to obtain the results - November 2022. In addition, we conducted an analysis of some aspects of the quality of Wikipedia articles on companies in different languages.

3 Wikipedia Articles Related to Companies

To find such articles, we used data from DBpedia and Wikidata. Data from these open databases are widely used in a number of domains, such as web search, life sciences, art market, digital libraries, and business networks [12,19,14,26].

DBpedia ontology has a hierarchical structure, and if some resource is aligned with other company-related classes, we can use connections between those classes to detect Wikipedia articles related to companies. For example, some organizations can be aligned to 'Bank', 'Publisher', 'BusCompany', or another company-related class of the DBpedia ontology, and after generalization, we can find that all of them belong just to the 'Company' class. Based on DBpedia dumps related to instance types [7] (the specific part of the dumps for each available language), we found that Wikipedia articles can be aligned directly to one of the 634 classes of the DBpedia ontology. After considering transitive DBpedia dumps, we have obtained resources in the 'Company' class. Next, we took similar data extracted by DBpedia from other Wikipedia languages and finally collected an extended list of articles related to companies.

In the next stage, we analyzed Wikidata items that were presented as a collection of different statements structured as Subject—Predicate—Object. Based on Wikidata statements, out of more than 100 million items, we determined more than 100 thousand items that were related to companies. Often they had a statement in the form `Property:P31 Q783794`, meaning 'instance of a company'. We also enriched our knowledge base with statements related to `business` (Q4830453), `enterprise` (Q6881511), `public company` (Q891723), `technology company` (Q18388277), and other similar items. The resulting list of Wikidata items about companies can provide links to related Wikipedia articles in different languages.

Compared to DBpedia ontology classes, Wikidata has roughly 100 times more possible alignments for different items [23]. There are various possibilities to automate the process of identifying company-related items in Wikidata. One of them is to analyze Wikidata items related to companies selected using DBpedia extraction and find the most popular alignments in `instance of` statements. In total, we collected more than 3000 various classes, and the most popular are business, enterprise, public company, company, automobile manufacturer, airline, record label, publisher, bus company, video game developer, organization, commercial organization, and bank.

Before we could identify relevant articles about companies, we introduced several tweaks to our procedure. First, we kept only alignments that appeared at least 200 times to avoid insignificant errors that could be introduced by less experienced users editing Wikidata. Furthermore, we removed the alignment to `organization` (Q43229) which was too general. As a result, we have more Wikidata items with articles on the list of companies; overall, 296,180 Wikidata items were identified with at least one related Wikipedia article in considered language versions. Since each Wikidata item can have one or more links to Wikipedia articles in some language versions, we were able to identify 516,586 articles related to companies in 310 language versions of Wikipedia. A more detailed description of the approach that allowed the search of Wikipedia articles on companies was described in our previous study [23].

Table 1 shows statistics for some of the language versions of Wikipedia (with more than 1000 articles related to companies). Please note that the average and median values were rounded to whole numbers. More extended results are available in the supplementary materials on the Web [8].

It is important to note that the other 30 language versions have only one article about a company, the next 68 languages have 2-10 articles related to companies, and 110 language editions of Wikipedia have over 10 but less than 1000 articles that describe various companies. There are also 51 language versions of Wikipedia that do not have any distinguished articles about a company.

The largest number of articles on companies was found in English Wikipedia – 133,220, which is 2.03% of all articles in that language version. The second largest number of articles about companies has German Wikipedia - 51,700 (2.08% share). Japanese Wikipedia is third in terms of the number of articles (37,292),

Table 1. Statistics on the identification of Wikipedia articles related to companies in different languages. Source: own calculations in November 2022.

Language	**Articles**		**Total Edits**	**Authors**		**Article Len.**		**Page Views**	
	number	**share**		**avg.**	**med.**	**avg.**	**med.**	**avg.**	**med.**
ar - Arabic	12,505	1.05%	1,371,350	8	4	9,882	4,704	5,188	577
arz - Egyptian Arabic	1,245	0.08%	50,018	3	3	2,123	1,406	318	41
az - Azerbaijani	1,357	0.72%	118,823	7	4	6,276	4,058	1,247	158
azb - South Azerbaijani	1,957	0.81%	101,985	4	4	3,047	2,812	38	20
be - Belarusian	1,216	0.54%	116,822	6	5	9,153	5,482	197	80
bg - Bulgarian	1,893	0.66%	263,727	11	8	9,680	5,855	3,033	546
bn - Bangla	2,391	1.85%	209,108	8	5	13,331	8,617	2,156	263
ca - Catalan	6,639	0.93%	778,498	8	5	7,128	4,004	410	84
cs - Czech	6,280	1.23%	826,522	13	9	8,385	4,980	2,689	596
da - Danish	4,183	1.46%	529,293	12	6	4,691	2,929	1,241	237
de - German	57,100	2.08%	11,280,527	32	20	8,597	5,313	5,951	1,011
el - Greek	2,200	1.03%	274,575	10	6	11,722	6,904	3,995	601
en - English	133,220	2.03%	38,459,600	46	24	10,636	6,540	21,652	2,890
eo - Esperanto	1,033	0.32%	105,768	6	5	4,934	2,832	92	34
es - Spanish	19,874	1.10%	3,876,285	19	8	10,393	6,286	12,410	1,363
et - Estonian	1,953	0.85%	225,326	8	6	4,686	2,383	599	143
fa - Persian	8,986	0.96%	847,192	9	3	5,945	3,514	4,485	288
fi - Finnish	9,567	1.77%	1,534,720	14	9	5,518	3,531	1,831	431
fr - French	36,652	1.49%	6,805,196	26	15	9,951	6,033	6,294	882
gl - Galician	1,607	0.84%	167,349	8	5	7,189	4,495	147	51
he - Hebrew	5,033	1.55%	776,861	23	13	10,064	6,692	3,420	702
hi - Hindi	1,266	0.82%	147,389	13	9	17,562	7,421	8,002	1,402
hr - Croatian	1,072	0.50%	136,106	10	6	6,771	4,045	3,093	590
hu - Hungarian	4,063	0.79%	607,256	16	8	10,034	6,090	3,139	479
hy - Armenian	2,078	0.71%	168,097	9	7	10,313	6,416	458	75
id - Indonesian	8,668	1.37%	1,017,284	8	4	8,086	4,424	4,344	412
it - Italian	17,486	0.98%	3,223,485	29	19	9,186	5,612	7,227	1,127
ja - Japanese	37,292	2.76%	7,862,134	26	13	13,342	6,931	12,526	2,659
ko - Korean	7,824	1.28%	1,357,212	14	7	7,458	4,410	5,225	637
lt - Lithuanian	1,461	0.71%	208,089	8	5	5,034	3,528	1,534	286
lv - Latvian	1,125	0.97%	132,502	8	5	6,938	5,053	979	207
ml - Malayalam	1,059	1.33%	107,189	7	5	10,722	6,898	589	126
ms - Malay	4,001	1.11%	308,476	5	3	7,293	4,268	801	120
nl - Dutch	9,939	0.47%	1,750,944	24	13	6,765	4,397	3,290	667
no - Norwegian	6,624	1.10%	1,021,525	18	12	4,383	2,619	1,029	220
pl - Polish	13,662	0.89%	2,181,898	16	9	7,211	4,213	5,063	769
pt - Portuguese	16,148	1.47%	2,591,292	14	7	7,470	4,406	6,260	687
ro - Romanian	5,017	1.15%	742,558	7	4	5,267	2,928	2,636	322
ru - Russian	22,012	1.18%	3,984,793	20	11	16,573	10,494	15,902	1,882
simple - Simple English	2,482	1.12%	271,982	13	7	4,118	2,712	757	139
sk - Slovak	1,251	0.52%	171,155	10	7	7,464	4,702	2,239	520
sr - Serbian	1,852	0.28%	233,473	10	7	12,641	7,582	2,432	508
sv - Swedish	10,597	0.41%	1,742,965	19	11	4,920	3,238	1,962	414
ta - Tamil	1,403	0.94%	146,982	6	4	12,879	6,964	1,297	236
th - Thai	2,114	1.40%	348,045	12	6	13,213	7,724	7,558	975
tr - Turkish	7,060	1.34%	783,320	13	6	6,338	3,692	6,723	648
uk - Ukrainian	9,928	0.83%	1,069,776	10	6	13,497	8,731	3,034	271
ur - Urdu	1,411	0.79%	113,174	3	3	3,869	2,096	168	32
uz - Uzbek	1,342	0.74%	44,999	5	4	13,978	7,579	646	34
vi - Vietnamese	4,061	0.32%	480,356	11	5	11,375	6,128	4,596	500
zh - Chinese	19,673	1.50%	3,329,049	18	9	9,222	5,067	8,289	1,409

