Peer-reviewed version available at Communications in Computer and Information Science 2018, 920, 546-558; doi:10.1007/978-3-319-99972-2_45

Article

Determining Quality of Articles in Polish Wikipedia Based on Linguistic Features

Włodzimierz Lewoniewski[1,†,‡]*, Krzysztof Węcel[1,‡] and Witold Abramowicz[1,‡]

[1] Poznań University of Economics and Business;
{wlodzimierz.lewoniewski,krzysztof.wecel,witold.abramowicz}@ue.poznan.pl
* Correspondence: wlodzimierz.lewoniewski@ue.poznan.pl; Tel.: +48 (61) 639-27-93
† Current address: al. Niepodległości 10, 61-875 Poznań, Poland
‡ These authors contributed equally to this work.

Abstract: Wikipedia is the most popular and the largest user-generated source of knowledge on the Web. Quality of the information in this encyclopedia is often questioned. Therefore, Wikipedians have developed an award system for high quality articles, which follows the specific style guidelines. Nevertheless, more than 1.2 million articles in Polish Wikipedia are unassessed. This paper considers over 100 linguistic features to determine the quality of Wikipedia articles in Polish language. We evaluate our models on 500 000 articles of Polish Wikipedia. Additionally, we discuss the importance of linguistic features for quality prediction.

Keywords: Wikipedia; Polish; Information Quality; Linguistic Features; Linguistics; Data Mining; NLP

1. Introduction

Information has become a commodity. The quantity and quality of information largely determine the quality of decisions in various business branches. On the one hand, managers care about having access to as wide range of information as possible. On the other hand, the quality of information determined by different features (such as relevance, accuracy, unambiguity and others) is also important.

Nowadays everyone can contribute to common human knowledge on the Web. One of the best examples of such online repositories is Wikipedia, which have more than 46 million articles in 288 active language editions[1]. Polish version of this encyclopedia is one of the biggest language editions. As of March 2017, the Polish-language edition of Wikipedia had more than 1.2 million articles and more than 250 million page views per month.

In Wikipedia there are systems of grades for article quality and particular language version can use own assessment standard [1,12]. Each language version have special awards for articles with the best quality. In English version such articles are called "Featured articles" (FA). In Polish Wikipedia articles with the highest quality have name "Artykuły na medal", what is essentially equivalent to FA grade in English. Such articles should be well written, in particular fulfil certain criteria: encyclopedic style, uniformity of tense, neutrality and other[2]. The language of encyclopedia article must be very precise. It is necessary to avoid clutter of thoughts or excessive coloration. Separate rules draw

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[2] https://pl.wikipedia.org/wiki/Pomoc:Styl_%E2%80%93_poradnik_dla_autor%C3%B3w

attention to lead section of article: it should be clear, not very long or detailed, written in accessible language, allowing the reader to learn the content without reading the whole article[3]. Articles that meet a core set of editorial standards but are not featured articles, qualify as "Good articles" (GA); in Polish language – "Dobre artykuły".

Usually in each language version of Wikipedia there are only about 0,4-0,6% of high-quality articles (marked as FA or GA). Other articles can get lower quality grades but still most of the articles are unevaluated. For Polish Wikipedia the share of articles without quality grade is about 99%. This number could be lowered by involving more experienced users and experts from different disciplines. Unfortunately, such experts are not always available.

Writing style of articles depends on the language characteristics. FA and GA articles cover more concepts, objects and facts than other ones [2]. So we can expect, that these articles used more nouns and verbs and less adjectives [3]. There are different studies that estimate the quality of articles in English Wikipedia based on various lexical features. However, there are no such studies related to Polish Wikipedia. Even more, there are no studies examining the use of specific Polish linguistic characteristics to evaluate the quality of texts in general.

In this paper we use over 100 linguistic features to determine quality of articles in Polish Wikipedia. Apart from the entire text of each of the article, we also analyzed features of a special part – the leading section. Our model achieved high precision and we test our model on selected unevaluated articles to predict the quality.

2. Related work

Existing studies describes various ways to predict the quality of the Wikipedia articles. Some of them determine the quality based on article's content, another uses the edit history, the article's talk page and other sources. In general, we can divide related studies into the two groups: content-based and user-based approaches. Existing research works proposed different feature sets for measuring quality of Wikipedia articles.

