Models for Arabic Document Quality Assessment

Adnan Yahya and Afnan Ahmad and Alaa Assaf and Rawan Khater and Ali Salhi
Electrical and Computer Department, Birzeit University, Birzeit, Palestine
yahya@birzeit.edu

Abstract. Digital content has been increasing rapidly. This content can be generated, accessed and used by anyone and thus the need for quality assessment of web content before usage becomes an important issue. Devising methods to assess quality of Arabic digital content is the focus of this paper. Our work was partially based on Wikipedia articles annotated into *featured* and *good* according to quality guidelines of the Wikipedia. Our analysis was directed at finding features that can serve as best quality indicators. Using the defined features we trained a high accuracy quality assessment model using machine-learning algorithms. Our work went beyond the Wikipedia documents to build a general model that can assess the quality of Arabic documents that lack Wikipedia metadata with acceptable accuracy. The model was trained and built using features from documents we collected from Arabic online news sites and blogs, and annotated in collaboration with university students.

Keywords: Document Quality Assessment, Arabic Wikipedia, Arabic Information Retrieval.

1 Introduction:

Due to the diversity of web content and the ease of posting on the web, one can expect diversity in web information quality and degree of trust. One cannot give the same trust to a social media post and an article in a well-known newspaper. The need for quality assessment of web content is paramount. But this is not a trivial task: manual quality annotation does not scale, so automatic quality assessment is needed.

In our research, we worked to build a model to assess Arabic document quality, first for the domain of Arabic Wikipedia characterized by abundant metadata but also large quality variations[10], then for general Arabic documents, that may lack Wikipedia style metadata.

Wikipedia has its own assessment system that classifies articles into quality classes according to specific criteria using manual judgement through a peer review process. The best articles are "feature" articles[1], after that there are the "good" articles[2] that don't meet the criteria of featured articles but still of high enough quality. The documents that are already assessed as high quality (featured or good) by human experts constitute a small fraction of the Arabic Wikipedia and will serve as the annotated data to train our Machine Learning (ML) based models for quality assessment of Wikipedia articles.

After building the Wikipedia model with high accuracy, we collected documents from different popular online blogs and news sites. Those documents do not have meta-data and are not classified based on quality. We manually annotated a number of such articles and trained a general model to assess document quality using features available for any article. We treated the quality level as a classification problem to place the article into one of the two quality classes: *high quality* and *Random*[2].

For ML we experimented with several algorithms and Support Vector Machine –SVM- (Sequential Minimal Optimization SMO implementation) and Random Forest were the main classifiers adopted for our models. As usual in machine learning, four measures for the classification effectiveness: *Precision* (P), *Recall* (R), *F-measure* (or F-score) and *Accuracy* (A) which can be read from the confusion matrix of individual experiments.
The rest of the paper is organized as follows. In section 2 we discuss related work on article quality assessment, general and Arabic. In section 3 we outline our ML approach to article quality assessment. I section 4 we give the results of our experiments for Wikipedia and General articles quality assessment and the role of features on the results. In section 5 we give our conclusions and point to possible directions of future research.