but it has the highest share of articles on companies among other Wikipedia languages - 2.76%.

Usually, the total number of edits correlates with the number of articles; therefore, we could expect that the largest number of edits will be in the English, German, and Japanese Wikipedia. However, if we analyze the number of unique authors who edited articles on companies, we can observe slightly different results. We considered only edits from registered authors (with an account on Wikipedia) and excluded bots (which also appear as separate accounts). It must be taken into account that one author may make many insignificant edits (e.g., adding a dot, removing spaces, etc.), while another author may include the entire section(s) in a single edit. In addition, the number of authors may also indicate the degree of objectivity of the content, because each of the authors may have their own opinion on the described organization and the way of presenting information about it. Taking into account the average number of unique authors per article, the top 5 Wikipedia languages include English (46 authors), German (32), Italian (29), French (26), and Japanese (25). This ranking looks similar in the case of median values. The lowest value of the average number of authors per article is in South Azerbaijani, Egyptian Arabic, and Urdu Wikipedia.

The length of the Wikipedia article can also be related to the quality of the content, e.g., completeness of the information about the described company. The length was measured as a volume in bytes of the wiki markup of the Wikipedia article. The largest average length values have the following Wikipedia languages: Hindi (17,562 bytes), Russian (16,573), Uzbek (13,978), Ukrainian (13,497), Japanese (13,342), Bangla (13,331), Thai (13,213), Tamil (12879), Serbian (12,641). When comparing median values, the longest articles belong to Russian (10,494 bytes), Ukrainian (8,730), and Bangla Wikipedia (8,617). Egyptian Arabic Wikipedia has the lowest average and median length of the articles: 2,123 bytes and 1,406 bytes per article, respectively.

The popularity of the articles can not only reflect the demand for information on Wikipedia in a specific language version but can also positively affect the quality of the content (especially on the timeliness of the information on current events). In this study, we considered only page views from real users (not automated or bots) from the last 12 months (November 2021 - October 2022). The largest number of page views per article (average and median) is available in English, Russian, Japanese, Spanish, Chinese, and Hindi Wikipedia.

4 Extraction of References

The following sections present results for the 51 language versions of Wikipedia with at least 1000 articles related to companies. To extract information on references, we prepared our own parser (implemented in Python) and applied it to Wikimedia dumps with articles in HTML format [38].

The presence of references in a Wikipedia article may indicate the degree of verifiability of information. More importantly, this information must come from reliable sources. External links (or URL addressees) in the references were used to indicate the main address of the source website. However, each web source can use a different structure of URL addresses. For example, some websites use subdomains for separate topics of information or news. Also, some organizational units (e.g., departments) of the same company may post their own information on separate subdomains of the main organization. To determine which level of domain indicates the source, we used the Public Suffix List, which is a cross-vendor initiative to provide an accurate list of domain name suffixes [31].

Some sources may have several different domains. For example, Google can be listed in sources as ‘google.com’, ‘google.pl’, ‘google.de’, etc. We, therefore, unified such sources to a single occurrence. Taking into account the fact that various useful services are placed under the ‘google.com’ (e.g., books) and separate blogs on ‘wordpress.com’ subdomains, we additionally provide subdomain distinction for these portals.

Table 2 presents the general extraction statistics. It has three groups of columns: 1) Total references – we count all references encountered in articles, without removing duplicates; 2) References tags share – share of references (in percent), described with respective tag; 3) Unique references – numbers after removing duplicated references, where duplicated were identified based on existing identifiers and similarity between references. The first and third groups comprise three columns: count, showing the absolute numbers; avg, the average number of references per article about a company; med, the median number of references per article about a company. The second group concerns the share of the following features of references: archived, books, and sci score (scientific references). ‘Archived’ means that the reference has a link to one of the archive services with the referenced web page. This often means that the original source may no longer be available or unavailable at the original URL address. In order to identify references related to ‘books,’ we analyzed if there is a link to the Google Books service. ‘Sci’ score counted based on references that contained the DOI identifier [20].

Taking into account the absolute numbers, the language with the highest number of references is English, both when unique and when all references are counted. The next with less than a quarter of references are German and Japanese, but German is using more unique references (second place). The number of references is a consequence of a large number of articles in these languages, therefore we also calculated the number of references per article. The highest number of average references per article, 21, is found in the Uzbek language, although it features only 27.8 thousand articles. The second place is taken by English with an average value of 18. The typical number is in the range of 5-7. However, the largest median is for English - 9. Taking into account the unique references, the situation is similar: English and Uzbek top the list. A “references tags share” promotes other languages. The highest share of archived references belongs to Polish (3.03%), Hindi (2.75%), and Malayalam (2.42%). The books are most often encountered in Indonesian (2.87%), Catalan (2.84%), and Serbian (2.70%). Scientific references are preferred in the following language versions: Arabic (1.75%), Serbian (1.38%), and Malayalam (1.37%).

Table 2. Statistics on references extraction from Wikipedia articles related to companies in different languages. Source: own calculations in November 2022.