Works related to user's (editor's) behavior, explore how the users experience and coordinate their activities in relation to article quality. These approaches use various characteristics related to a user reputation and changes that they made [4,5]. Usually high quality articles have a large number of editors and edits [6]. Interaction among editors and articles can be visualized as a network, and using graph theory structural features associated to articles quality can be determined [7]. There is also artificial intelligence service involved to discover damaging edits, which can be used to immediately score the quality [8]. However, such user-based methods often require complex calculations and they do not analyze article itself, which would indicate what needs to be changed to improve its quality.

Another group of the scientific works analyzed the article content. One of the first studies showed that longer articles in Wikipedia often had higher quality grades [9]. Later works identified other features related to various constituents of the article: the best articles have more images, sections, use bigger number of references than articles with lower quality [1,10,11]. Online service WikiRank[4] use some of the content quantitative features to assess relative quality of Wikipedia articles in different languages [12]. Special quality flaw templates can also help in articles assessment in Wikipedia [13]. A few works try to combine features from edition history and articles content [14,15].

Recently, in scientific works more and more attention is paid to analyzing not only the quantitative features of the text of articles but also qualitative. One of the studies used character trigram feature to analyze article writing style, which can be a predictor for its quality [16]. Another study used 8 basic lexical metrics derived from the statistic on word usages in Wikipedia articles as the factors that can reflect its quality [3]. Linguistic features can also be used to examine how density of factual

[3] https://pl.wikipedia.org/wiki/Pomoc:Jak_napisa%C4%87_dobr%C4%85_definicj%C4%99
[4] http://www.wikirank.net

information impacts articles quality: bigger relative number of facts in a document indicates its more informative [2].

Concluding, existing studies propose different feature sets for assessing quality of articles in Wikipedia. However, there is no single universal feature set for doing it [15], especially if we consider different language versions [1,10]. Nowadays, Wikipedia contains articles in over 290 languages[5]. Most of the research related to automatic quality assessment based on linguistic parameters is associated with the biggest language version of Wikipedia – English [2,3,16], which has more than 5.5 million articles. The most similar work for this study was done recently for articles in Russian Wikipedia [17]. However, we did not find such studies related to Polish language.

Our work is related to content-based approach and using more than 100 linguistic features to predict the quality of Polish articles in Wikipedia. In addition to the entire articles text, we decided to separately explore the linguistic features of one of its most important parts – leading section.

3. Preparing the dataset

In Polish Wikipedia there are over 2300 GA and over 700 FA articles, which cover topics such as: biographies, humanities, arts, social sciences, earth sciences, mathematics, physical sciences, chemical sciences, biological sciences, medical sciences, technical sciences, military, and others. Depending on the language version, Wikipedia articles can have other (lower) quality grades [1,10,12]. In the case of the Polish Wikipedia there are additional 5 quality grades, which show the degree of development of the article: four, correct, sufficient, start, stub. However, despite the use of different names in separate projects of the Polish Wikipedia, some grades are similar in their criteria. As a result, we group them to 3 lower quality grades: correct (four), start (sufficient), stub. The number of articles with certain quality grade can be seen in Table 1.

Table 1. Articles count with quality grade in Polish Wikipedia. Source: own calculations in May 2017.

Quality Grade	Articles count
Featured Article	723
Good Article	2 303
Correct	785
Start	1 246
Stub	1 635
Without grade	1 212 559

According to previous works [1,2,10,11,15,17], we divide all grades into two classes:

- Complete articles (with FA and GA).
- To-improve (with Correct, Start, Stub).

To train our model we also decided to use two sampling methods:

- Imbalanced - with all assessed articles in each class.
- Balanced - with equal number of randomly extracted articles in each class (stratified sampling).

In case of Balanced dataset, the number of articles of each grade was based on maximum articles count in the smallest class (Complete), and taking into account the equal representation of each grade within each class. Imbalanced dataset used all the available evaluated articles. Details are presented in Table 2.