2 Related Work

2.1 Article Quality Assessment

Studying the quality of web content, and specially the Wikipedia articles quality, and the best information quality metrics and features that can best capture the degree of article quality has been under extensive research. In [4], Blumenstock discusses the simplest metric that can be used to classify the articles of English Wikipedia to *featured* (the highest quality articles in Wikipedia) and *random* articles. Blumenstock got 96.31% accuracy when applying binary classification on his dataset depending on a specific threshold of word count equals to 2000 words, but as expected, this method has drawbacks and can be fooled easily. Lipka and Stein[9] give a more advanced step in identifying articles by analyzing writing style feature which is character tri-gram. Their results improve on the word count (naïve) approach. Yahya and Salhi studied Arabic Wikipedia articles quality with emphasis on features from 3 groups; textual content features, non-textual content features and features related to contributors and editors but didn't use that to build models for quality based classification using machine learning[16].
Stivilia, Twidale, Smith and Gasser present seven information quality metrics that is Authority, completeness, complexity, informativeness, consistency, currency and volatility[13]. They define these metrics using 19 statistical measures from both the content and metadata of articles. The experiments show that the developed model can capture big differences between featured and random articles[13]. In [14], Warncke-Wrang, Cosley and Riedl did many experiments and investigations after Stivilia [12] to come up with an actionable model for assessing Wikipedia articles and it worked with completeness, informativeness, number of headings, article length and number of references features. The results of this model are close to the other models, which have larger number of features. In [7], De La Calzada and Dekhtyar argue that not all articles in Wikipedia are the same and define 2 article categories: *stabilized* and *controversial* articles and use different models to measure the quality of article in each category. The stabilized model uses measures related to the structure and construction of article while controversial model depends on the history log of revisions for article. Lim, Vuong, Lauw and Sun present two quality models: the basic model and peer review model to measure the quality of article depending on the authorities of its contributors[8]. In [5] it is argued that the quality of content of medical Web documents is affected by domain features and use specific vocabulary and codes and document type for improved results. In [15] the authors attempt to improve the quality of DBpedia by analyzing features and models that can be used to evaluate the quality of articles, providing foundation for the relative quality assessment of infobox attributes. In [11] a combination of the usual features and deep learning for better results in Wikipedia article quality assessment is used. In [6] Wikipedia article quality is assessed by analyzing article content rather than the

feature set and utilizes NLP deep learning to achieve the reported results. Some of the features used in the surveyed literature are easy to calculate while the others are more sophisticated.
We tried our best to find the best combination from different models and studied the importance and effect of these features in the quality of Arabic articles. We propose new features not mentioned in the above research which we found useful in the case of general documents.

3 Our Article Quality Assessment Approach:

3.1 Classification Models Using Machine Learning (ML):

The classification problem can be described as the process of choosing the category an instance belongs to from a list of categories. The main steps used to solve the classification problem and building a classification model using machine learning algorithms (implemented using software like WEKA) is as follows:
1. An annotated dataset (with instances classified by humans) is needed to train the model.
2. Feature extraction: the training dataset must be presented as features related to the classes of the model. When the classes of the dataset are related to quality (high quality and low quality), then features related to quality must be extracted from each instance to represent the dataset elements.
3. Then a classification algorithm (classifier) is used on the extracted features for each instance in the dataset to build a model that can classify any external instance.
4. After that the model must be tested (can be tested using external testing set or using cross-validation method) to get the evaluation measurements about the performance and decide if the model is acceptable or needs fine tuning.

Figure 1 shows the basic steps in building a classification model using ML. It is used in building the two quality assessment models for Wikipedia and general articles reported here.
Next we describe the two machine learning models we built, first for Arabic Wikipedia articles with extensive metadata then for general Arabic documents.

3.2 Wikipedia Based Quality Model.

We experimented with features related to the textual content of articles like writing style, spelling errors and the metadata provided with them like links, multimedia content, edits, contributors and authors and social media effect. Then the best combination of features was selected to help in building a high accuracy quality assessment model.

Naïve Approach: Word Count. We started with the simplest method, referred to as the *Naïve Approach* using word count as the sole measure for quality classification[4]. We applied this method to a balanced dataset with four classes of articles "Feature", "Good", "Random for feature" and "Random for good". We found that the average length of featured articles is 9176 words, for good class articles is 4694, for HQ class (featured and good combined) is 7038 and for the random class it is 653 words (Table 1). These numbers indicate that the word count seems to be a good indicator for article quality.

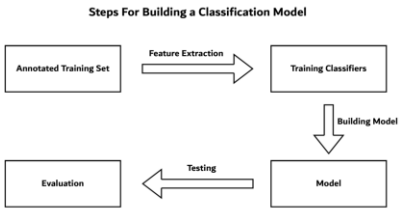

Fig. 1. Steps to Build a Classification Model.

Class	Word Count
Feature	9176
Good	4694
High Quality (Featured + Good)	7038
Random for Feature	749
Random for Good	548
Random for High Quality	653

Table 1. Average Word Count for Quality Class.