Language	Total references			References tags share			Unique references		
	count	avg	med	archived	books	sci	count	avg	med
ar - Arabic	123,743	10	4	1.28%	1.28%	1.75%	105,572	8	3
arz - Egyptian Arabic	5,732	5	3	1.31%	0.42%	0.85%	4,053	3	3
az - Azerbaijani	9,141	7	3	1.61%	1.43%	0.34%	7,542	6	3
azb - South Azerbaijani	7,665	4	3	0.51%	0.09%	0.04%	4,168	2	2
be - Belarusian	9,114	7	4	0.89%	0.21%	0.34%	7,513	6	3
bg - Bulgarian	13,523	7	3	0.92%	0.71%	0.30%	11,379	6	3
bn - Bangla	23,551	10	5	2.06%	1.30%	0.96%	19,750	8	5
ca - Catalan	61,134	9	5	0.56%	2.84%	0.80%	50,321	8	4
cs - Czech	67,032	11	5	0.57%	0.38%	0.42%	50,267	8	4
da - Danish	23,653	6	3	1.84%	0.63%	0.22%	20,386	5	3
de - German	602,498	11	6	0.52%	0.93%	0.25%	488,127	9	4
el - Greek	20,629	9	5	1.47%	1.62%	1.21%	18,184	8	4
en - English	2,344,978	18	9	1.62%	2.01%	0.82%	1,857,221	14	8
eo - Esperanto	4,954	5	2	1.11%	2.04%	0.52%	4,396	4	2
es - Spanish	242,210	12	6	1.34%	1.29%	0.46%	200,098	10	5
et - Estonian	12,994	7	3	0.74%	0.50%	0.06%	10,073	5	2
fa - Persian	46,403	5	2	1.15%	1.22%	0.56%	38,994	4	2
fi - Finnish	88,940	9	5	0.30%	0.23%	0.13%	60,889	6	4
fr - French	460,496	13	6	0.00%	1.41%	0.41%	366,338	10	5
gl - Galician	12,496	8	3	1.43%	1.26%	0.47%	9,989	6	3
he - Hebrew	48,761	10	5	0.51%	0.61%	0.23%	45,011	9	5
hi - Hindi	14,203	11	4	2.75%	0.78%	0.37%	12,044	10	4
hr - Croatian	7,183	7	3	0.72%	0.81%	0.35%	6,047	6	3
hu - Hungarian	41,389	10	5	0.73%	0.60%	0.33%	34,137	8	4
hy - Armenian	22,295	11	5	2.07%	1.43%	1.09%	18,759	9	5
id - Indonesian	102,297	12	5	1.49%	2.87%	0.51%	80,855	9	4
it - Italian	195,831	11	5	1.65%	1.19%	0.32%	155,619	9	4
ja - Japanese	620,815	17	6	0.37%	0.21%	0.27%	408,157	11	4
ko - Korean	54,797	7	3	0.83%	0.62%	0.62%	46,079	6	3
lt - Lithuanian	7,657	5	3	0.89%	0.57%	0.25%	6,935	5	3
lv - Latvian	7,579	7	4	0.75%	0.50%	0.21%	6,394	6	3
ml - Malayalam	9,683	9	5	2.42%	1.40%	1.37%	7,828	7	4
ms - Malay	41,866	10	5	0.90%	0.94%	0.22%	35,138	9	4
nl - Dutch	64,431	6	3	0.50%	0.48%	0.14%	52,548	5	3
no - Norwegian	37,836	6	3	0.65%	0.61%	0.23%	32,232	5	3
pl - Polish	131,539	10	4	3.03%	0.78%	0.17%	98,102	7	3
pt - Portuguese	169,459	10	5	0.80%	0.84%	0.45%	136,242	8	4
ro - Romanian	43,055	9	5	0.48%	0.64%	0.30%	29,221	6	3
ru - Russian	333,347	15	8	1.30%	0.80%	0.40%	259,880	12	6
simple - Simple English	14,560	6	3	1.74%	1.34%	0.89%	12,051	5	3
sk - Slovak	9,785	8	4	0.14%	0.72%	0.22%	7,725	6	3
sr - Serbian	20,129	11	5	1.84%	2.70%	1.38%	16,711	9	4
sv - Swedish	66,446	6	3	1.59%	0.39%	0.21%	55,124	5	3
ta - Tamil	10,742	8	4	1.70%	1.31%	0.66%	9,324	7	4
th - Thai	19,161	9	5	0.61%	1.27%	0.61%	15,940	8	4
tr - Turkish	47,141	7	3	1.24%	0.94%	0.41%	39,977	6	3
uk - Ukrainian	103,271	10	5	1.20%	0.62%	0.75%	85,099	9	4
ur - Urdu	4,769	3	1	0.67%	0.84%	0.27%	4,119	3	1
uz - Uzbek	27,828	21	8	1.43%	1.23%	0.49%	24,923	19	8
vi - Vietnamese	53,908	13	6	1.51%	1.29%	0.57%	44,584	11	5
zh - Chinese	220,141	11	5	1.43%	0.85%	0.18%	174,152	9	4

5 The Information Sources in Wikipedia about Companies

This section presents the results of the evaluation of the most important sources of information about companies described in various Wikipedia languages and assessed using different models.

It is important to note that archive services (e.g., archive.org, archive.today) were excluded from the analysis, due to the frequent occurrence of such links alongside the original sources in the same reference. If the original source is no longer available, such archive services are very important because Wikipedia readers can verify information, but unavailable original web sources are not in the scope of this research. References to Wikipedia itself and Wikidata were also excluded. Many references contained links that are automatically inserted based on such identifiers as DOI (often links to doi.org) or ISBN (often links to books.google.com).

This work used the following modified and improved models from our previous articles on source assessment [22,23]:

1. **F**-model – how frequently (F) considered source appears in references.
2. **LRP**-model – how popular (P) Wikipedia articles are, in which the considered source appears.
3. **LRA**-model – how many authors (A) edited the articles, in which the considered source appears.

5.1 F-model

One of the most basic and commonly used approaches to assessing the importance of a web source is to count how frequently it was used in Wikipedia articles. This principle was used in relevant studies [28,21,32,16]. Therefore, the **F**-model assesses how many times a specific web domain occurred within the external links of the references. For example, if the same source is cited 50 times in 44 Wikipedia articles (each contains at least one reference with such web source), we count the (cumulative) frequency as 50. Equation 1 shows the calculation for the F-model.

$$F(s) = \sum_{i=1}^{n} X_s(i), \quad \begin{array}{l} \text{where:} \\ s \text{ is the source (website or web domain),} \\ n \text{ is the number of the considered Wikipedia articles,} \\ X_s(i) \text{ is the number of references that the source } s \text{ uses} \\ \text{ (e.g. domain in URL) in the article } i. \end{array} \tag{1}$$

The top web sources according to the F-model include websites such as nytimes.com (American daily newspaper: 76,072 references), worldcat.org (international union library catalog: 70,784), reuters.com (international news agency:

45,520), bloomberg.com (American multinational mass media corporation: 32,675), forbes.com (American business magazine: 29,552), bbc.co.uk (British public service broadcaster: 28,729), techcrunch.com (American technology news website: 25,962), wsj.com (American business-focused daily newspaper: 25,703).

Next, we created separate Web sources rankings for each language version. To provide cross-lingual analysis and due to the limited space in the next graph, we selected only websites that appeared at least seven times among the top 100 websites for each of the 51 selected language versions of Wikipedia (see Table 2). Websites that appear in the top 100 of each of the 51 languages are the following: nytimes.com, reuters.com, bloomberg.com, forbes.com. Figure 1 shows the positions in the ranking of the best web sources of information on companies in each of the 51 languages on Wikipedia according to the F-model.

5.2 LRP-Model

LRP-model uses page views (or visits) of Wikipedia articles within a certain period divided by the total number of references in each Wikipedia article considered. Some studies found a correlation between information quality and page views in Wikipedia articles [18,1]. Such a measure as page views can be considered a public interest in a specific topic [41,37]. The more people read a specific Wikipedia article, the more likely its content was checked by part of them (including the presence of reliable sources in references). So, the more readers see particular facts in Wikipedia, the bigger the probability that one of such readers will make an appropriate edit if such facts are incorrect (or if the source of information is inappropriate).