[5] https://meta.wikimedia.org/wiki/List_of_Wikipedias

Table 2. Articles count in each class in two data sets

Quality Grade	Imbalanced	Balanced	Class
Featured Article	723	720	Complete
Good Article	2 303	720	Complete
Correct	785	480	Uncomplete
Start	1 246	480	Uncomplete
Stub	1 635	480	Uncomplete
Total	**6 692**	**2 880**	

4. Features extraction

Wikipedia have special web service that provides convenient access to wiki features, data, and meta-data over HTTP. We use it to extract the texts of the articles (with leading section). We implemented own application to extract linguistic features based on morphological vocabulary - PoliMorf [18]. Most of the parameters have the notation according to morphosyntactic marker system used in the corpus of Institute of Computer Science, Polish Academy of Sciences [19].

Aside from the standard markup for the Polish language, we included additional features such as: number of unique words[6] (and separately verbs, nouns, adjectives), noun to verb ratio, long words (with over 3 syllables) and others. Some of these additional features are also used in assessing the readability of Polish texts [20]. List of analyzed features can be found in Table 3.

We also used frequency lists extracted from large texts corpora, including National Corpus of Polish Language, Rzeczpospolita (newspaper), Polish Wikipedia and others. These corpora have about 1.8 billion tokens in common and SuperMatrix utilities were used to generate the frequency list [21]. We created features that count separately words from top 50, 100, 200, 300, 400, 500 and 1000 words of frequency list of base forms used in each article (*f50*, *f100*, *f200*, *f300*, *f400*, *f500*, *f1000*). We did similarly for popular words taken from frequency list of Polish Wiktionary[7] (*w50*, *w100*, *w200*, *w300*, *w400*, *w500*, *w1000*). Before the feature calculation, we converted each word in articles into the base form using the PoliMorf vocabulary mentioned before.

[6] Unique words was counted on base forms each of each in the texts.
[7] https://pl.wiktionary.org/wiki/Indeks:Polski_-_Najpopularniejsze_s%C5%82owa_1-2000

Table 3. List of analyzed linguistic features. Source: own work.

Name	Description	Name	Description
acc	accusative case	*nie*	words with "nie" (negation)
add	additional term	*nom*	nominative case
adj	adjective	*noun*	subst + depr
adja	adjectival adjective	*noun/verb*	noun to verb ration
adjc	predicative adjective	*npraep*	non-post-prepositional
adjp	post-prepositional adjective	*num*	numeral
adv	adverb	*nwok*	non-vocalic
aff	affirmative	*org*	organization
agl	agglutinative	*osoba*	person
aglt	agglutinate form "być"	*own*	proper name
akc	accented	*p1*	personal plurale tantum gender
AWLS	avg. words lengths (in syllabs)	*p2*	second plurale tantum gender
bedzie	future form "być"	*p3*	third plurale tantum gender
burk	bound word	*pact*	active adj. participle
com	comparative degree	*pant*	anterior adv. participle
comp	subordinating conjunction	*pcon*	contemporary adv. participle
congr	agreeing accommodability	*perf*	perfective
conj	coordinating conjunction	*pl*	plural
dat	dative case	*pos*	positive degree
depr	depreciative noun	*posp*	common words
etn	etnonim	*ppas*	passive adj. participle
f	feminine	*ppron12*	non-3rd person pronoun
f100	top 100 of frequency list	*ppron3*	3rd person pronoun
f1000	top 1000 of frequency list	*praep*	post-prepositional
f200	top 200 of frequency list	*praet*	l-participle
f300	top 300 of frequency list	*prd*	product
f400	top 400 of frequency list	*pred*	predicative
f50	top 50 of frequency list	*prep*	preposition
f500	top 500 of frequency list	*pri*	first person
fin	non-past form	*qub*	particle-adverb
gen	genitive case	*rec*	governing accommodability
geo	geographical name	*sec*	second person
ger	gerund	*sg*	singular
imperf	imperfective	*sie*	word "się"
imps	impersonal	*sname*	surname
impt	imperative	*subst*	noun
inf	infinitive	*sup*	superlative degree
inst	instrumental case	*ter*	third person
interj	interjection	*uadj*	unique adjective count
ladj	long adjectives	*unoun*	unique noun count
lnoun	long nouns	*uverb*	unique verb count
loc	locative case	*verb*	verb
lverb	long verbs	*voc*	vocative case
lword	long words	*w100*	top 100 of wiki-frequency list
m1	personal masculine gender	*w1000*	top 1000 of wiki-frequency list
m2	animate non-personal masculine gender	*w200*	top 200 of wiki-frequency list
m3	inanimate masculine gender	*w300*	top 300 of wiki-frequency list
n1	first neuter gender	*w400*	top 400 of wiki-frequency list
n2	second neuter gender	*w50*	top 50 of wiki-frequency list
nagl	non-agglutinative	*w500*	top 500 of wiki-frequency list
nakc	non-accented	*winien*	word "winien"
name	name	*wok*	vocalic
neg	negative	*wyd*	event