We were interested in calculating the threshold value for word count that gives minimum classification error rate for the two class case: *High Quality* if the word count exceeds this threshold, *Random* otherwise. We considered word count threshold levels from 1000 to 3000 and the accuracy was calculated. Figure 4.2 shows the results. The threshold that leads to minimum error rate is 2100 words.

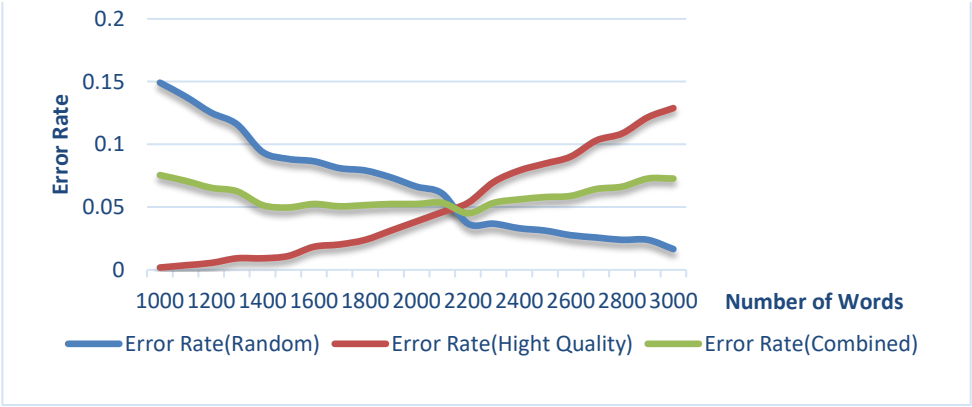

Fig. 2. Error Rate for Classification by Word Count.

The article word count was also used as a feature to train a classifier using the WEKA for the two classes High Quality vs. Random and Featured vs. Good. Many classifiers were applied in different modes of training and testing (10 folds cross validation, splitting data at 66%). As expected the classifiers predicted class with high accuracy reaching 97.5% in the case of High Quality vs. Random since the feature of word count is very distinguishable between them. For Featured vs Good case, the accuracy dropped to 83%.

Building our Dataset. We wrote a PHP code with Media-Wiki APIs[3] to extract 309 Featured and 305 Good articles[1] from the Arabic Wikipedia. We considered two types of random articles:

[1]The numbers when we built our dataset. The current numbers are: 620 and 634, respectively.

the first is arbitrary length (AL) random articles, the other has random articles close in length (CL) to the two high quality articles to reduce the effect of article length feature. To maintain balance, we extracted 309 articles and called the set "random for feature" and another 305 called "random for good". For the AL case the selection was random, and for the CL case the selected articles were of comparable length and length distribution. Thus we selected a total of 1228 random (unknown quality) articles.

Analyzed Features: We studied the features in three main categories: textual features, non-textual features and contributors and editors. The following features were calculated for feature, good, random-for-featured (with close length to feature), random-for-good (with close length to good), random-for-featured (any length), random-for-good (any length). The features were retrieved/calculated using PHP code written with the Wikipedia APIs.

Textual Features: Textual content is the basic parameter to analyze in each article; we can use it to find specific parameters that may reflect the writing style, spelling errors and other quality features. The following textual parameters were extracted and studied:

1. Article Length in words.
2. Average Sentence Length: the sentence is a sequence of words ending with one of the following punctuation marks dot " . " , question mark " ? " or exclamation mark " ! ".
3. Average Number of Words per Comma.
4. Number of Paragraphs.
5. Average Paragraph Length (in Words).

We also considered Arabic confusion letters frequencies. Confusion letters are the Arabic letters that cause confusion to the writer and lead to spelling errors, these letters are "أ","ئ","ء","إ","ؤ","آ","ا","ى","ي","ه","ة". The results didn't support our assumption that these letters and their relative frequencies are different for various quality classes.