The visibility of a single reference is also important. If more references are present in the article, then a specific source for the particular reader (visitor) is less visible. At the same time, the more references Wikipedia articles have, the more visible a particular source is. Equation 2 shows the calculation using the *RLA*-model.

$$LRP(s) = \sum_{i=1}^{n} \frac{L(i)}{X(i)} \cdot X_s(i) \cdot P(i), \tag{2}$$

where:
s is the source (website or web domain),
n is the number of considered articles,
$X_s(i)$ is the number of references using the source s in the Wikipedia article i,
$X(i)$ is the total number of references in i,
$L(i)$ length of the Wikipedia article i,
$P(i)$ number of page views of the article i.

This model uses cumulative page views P from human users (excluding bots) in November 2021 - October 2022. Figure 2 shows the positions in the ranking

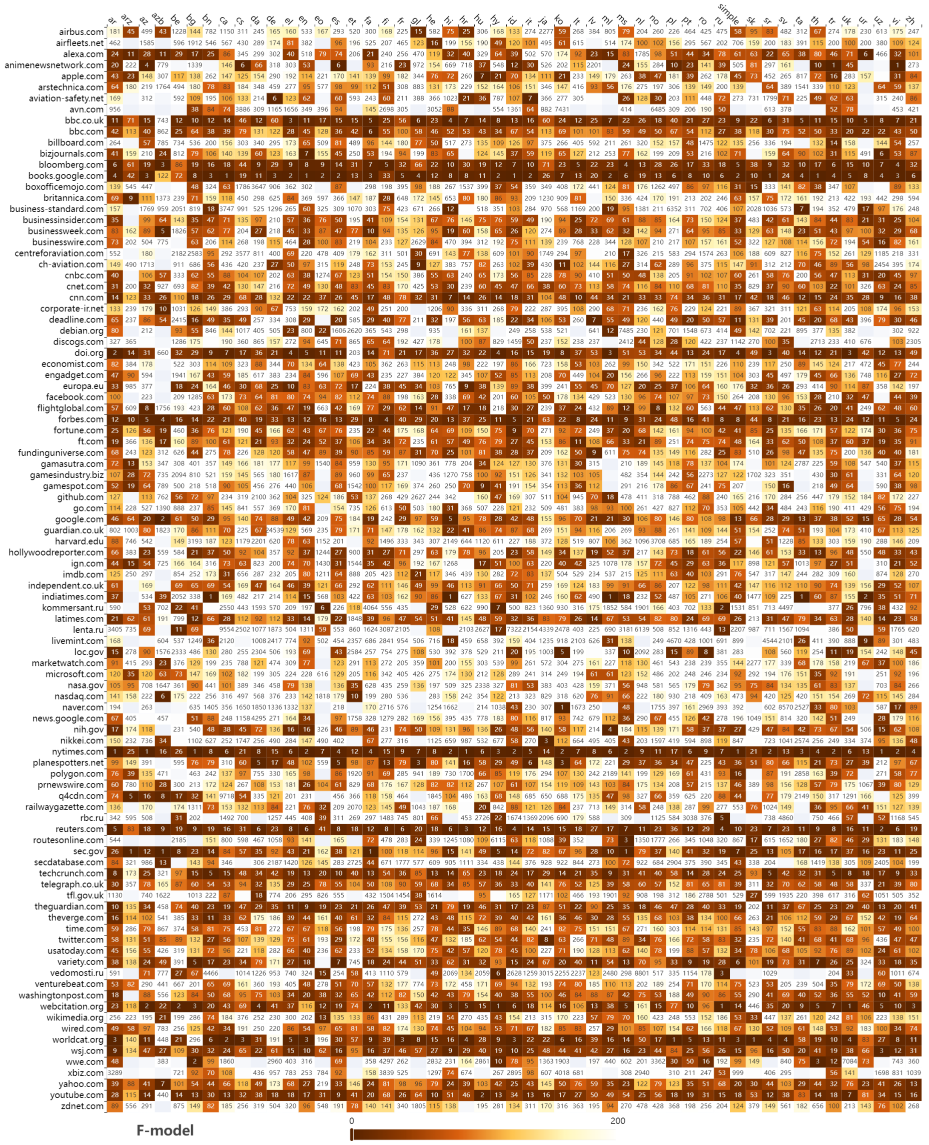

Fig. 1. Positions in rankings of the best web sources of information about companies in each of 51 languages on Wikipedia according to F-model. Interactive and extended versions of this chart can be found on [8]

of the best Web sources of information about companies in each of 51 language versions according to the LRP-model.

Comparing the results between LRP-model and F-model, we can find some important changes in the web sources rankings. These are some examples of such changes in the multilingual ranking (in all Wikipedia languages):

- statista.com (platform specialized in market and consumer data): the 287th place according to F-model and the 127th place according to LRP-model
- imdb.com (online database of information related to films, television series, and video games): the 86th place according to F-model and the 233rd place according to LRP-model
- mashable.com (digital media platform, news website, and entertainment company): the 155th place according to F-model and the 31st place according to LRP-model.
- discogs.com (website and database about audio recordings): the 101st place according to F-model and the 968th place according to LRP-model.
- fb.com (online social media and social networking service): the 1126th place according to F-model and the 76th place according to LRP-model.

5.3 LRA-Model

The quality of Wikipedia articles also depends on the number of authors who contributed to the content and their experience. Wikipedia articles of high quality are often edited jointly by a large number of different authors. This correlation was observed by many authors [24,40,4,17,25]. To assess the popularity of an article among editing users, there is the possibility of analyzing the revision history of the article to find how many authors were involved in content creation and editing. So, the *AR*-model characterizes how popular the article is among Wikipedia volunteer editors. Equation 3 presents this model in mathematical form.

$$LRA(s) = \sum_{i=1}^{n} \frac{L(i)}{X(i)} \cdot X_s(i) \cdot A(i), \tag{3}$$

where:
s is the source (website or web domain),
n is the number of considered articles,
$X_s(i)$ is the number of references using the source s in the Wikipedia article i,
$X(i)$ is the total number of references in i,
$L(i)$ length of the Wikipedia article i,
$A(i)$ number of authors of the article i.

Unlike our previous work, the *LRA*-model in this study uses the number of authors A who are registered on Wikipedia as users, excluding bots.