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2018 doi:10.20944/preprints201801.0017.v1

Peer-reviewed version available at Communications in Computer and Information Science 2018, 920, 546-558; doi:10.1007/978-3-319-99972-2_45

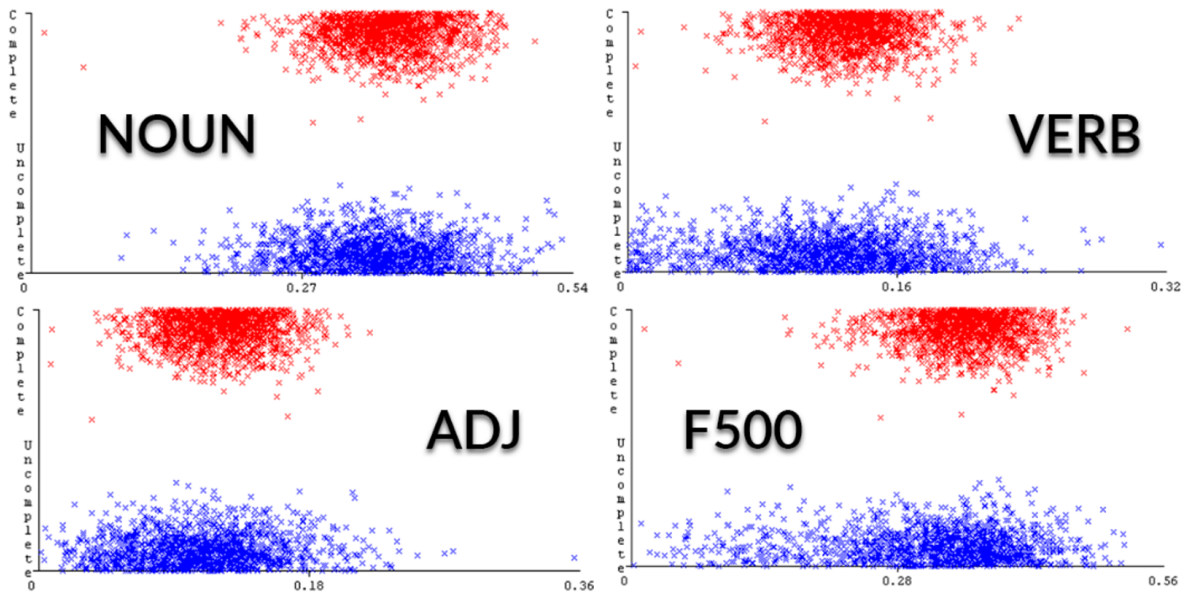

Figure 1. Feature values (normalized by word count) by quality classes in Balanced dataset: nouns (NOUN), verbs (VERB), adjectives (ADJ), words from frequency list (F500 - first 500 words). Source: own work using WEKA software.

Previous works showed that articles with high quality usually have more text (length, number of words) [9,12,16]. Therefore we can expect that the value of a majority of analyzed linguistic features will correlated with the quality of the article - the larger the article the bigger will be the number of nouns, verbs and other parts of speech. Therefore, we decided to normalize all features by words count. In this case, the density of these parts of speech will play a greater role. We expect predicting quality with such normalized features be a challenging task.

We have analysed the distribution of values of various features in relation to quality classes. It is basically not possible to unambiguously assign article to the class, based only on one feature. Distribution of some features is presented in Figure 1.

In addition, we took into account features from only leading sections of a Wikipedia articles. This sections placed before the table of contents and the first heading, serves as an introduction to the article and a summary of its most important contents. Wikipedia community have separate rules for writing the leading section[8].