Non-Textual Features: We considered non-textual features and wrote the code to extract them from Wikipedia html pages using Media-Wiki APIs. These features were:

1. Categories: Wikipedia articles can be tagged to one or more categories related to the topic of article. We think that the number of categories an article tagged to may serve as a quality measure as it could be an indicator for the degree of article specialization, so we extracted all categories for each article in the training set.
2. Links: Links are among the most important features in studying any web content. Wikipedia itself has many types of links that can be analyzed to find if any of them has an effect on article quality. The links in Wikipedia that we extracted are:
   a. Language Links: links that connect an article with parallel articles in other languages.
   b. External Links: represent links that redirect users from a Wikipedia article to locations outside the Wikipedia. This feature indicates how much an article is connected to, and supported by, other web content.
   c. Internal Links: links going from a specific Wikipedia article to other Wikipedia articles.
   d. Back links: links that come from Wikipedia articles to a specific Wikipedia article. Back links may indicate trust in content.

3. Multimedia Content: Multimedia content consists of tables, pictures, videos and sounds in articles. We conjectured that articles with multimedia content are expected to be of higher quality since multimedia elements can serve to support the article content.
4. Number of High Quality Articles Related to an Article in other Languages: As we know, each article in the Wikipedia is connected to other articles related in content but in different languages using language links. We think that if an article is considered of high quality in other languages that will support the writers and help them in improving the quality of the article. Therefore, the more high quality articles related to an article in other languages the better quality an article has.
5. Number of References: This parameter can be considered as one of the most important indicators of quality. More references may mean more work and effort the writer has done, and the more trust in the facts of the article.
6. Number of Shares in Social Media (Facebook, Twitter and Google+): Social Media is an essential part of today's life. Therefore, we think that the shares of an article in such social media may be an indicator of quality. Code for counting the number of shares was written in PHP and using the APIs of Facebook, Twitter and Google+ offers.
7. Age of Article: We looked at the age of Wikipedia article as a quality indicator assuming that high quality articles must have an age that allows for modifications and improvements.

Edits and Contributors: Contributors to Wikipedia can be registered users, anonymous users or bots. Wikipedia stores all the users and their contributions to articles in the history log (known as revisions of articles) which can be a source of information about article contributors, edits and their numbers, sizes and dates. The following features were studied:

1. Number of Contributors: anonymous, registered and bot contributors, the latter being programs that perform specific actions to articles such as spelling correction.
2. Average Edit Size: The average edit size for each article was computed from the history logs.
3. Number of High Quality Articles Related to an Article by the Authors: We know that the writer expertise can reflect strongly on the quality of his/her writing. The more high quality content an author publishes the more expertise he/she has. We took all the contributors to featured and good articles and for each author we counted the number of high quality articles he/she has contributed to, so we got a list that has all the authors and the number of high quality articles each author contributed to. Then we computed the sum the number of high quality articles related to an article by its authors. This was implemented using PHP and the Media-Wiki APIs.

3.3 General Article Quality Model:

After building the Wikipedia based model, we moved to building a model to assess the quality of any text document. For that we needed a dataset with articles from different sources to neutralize the effects of Wikipedia authoring guidelines, and limited features to those related to textual content alone to make the results easier applicable to any document.
Next we describe the dataset we collected and the features we studied and analyzed for use in the general article model.

General Dataset Collection: We collected articles in modern standard Arabic from popular news sites (https://www.alquds.co.uk/, https://www.aljazeera.net/, https://www.egyptwindow.net/, http://www.alriyadh.com/, https://elaph.com/ and prominent bloggers with large enough articles (http://ahmedjedou.blogspot.com/, https://www.essamalzamel.com/). The average word count was about 1050 words per document. Then we set to label the collected articles manually by volunteer university students. Each article had at least three assessments and used a threshold for the average evaluation score to get a balanced dataset of 112 articles labeled as High Quality (HQ) and 113 articles labeled as Low Quality (Random).

Analyzed Features: The features we analyzed for the new dataset articles are:

Features Used in the Wikipedia Model: We used the same textual features used in the Wikipedia model available for general documents: average sentence length, number of commas, number of words, number of sentences, average number of words per comma and number of shares.