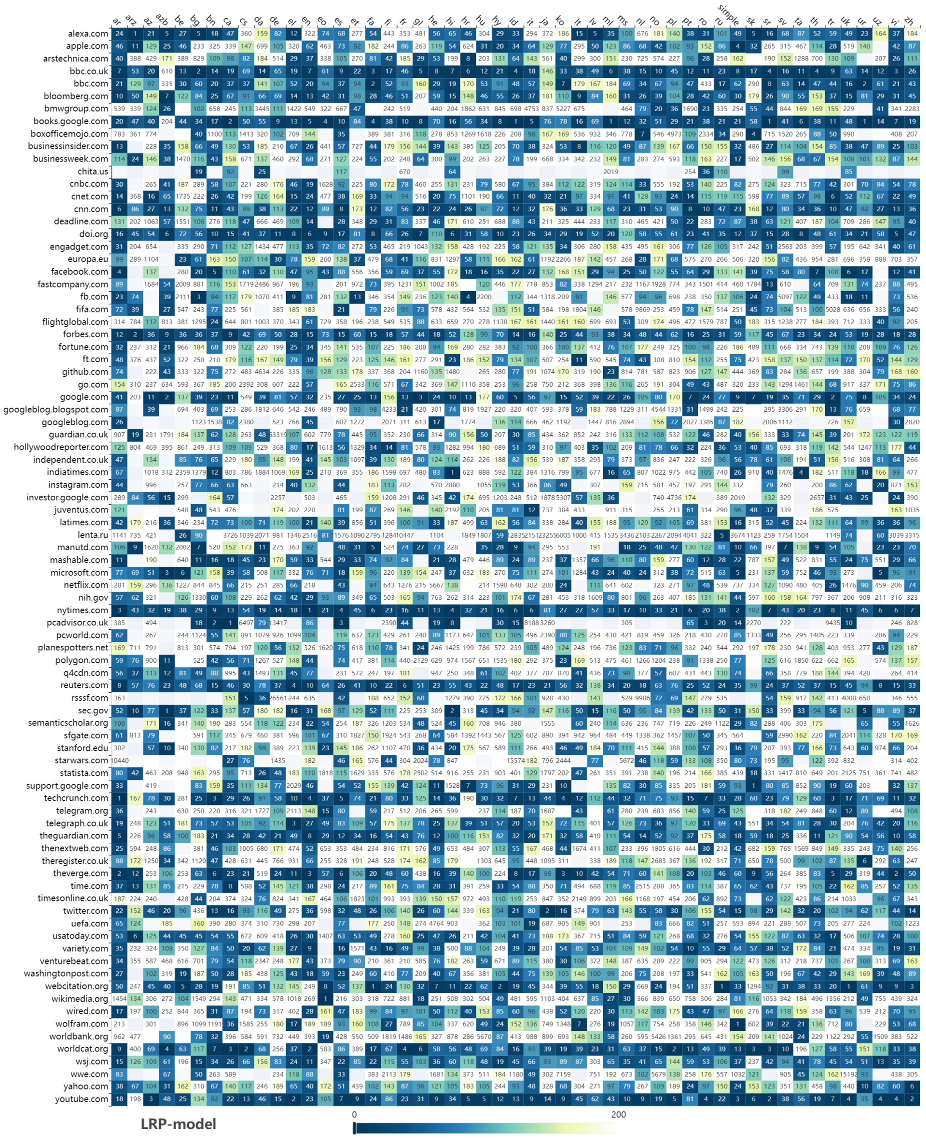

Fig. 2. Positions in rankings of the best web sources of information about companies in each of 51 language versions according to the LRP-model. Interactive and extended versions of this chart can be found on [8]

Figure 3 shows the positions in the ranking of the best Web sources of information about companies in each of the 51 language versions of Wikipedia according to the LRA-model.

Comparing results between LRA-model and F-model, we can also find some important changes in the web sources rankings. Below is a summary of the differences:

- uefa.com (website of one of six continental bodies of governance in association football): the 736th place according to F-model and the 162nd place according to LRA-model
- nasdaq.com (American stock exchange): the 162nd place according to F-model and the 270th place according to LRA-model.
- loc.gov (research library of the United States Congress): the 52nd place according to F-model and the 105th place according to LRA-model.
- harvard.edu (research university in Cambridge, Massachusetts): the 103rd place according to F-model and the 70th place according to LRA-model.

6 Discussion of the Results

Some important websites for separate language versions are not presented in the heat maps 1, 2, and 2, due to poor support among at least six additional language versions. It means that some of the Web sources can be widely used in only one or two language versions of Wikipedia. For example, newspapers.com is the fourth most important source of information on companies in the English Wikipedia according to the F-model, but only one additional language version of Wikipedia (among the 51 languages considered) has this source in the top 100. Depending on the model, we can observe some differences between lists of sources that were selected for the heat maps (only websites that appeared at least seven times among the top 100 websites for each of 51 languages). However, there are many important sources for separate Wikipedia languages that are not presented in such heat maps.

The following is the list of sources that did not meet the threshold required to be placed in presented heat maps but are placed in the top 10 important sources according to one of the three models in some Wikipedia languages (the highest position among models in the local ranking is given in brackets):

- **ar (Arabic)**: grid.ac (7)
- **arz (Egyptian Arabic)**: grid.ac (2), charitynavigator.org (3), csfd.cz (3), wikisource.org (3), justice.cz (4), purl.org (4), youm7.com (4), staralliance.com (5), miningreece.com (6), ralphlauren.com (7), creativecommons.org (7), nashvillezoo.org (8), cbc.ca (8)
- **az (Azerbaijani)**: e-qanun.az (1), lent.az (2), president.az (3), deyerler.org (5), virtualaz.org (6), apa.az (6), shanghairanking.com (7), azertag.az (7), mediaforum.az (8), mta.info (10), qafqazinfo.az (10)
- **azb (South Azerbaijani)**: domaintools.com (8), the-afc.com (10)
- **be (Belarusian)**: zviazda.by (1), marketing.by (1), svaboda.org (2), tut.by (3), belta.by (3), minsk.by (4), gortransport.kharkov.ua (5), sigla.ru (6), nbrb.by (7), yandex.ru (7), rt.com (9), nn.by (9), metropoliten.by (10)
- **bg (Bulgarian)**: dnevnik.bg (1), mersenne.org (1), capital.bg (4), technologyreview.com (6), brra.bg (7), btv.bg (8), bas.bg (10)
- **bn (Bangla)**: thedailystar.net (1), bdnews24.com (4), indianrailways.gov.in (7), prothomalo.com (8), irfca.org (10)