5. Building the quality models

We decide to use Random Forest algorithm to analyze the described linguistic features in order to automatically determine quality classes of Wikipedia articles. This data mining algorithm is robust, i.e. tolerates correlated variables and noise, so it shows the highest precision on similar tasks [1,10,11,17]. We applied its implementation in WEKA software[9] using default settings - 100 trees, cross-validation with 10 folds. In order to build a model, we take into account 106 linguistic features as independent variables and the quality class as the dependent variable. Cross-validation allows to test quality models on independent data sets. Classification accuracy of models with various datasets is presented in Table 4.

The evaluation shows that higher precision can be achieved when analyzing the whole text compared to the leading section - difference is of about 10 percent. We also observe that one can

[8] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section
[9] http://www.cs.waikato.ac.nz/ml/weka/

Table 4. Classification accuracy (weighted avg.) in percent for Balanced and Imbalanced datasets.

Index	Balanced		Imbalanced	
	Lead. sect.	Text	Lead. sect.	Text
TP Rate	81.3	91.8	81.5	93.1
FP Rate	18.8	8.2	18.5	6.8
Precision	81.5	92.0	81.6	93.3
Recall	81.3	91.8	81.5	93.1
F-Measure	81.2	91.8	81.5	93.1
MCC	62.8	83.8	62.8	86.4
ROC Area	89.4	97.1	90.3	97.5
PRC Area	89.3	96.9	90.1	97.3

achieve a slightly larger value of classification accuracy using imbalanced dataset, which is associated with a large number of training samples.

One of the advantages of using the Random Forest is the ability to identify the most significant features for prediction of quality. In Figure 2 we show importance of all considered features in scale 0-100 (a higher value indicates a higher importance) in different datasets and different analyzed parts (text and leading section). In Table 5 we show the top 10 most significant features for each dataset and analyzed part.

Table 5. The top 10 most significant features in predicting articles quality in Polish Wikipedia. Source: own calculation using Random Forest algorithm

Balanced		Imbalanced	
Lead. sect.	Text	Lead. sect.	Text
posp	unoun	posp	ter
loc	ter	pos	unoun
voc	uverb	voc	uverb
m2	imps	ter	posp
ter	conj	loc	imps
pos	inf	conj	adja
m3	congr	m3	pos
subst	posp	geo	voc
geo	adja	m2	praep
m1	praep	nwok	conj

The results showed that there are differences between the quality models based on features taken from the text and from leading section, but are similar if comparing Balanced and Imbalanced dataset. If we consider leading section of articles, the highest importance in prediction quality are relative quantity of common words, locatives, vocatives, third person words, and ordinal. In case of text features, the most important are relative number of unique nouns, unique verbs, 3rd person verbs, common words, and impersonal verbs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2018 doi:10.20944/preprints201801.0017.v1

Peer-reviewed version available at Communications in Computer and Information Science 2018, 920, 546-558; doi:10.1007/978-3-319-99972-2_45

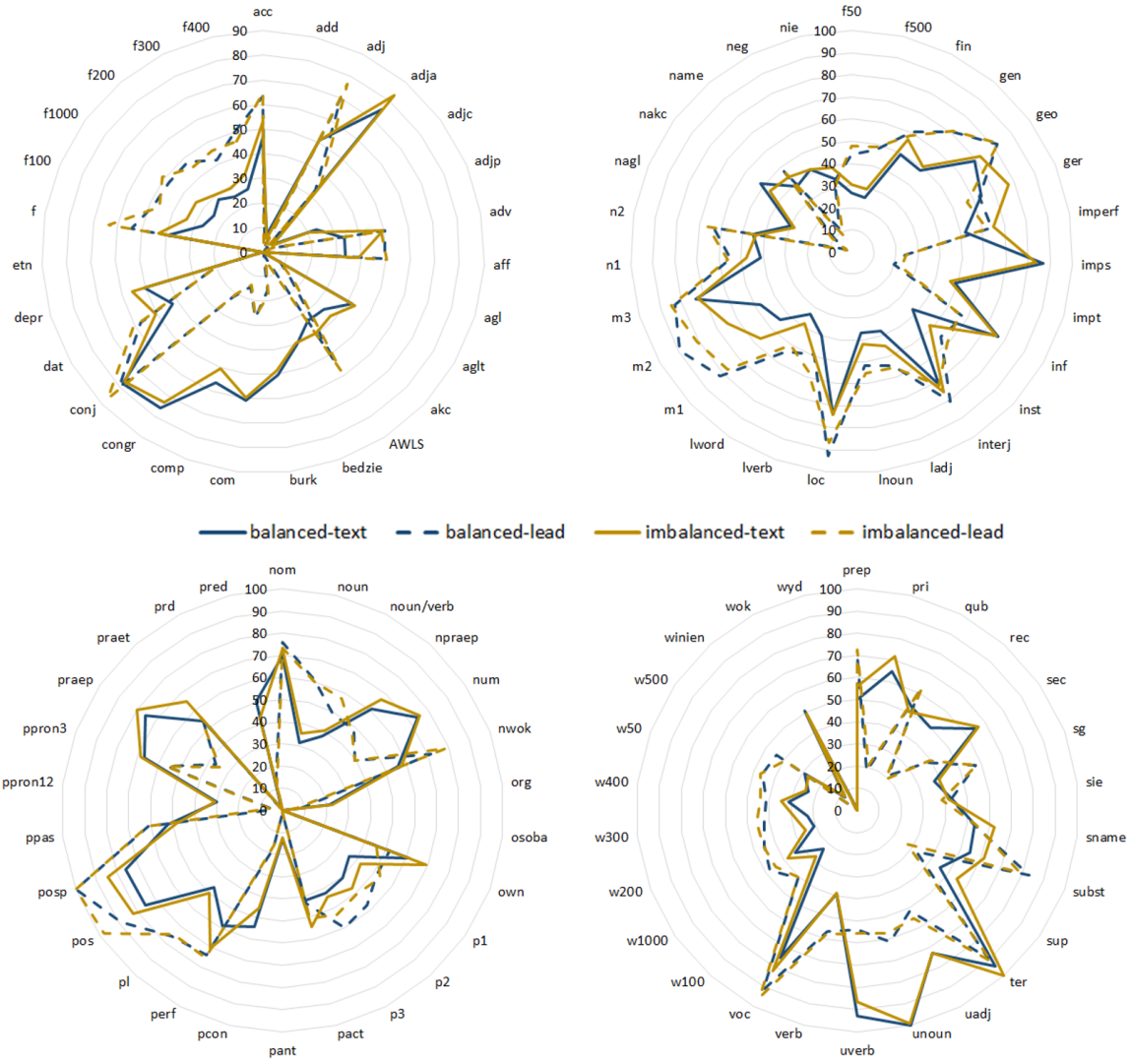

Figure 2. Feature importance (normalized by number of nodes using that features) in classification model using Random Forest algorithm in different datasets.

6. Articles evaluation

Quality models based on text features from Balanced and Imbalanced dataset have similar precision. Therefore we used both to assess 500 000 randomly chosen unevaluated articles from Polish Wikipedia. About 4-5% of these articles were determined by both models as Complete. Results are shown in Table 6.

Table 6. Classification results for 500 000 unevaluated articles in Polish Wikipedia using quality models from Balanced and Imbalanced dataset.

Balanced		**Imbalanced**	
Lead. sect.	Text	Lead. sect.	Text
18 306	481 694	25 209	474 791

Balanced model, which show slightly lower precision than Imbalanced, marks near 3,7% of analyzing articles as Complete. Imbalanced model classified about 5% of these as good quality articles.

7. Conclusions and Future Work

Use of linguistic features is valuable for automatic determination of quality of Wikipedia articles in Polish language. Better results in terms of precision can be achieved when the whole text of article is taken into the account. Then our model shows over 93% classification precision using such features as relative number of unique nouns and verbs (unique, 3rd person, impersonal). However, if we take into account only leading section of an article, relative quantity of common words, locatives, vocatives and third person words are the most significant for determination of quality.

Using the obtained quality models we asses 500 000 randomly chosen unevaluated articles from Polish Wikipedia. According to result, about 4-5% of assessed articles can be considered by Wikipedia community as high quality articles.

We plan to expand the number of considered independent variables in the quality model and complement them with semantic features and readability formulas specific for Polish language. We also plan to increase the number of frequency and topics dictionaries.