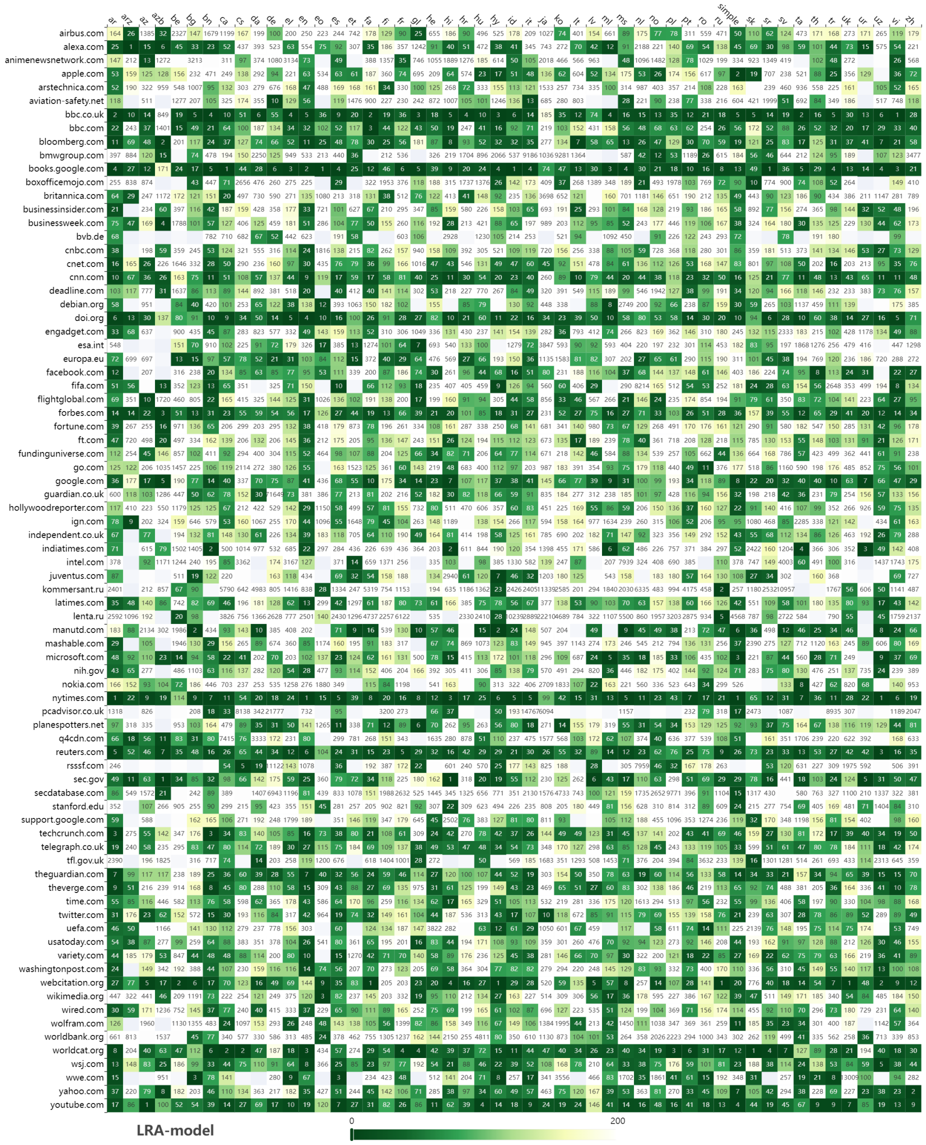

Fig. 3. Positions in rankings of the best web sources of information about companies in each of 51 Wikipedia language versions according to LRA-model. Interactive and extended versions of this chart can be found on [8]

- **ca (Catalan)**: elpais.com (3), gencat.cat (4), lavanguardia.com (5), enciclopedia.cat (6), ara.cat (7), vilaweb.cat (8), bcn.cat (10)
- **cs (Czech)**: justice.cz (1), idnes.cz (1), lupa.cz (3), denik.cz (4), ihned.cz (5), zdopravy.cz (6), ceskatelevize.cz (6), lidovky.cz (7), e15.cz (8), aktualne.cz (9), novinky.cz (10)
- **da (Danish)**: dr.dk (1), business.dk (2), dsb.dk (2), brondby.com (2), politiken.dk (3), finans.dk (4), starwarsplaces.com (5), berlingske.dk (6), borsen.dk (7), tv2.dk (8), computerworld.dk (9)
- **de (German)**: spiegel.de (1), zdb-katalog.de (1), mementoweb.org (2), heise.de (3), handelsblatt.com (4), tagesspiegel.de (4), faz.net (6), welt.de (7), sueddeutsche.de (7), bundesbank.de (10), zeit.de (10)
- **el (Greek)**: et.gr (1), similarweb.com (1), kathimerini.gr (2), utm.edu (6), e-tetradio.gr (6), typologies.gr (7)
- **en (English)**: newspapers.com (4)
- **eo (Esperanto)**: yandex.ru (1), wikiwix.com (2), liberafolio.org (3), vc.ru (4), staralliance.com (5), creativecommons.org (5), aidh.org (6), rezo.net (7), metromadrid.es (8), zelpage.cz (10)
- **es (Spanish)**: elpais.com (1), issn.org (2), elmundo.es (6), lanacion.com.ar (8)
- **et (Estonian)**: postimees.ee (1), delfi.ee (2), err.ee (3), aripaev.ee (4), muinas.ee (4), riigiteataja.ee (4), nasdaqbaltic.com (5), digar.ee (5), efis.ee (8), swedbank.ee (8), staralliance.com (10)
- **fa (Persian)**: tehran.ir (4), radiofarda.com (5), hamshahrionline.ir (6), isna.ir (7), mehrnews.com (9)
- **fi (Finnish)**: yle.fi (1), hs.fi (1), kauppalehti.fi (3), hel.fi (3), is.fi (4), talouselama.fi (6), iltalehti.fi (7), stat.fi (8), habbo.fi (8), finder.fi (9), espoo.fi (9), tekniikkatalous.fi (10), taloussanomat.fi (10)
- **fr (French)**: lesechos.fr (1), lemonde.fr (1), lefigaro.fr (2), wikiwix.com (5), bnf.fr (7), googleusercontent.com (7), ozap.com (7), societe.com (8), liberation.fr (8)
- **gl (Galician)**: elpais.com (1), lavozdegalicia.es (1), skyrocket.de (1), jstor.org (4), rinoceronte.gal (9), formulatv.com (9), numista.com (10), laopinioncoruna.es (10)
- **he (Hebrew)**: globes.co.il (1), nli.org.il (2), themarker.com (2), haaretz.co.il (4), ynet.co.il (4), calcalist.co.il (5), tase.co.il (7), walla.co.il (7), mako.co.il (8), makorrishon.co.il (10)
- **hi (Hindi)**: pib.nic.in (3), rbi.org.in (5), annualreports.com (6), ndtv.com (9)
- **hr (Croatian)**: hrt.hr (1), mojarijeka.hr (2), vecernji.hr (3), dnevnik.hr (3), jutarnji.hr (6), casopis-gradjevinar.hr (6), enciklopedija.hr (7), htmlgoodies.com (9), rtl.hr (9), poslovni.hr (10), tportal.hr (10)
- **hu (Hungarian)**: index.hu (1), origo.hu (2), hvg.hu (2), iho.hu (3), kaze.fr (4), telex.hu (4), villamosok.hu (6), 24.hu (6), sg.hu (7), mavcsoport.hu (7), crt-tv.com (8), blog.hu (9), media1.hu (10)
- **hy (Armenian)**: matenadaran.am (1), 1tv.am (3), cba.am (4), amazon.fr (5), asj-oa.am (6), csufresno.edu (7), unesco.org (9), stretfordend.co.uk (10)
- **id (Indonesian)**: detik.com (2), kompas.com (3), tempo.co (4), liputan6.com (7), tribunnews.com (8), thejakartapost.com (9), transjakarta.co.id (10)
- **it (Italian)**: repubblica.it (1), corriere.it (2), ilsole24ore.com (2), rai.it (4), ansa.it (6), lastampa.it (7), beniculturali.it (7), cm-lisboa.pt (8), primaonline.it (10)
- **ja (Japanese)**: catr.jp (1), ndl.go.jp (2), sponichi.co.jp (3), edinet-fsa.go.jp (4), impress.co.jp (4), asahi.com (5), itmedia.co.jp (5), nikkansports.com (7), jreast.co.jp (8), prtimes.jp (9), eirparts.net (10)
- **ko (Korean)**: kbs.co.kr (2), gg.go.kr (2), chosun.com (3), donga.com (4), yonhapnews.co.kr (5), ytn.co.kr (6), mt.co.kr (6), joins.com (7), hani.co.kr (8), hankyung.com (9), mois.go.kr (9), mk.co.kr (10)
- **lt (Lithuanian)**: vz.lt (1), delfi.lt (2), 15min.lt (3), vle.lt (4), litrail.lt (4), lrt.lt (5), lrytas.lt (6), lrs.lt (9)
- **lv (Latvian)**: db.lv (1), delfi.lv (2), tvnet.lv (3), lursoft.lv (4), lsm.lv (4), diena.lv (6), airbaltic.com (6), lattelecom.lv (8), inbox.lv (9), ldz.lv (9), ltv.lv (10), porsche.com (10)
- **ml (Malayalam)**: mathrubhumi.com (1), manoramaonline.com (2), madhyamam.com (4), thehindu.com (5), kerala.gov.in (5), nhrc.nic.in (7), jal.co.jp (8), eci.nic.in (8), ncert.nic.in (9), kseb.in (10)
- **ms (Malay)**: thestar.com.my (1), utusan.com.my (2), malaysiaairlines.com (2), mstar.com.my (4), airasia.com (5), bernama.com (6), rtm.gov.my (6), themalaysianinsider.com (7), astroawani.com (8), bharian.com.my (9)
- **nl (Dutch)**: nrc.nl (1), volkskrant.nl (1), nos.nl (2), nu.nl (2), fd.nl (3), kb.nl (5), ad.nl (5), telegraaf.nl (6), standaard.be (8), trouw.nl (8), tijd.be (10)
- **no (Norwegian)**: nb.no (1), nrk.no (2), brreg.no (3), regjeringen.no (3), aftenposten.no (4), e24.no (5), dn.no (6), stretfordend.co.uk (6), snl.no (8), proff.no (9), dagbladet.no (9), hafenhamburg.de (10)
- **pl (Polish)**: wirtualnemedia.pl (1), wyborcza.pl (2), plk-sa.pl (3), rynek-kolejowy.pl (4), wp.pl (4), sejm.gov.pl (6), satkurier.pl (7), mma.pl (7), transinfo.pl (8), transport-publiczny.pl (8), onet.pl (9), pwn.pl (10), tvp.pl (10)
- **pt (Portuguese)**: uol.com.br (1), globo.com (2), cm-lisboa.pt (3), abril.com.br (3), estadao.com.br (4), mziq.com (5), terra.com.br (5), sapo.pt (8), tecmundo.com.br (9)