Previous work and current study shows that it is possible to build similar models for other languages, so we also plan to compare the quality of information in different language versions of the same article in Wikipedia. However, it requires availability of specialized dictionaries (resources) and language-specific techniques. This is especially challenging task for less developed language editions of Wikipedia such as Belarusian, Ukrainian, Czech and other.

The linguistic features considered in this study can be useful for building more complex quality models of articles in different languages of Wikipedia. This will not only increase the accuracy of determining the quality of articles but also progress from a binary quality classification to a more detailed grading scheme.

References

1. Węcel, K.; Lewoniewski, W. Modelling the Quality of Attributes in Wikipedia Infoboxes. In *Business Information Systems Workshops*; Abramowicz, W., Ed.; Springer International Publishing, 2015; Vol. 228, *Lecture Notes in Business Information Processing*, pp. 308–320.
2. Lex, E.; Voelske, M.; Errecalde, M.; Ferretti, E.; Cagnina, L.; Horn, C.; Stein, B.; Granitzer, M. Measuring the quality of web content using factual information. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12* **2012**, p. 7.
3. Xu, Y.; Luo, T. Measuring article quality in Wikipedia: Lexical clue model. Web Society (SWS), 2011 3rd Symposium on. IEEE, 2011, pp. 141–146.
4. Wu, G.; Harrigan, M.; Cunningham, P. Characterizing Wikipedia Pages Using Edit Network Motif Profiles. Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents; ACM: New York, NY, USA, 2011; SMUC '11, pp. 45–52.
5. Suzuki, Y.; Nakamura, S. Assessing the Quality of Wikipedia Editors Through Crowdsourcing. Proceedings of the 25th International Conference Companion on World Wide Web; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, Switzerland, 2016; WWW '16 Companion, pp. 1001–1006.
6. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis WikiSym 07* **2007**, pp. 157–164.
7. Ingawale, M.; Dutta, A.; Roy, R.; Seetharaman, P. Network analysis of user generated content quality in Wikipedia. *Online Information Review* **2013**, *37*, 602–619.
8. Halfaker, A.; Taraborelli, D. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/, accessed on 2017-12-31.
9. Blumenstock, J.E. Size matters: word count as a measure of quality on wikipedia. WWW, 2008, pp. 1095–1096.
10. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Quality and Importance of Wikipedia Articles in Different Languages. In *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings*; Springer International Publishing: Cham, 2016; pp. 613–624.

11. Warncke-wang, M.; Cosley, D.; Riedl, J. Tell Me More : An Actionable Quality Model for Wikipedia. WikiSym 2013, 2013, pp. 1–10.
12. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* **2017**, *4*.
13. Anderka, M. Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany, 2013.
14. Dalip, D.H.; Lima, H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Quality assessment of collaborative content with minimal information. IEEE/ACM Joint Conference on Digital Libraries, 2014, pp. 201–210.
15. Dang, Q.V.; Ignat, C.L. Quality assessment of Wikipedia articles without feature engineering. 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, pp. 27–30.
16. Lipka, N.; Stein, B. Identifying Featured Articles in Wikipedia: Writing Style Matters. *Proceedings of the 19th International Conference on World Wide Web (2010)* **2010**, pp. 1147–1148.
17. Lewoniewski, W.; Khairova, N.; Węcel, K.; Stratiienko, N.; Abramowicz, W., Using Morphological and Semantic Features for the Quality Assessment of Russian Wikipedia. In *Information and Software Technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings*; Damaševičius, R.; Mikašytė, V., Eds.; Springer International Publishing: Cham, 2017; pp. 550–560.
18. Wolinski, M.; Milkowski, M.; Ogrodniczuk, M.; Przepiórkowski, A. PoliMorf: a (not so) new open morphological dictionary for Polish. LREC, 2012, pp. 860–864.
19. Woliński, M. System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica XXII–XXIII* **2003**, pp. 39–55.
20. Gruszczyński, W.; Broda, B.; Charzyńska, E.; Dębowski, u.; Hadryan, M.; Nitoń, B.; Ogrodniczuk, M. Measuring readability of Polish texts. Proceedings of the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 2015.
21. Broda, B.; Piasecki, M. Parallel, massive processing in SuperMatrix: a general tool for distributional semantic analysis of corpora. *International Journal of Data Mining, Modelling and Management* **2013**, *5*, 1–19.