- **ro (Romanian)**: wall-street.ro (1), zf.ro (1), money.ro (3), adevarul.ro (4), arma.org.ro (4), afi.com (5), arabcrunch.com (5), capital.ro (6), mediafax.ro (6), evz.ro (7), paginademedia.ro (7), hotnews.ro (8), acasatv.ro (9), metrorex.ro (10)
- **ru (Russian)**: ria.ru (6), vk.com (6), tass.ru (7), forbes.ru (9), vkontakte.ru (10), gazeta.ru (10)
- **simple (Simple English)**: mathvault.ca (6), baskinrobbins.com (7)
- **sk (Slovak)**: sme.sk (2), socialblade.com (2), dennikn.sk (3), finstat.sk (4), imhd.sk (4), orsr.sk (5), pravda.sk (6), aktuality.sk (6), etrend.sk (7), hnonline.sk (8), techbyte.sk (8), zoznam.sk (9), visibility.sk (10)
- **sr (Serbian)**: b92.net (2), rts.rs (3), novosti.rs (5), exyuaviation.com (7), nb.rs (8), nbs.rs (9), blic.rs (9)
- **sv (Swedish)**: allabolag.se (1), svt.se (1), kb.se (2), dn.se (2), svd.se (3), sverigesradio.se (4), resume.se (6), historiskt.nu (8), expressen.se (8), trafikverket.se (8), di.se (9), aftonbladet.se (9), runeberg.org (10)
- **ta (Tamil)**: indianrailways.gov.in (1), tn.gov.in (1), thehindu.com (2), rbi.org.in (5), theekkathir.org (7), thehindubusinessline.com (8), dinamani.com (9)
- **th (Thai)**: soc.go.th (1), mcot.net (1), listedcompany.com (2), gotomanager.com (2), set.or.th (3), mgronline.com (4), thairath.co.th (6), prachachat.net (7), settrade.com (8), dbd.go.th (10), positioningmag.com (10)
- **tr (Turkish)**: hurriyet.com.tr (1), milliyet.com.tr (2), haberturk.com (3), ntv.com.tr (6)
- **uk (Ukrainian)**: rada.gov.ua (2), rbc.ua (3), uprom.info (3), epravda.com.ua (4), pravda.com.ua (5), detector.media (7), ukrinform.ua (8), anisearch.de (9), president.gov.ua (10)
- **ur (Urdu)**: ourairports.com (1), booleanstrings.com (2), zerohedge.com (3), dawn.com (4), nlpd.gov.pk (4), tribune.com.pk (7), radio.gov.pk (7), pakrail.com (8), piac.com.pk (10)
- **uz (Uzbek)**: kun.uz (2), ziyouz.com (3), uztelecom.uz (7)
- **vi (Vietnamese)**: vtv.vn (1), tuoitre.vn (2), vnexpress.net (4), hoinhabaovietnam.vn (8)
- **zh (Chinese)**: nii.ac.jp (1), sina.com.cn (1), qq.com (3), udn.com (4), ltn.com.tw (4), on.cc (5), xinhuanet.com (6), hk01.com (7), hkexnews.hk (8), sohu.com (9), people.com.cn (10), appledaily.com (10)

The results also showed that many references contained links that are automatically inserted based on such identifiers as DOI and ISBN numbers, which often link to doi.org and books.google.com, respectively. Such web services provide works written by different authors and shared by various organizations (including publishing houses). In that case, a more detailed analysis can be performed in future work.

In this paper, we present a cross-lingual comparison only for 51 selected language versions of Wikipedia. Extended and interactive results for all 310 language versions of Wikipedia can be found in the supplementary material [8].

7 Conclusion and Future Work

This study focused on the analysis of the quality of Wikipedia articles on companies and their sources of information in different languages. Using the semantic representation of information in DBpedia and user-generated knowledge in Wikidata, this study provides the method for identifying Wikipedia articles that describe separate companies. After determining the titles of Wikipedia articles and extracting references from their content, we traced the URLs of these references and determined the main site addresses. As a result, we identified the websites of the sources considered. Each identified web source of information was assessed using an improved version of the three models from our previous research.

The approach presented in this work can help not only Wikipedia volunteer editors in selecting websites that can provide valuable information on companies, but also help other Internet users better understand how to find valuable sources of information for a specific topic on the Web using open data from Wikipedia.

The models we used in the research have some limitations. Some of them use page views and the number of authors of Wikipedia articles. These measures can be imprecise or not always available. For example, some of the page views can be accidental: the Internet user, shortly after visiting the Wikipedia article, can realize that information is not relevant and search for another page in the encyclopedia or move on to another website. Another example is "short page views", where the reader spends a relatively short time studying only a few sentences and their sources from the beginning of the article. In this case, the reader will not see all the content and references to sources in the Wikipedia article. Unfortunately, Wikipedia does not provide data on the duration of each user's visit to the website. Regarding data on Wikipedia authors, it allows for analysis of the reputation of the particular user, who provides some changes to the Wikipedia article. It is even possible to analyze each contribution of any author. Therefore, some of the models presented can be improved in future work by providing more complex measurements of some features.

We plan to extend this research in the future by providing additional features on the identification of companies on Wikipedia. So far, we have been concerned with various aspects of the quality of Wikipedia articles on companies, including objectiveness, completeness, timeliness, and verifiability. In addition, we will group organizations by sectors (industries) to find the differences in the reliability of information sources. Future work will also focus on the extension of reliability models and the use of different methods in topic classification. One of the directions is to develop ways of weighing the importance of a reference based on its position within a Wikipedia article.

Acknowledgements

This research is supported by the project "OpenFact – artificial intelligence tools for verification of the veracity of information sources and fake news detection" (INFOSTRATEG-I/0035/2021-00), granted within the INFOSTRATEG I program of the National Center for Research and Development, under the topic: Verifying information sources and detecting fake news.

References

1. Apollonio, D.E., Broyde, K., Azzam, A., De Guia, M., Heilman, J., Brock, T.: Pharmacy students can improve access to quality medicines information by editing wikipedia articles. BMC medical education **18**(1), 1–8 (2018). https://doi.org/10.1186/s12909-018-1375-z
2. BestRef: Popularity and Reliability Assessment of Wikipedia Sources. https://bestref.net (2022)
3. Blumenstock, J.E.: Size matters: word count as a measure of quality on Wikipedia. In: Proceedings of the 17th international conference on World Wide Web. pp. 1095–1096. ACM (2008). https://doi.org/10.1145/1367497.1367673

4. Callahan, E.S., Herring, S.C.: Cultural bias in wikipedia content on famous persons. Journal of the American society for information science and technology **62**(10), 1899–1915 (2011). https://doi.org/10.1002/asi.21577
5. Colavizza, G.: COVID-19 research in Wikipedia. Quantitative Science Studies **1**(4), 1349–1380 (12 2020). https://doi.org/10.1162/qss_a_00080
6. Conti, R., Marzini, E., Spognardi, A., Matteucci, I., Mori, P., Petrocchi, M.: Maturity assessment of Wikipedia medical articles. In: Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on. pp. 281–286. IEEE (2014). https://doi.org/10.1109/CBMS.2014.69
7. Databus: DBpedia Ontology instance types. https://databus.dbpedia.org/dbpedia/mappings/instance-types/ (2022)
8. data.lewoniewski.info: Supplementary materials for this research. https://data.lewoniewski.info/company/ (2022)
9. English Wikipedia: Wikipedia:Reliable sources. https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources (2022)
10. English Wikipedia: Wikipedia:Reliable sources/Perennial sources. https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources (2022)
11. English Wikipedia: Wikipedia:Verifiability. https://en.wikipedia.org/wiki/Wikipedia:Verifiability (2022)
12. Färber, M., Ell, B., Menne, C., Rettinger, A.: A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web Journal **1**(1), 1–5 (2015)
13. Fetahu, B., Markert, K., Nejdl, W., Anand, A.: Finding news citations for wikipedia. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 337–346 (2016)
14. Filipiak, D., Filipowska, A.: Improving the quality of art market data using linked open data and machine learning. In: Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19. pp. 418–428. Springer (2017). https://doi.org/10.1007/978-3-319-52464-1_39
15. Internet Live Stats: Total number of Websites. https://www.internetlivestats.com/total-number-of-websites/ (2022)
16. Jemielniak, D., Masukume, G., Wilamowski, M.: The most influential medical journals according to Wikipedia: quantitative analysis. Journal of medical Internet research **21**(1), e11429 (2019). https://doi.org/10.2196/11429
17. Kane, G.C.: A multimethod study of information quality in wiki collaboration. ACM Transactions on Management Information Systems (TMIS) **2**(1), 4 (2011). https://doi.org/10.1145/1929916.1929920
18. Lerner, J., Lomi, A.: Knowledge categorization affects popularity and quality of Wikipedia articles. PloS one **13**(1), e0190674 (2018). https://doi.org/10.1371/journal.pone.0190674
19. Lewańska, E.: Towards automatic business networks identification. In: Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19. pp. 389–398. Springer (2017). https://doi.org/10.1007/978-3-319-52464-1_36
20. Lewoniewski, W.: Identification of important web sources of information on wikipedia across various topics and languages. Procedia Computer Science **207**, 3290–3299 (2022)
21. Lewoniewski, W., Węcel, K., Abramowicz, W.: Analysis of references across Wikipedia languages. In: International Conference on Information and Software Technologies. pp. 561–573. Springer (2017). https://doi.org/10.1007/978-3-319-67642-5_47

22. Lewoniewski, W., Węcel, K., Abramowicz, W.: Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information **11**(5), 263 (2020). https://doi.org/10.3390/info11050263
23. Lewoniewski, W., Węcel, K., Abramowicz, W.: Identifying reliable sources of information about companies in multilingual wikipedia. In: 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS). pp. 705–714. IEEE (2022). https://doi.org/10.15439/2022F259
24. Lih, A.: Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. 5th International Symposium on Online Journalism p. 31 (2004)
25. Liu, J., Ram, S.: Using big data and network analysis to understand Wikipedia article quality. Data & Knowledge Engineering (2018). https://doi.org/10.1016/j.datak.2018.02.004
26. Metilli, D., Bartalesi, V., Meghini, C.: A wikidata-based tool for building and visualising narratives. International Journal on Digital Libraries **20**, 417–432 (2019). https://doi.org/10.1007/s00799-019-00266-3
27. Netcraft: August 2021 Web Server Survey. https://news.netcraft.com/archives/2021/08/25/august-2021-web-server-survey.html (2021)
28. Nielsen, F.Å.: Scientific citations in Wikipedia. arXiv preprint arXiv:0705.2106 (2007). https://doi.org/10.48550/arXiv.0705.2106
29. Nielsen, F.Å., Mietchen, D., Willighagen, E.: Scholia, scientometrics and Wikidata. In: European Semantic Web Conference. pp. 237–259. Springer (2017). https://doi.org/10.1007/978-3-319-70407-4_36
30. Piccardi, T., Redi, M., Colavizza, G., West, R.: Quantifying engagement with citations on Wikipedia. In: Proceedings of The Web Conference 2020. pp. 2365–2376 (2020). https://doi.org/10.1145/3366423.3380300
31. Public Suffix List: List. https://publicsuffix.org/learn/ (2022)
32. Redi, M.: Characterizing Wikipedia Citation Usage. Analyzing Reading Sessions. https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Citation_Usage/Analyzing_Reading_Sessions (2019), [Online; accessed 01-Sep-2021]
33. Singh, H., West, R., Colavizza, G.: Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. Quantitative Science Studies **2**(1), 1–19 (2021). https://doi.org/10.1162/qss_a_00105
34. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Assessing information quality of a community-based encyclopedia. Proc. ICIQ pp. 442–454 (2005)
35. Teplitskiy, M., Lu, G., Duede, E.: Amplifying the impact of open access: Wikipedia and the diffusion of science. Journal of the Association for Information Science and Technology **68**(9), 2116–2127 (2017). https://doi.org/10.1002/asi.23687
36. Tzekou, P., Stamou, S., Kirtsis, N., Zotos, N.: Quality Assessment of Wikipedia External Links. In: WEBIST. pp. 248–254 (2011)
37. Weiner, S.S., Horbacewicz, J., Rasberry, L., Bensinger-Brody, Y.: Improving the quality of consumer health information on wikipedia: case series. Journal of medical Internet research **21**(3), e12450 (2019). https://doi.org/10.2196/12450
38. Wikimedia Downloads: Main page. https://dumps.wikimedia.org (2021)
39. WikiRank: Quality and Popularity Assessment of Wikipedia Articles. https://wikirank.net/ (2022)
40. Wilkinson, D.M., Huberman, B.a.: Cooperation and quality in wikipedia. Proceedings of the 2007 international symposium on Wikis WikiSym 07 pp. 157–164 (2007). https://doi.org/10.1145/1296951.1296968

41. Wulczyn, E., West, R., Zia, L., Leskovec, J.: Growing wikipedia across languages via recommendation. In: Proceedings of the 25th International Conference on World Wide Web. pp. 975–985 (2016). https://doi.org/10.1145/2872427.2883077
42. Yaari, E., Baruchson-Arbib, S., Bar-Ilan, J.: Information quality assessment of community generated content: A user study of Wikipedia. Journal of Information Science **37**(5), 487–498 (2011). https://doi.org/10.1177/0165551511416065