Other Features: The other features we used for the new model are:

1. Title Relation with Article Body: this feature describes the relative degree of overlap between the title to the body of article. We calculated it by counting the number of occurrences of the title words in the article body.
2. Part of Speech (POS) Tagging: Each word in Arabic sentence can be classified to Verb, Noun, Pronoun, Adverb, Adjective and Numbers or others describing the word part of speech. We decided to find the distribution of each type of these parts of speech in the quality classes we have to find if that affects model results. To tag article words we used Stanford POS Tagger[12], which is a widely used open source tagger. We fed the sentences of each article and got the tag of each word. Then we computed the percentage of the occurrence of each tag in the article and considered them as attributes.

4 Results and Discussion:

In this section, we present an analysis of features studied and the results using machine learning models using WEKA on features mentioned in the previous section for both models.

4.1 Wikipedia Model Features and Classification Results:

First we present the results of studying the features mentioned in the previous chapter for the Arabic Wikipedia articles, and then we will talk about the WEKA classification results for the quality assessment model for Arabic Wikipedia articles.

Results for Textual Features: Table 2 contains the results of analyzing the textual content features which are: word count, number of paragraphs, paragraph length, number of words per comma and sentence length. The average word count for the high quality class was found to be 4914 words, for the random with close length (CL) to high quality articles the word count was 4889 words as expected, with the real random articles (any length) the average word count was 1261 words. This reflects the results of the naïve approach discussed earlier.

For the number of paragraphs and paragraph length we can see that the result is distinguishable between the high quality class (with 57 paragraphs and 85 words per paragraph) and the random of any length class (with 19 paragraphs and 68 words per paragraph), but averages with random close in length is very close to high quality (with 57 paragraphs and 86 words per paragraph).
The high quality class has a lower number of words per comma compared to the two random classes (high quality with 18 words per comma, the two random classes with 23words per comma). Sentence length seems not distinguishable between the three classes with HQ having 27 words and the two random classes with 26 words.
For the case of high quality and close length random articles the results of analyzing the texts are close, probably reflecting a common style of writing for the Wikipedia enforced by editors or it may indicate that the effort used to write long articles is close to the effort used to write high quality articles. However, in the case of random articles of any length the results show differences in some features.
Though not detailed here, for Arabic confusion letters we found that the relative frequencies and distribution of shapes for high quality class are very close to that for the random classes so we can say that the articles in Wikipedia do not have spelling errors related to confusion letters.

Quality Class → Feature ↓	**HQ-High Quality**	**Random CL**	**Random AL**
Word Count	4914	4889.00	1261.00
Paragraphs	57.46	56.99	18.62
Paragraph Length	85.52	85.80	67.74
Words/Comma	18.36	22.88	23.15
Sentence Length	26.85	26.26	25.84
Language Links	60.34	66.70	19.47
Categories	13.13	10.62	6.74
External Links	70.70	33.28	6.40
Internal Links	382.00	271	93.00
Back Links	459.00	394	118.00
Multimedia Elements	25.24	12.37	3.63
References	126.20	50.11	7.76
Others HQ	2.32	1.65	0.27
Shares of Article	47.20	71.67	8.32
Age of Article	2587.00	2742.00	1856.00
Bot Contributors	26.39	26.82	21.98
Registered Contributors	40.95	42.32	29.01
Anonymous Contributors	46.03	58.72	26.95
Bot Edits	82.44	82.85	61.04
Registered Users Edits	26.39	26.82	21.98
Anonymous Users Edits	40.95	42.32	29.01
Edit Size (Bytes)	46.03	58.72	26.95
NHQAA	3154.00	2435.00	738.00

Table 2. Select Features for Wikipedia High Quality, Random of Comparable-Length (CL) &Random Any-Length (AL).

Results for Non-Textual Features: Table 2 also contains the results for analyzing the non-textual features which are: References, Number of High Quality Articles Related to an Article in Other Languages (OthersHQ), Shares, Age of Article.

From the results in the above table, we can see that the average number of language links for high quality articles is 60 and for random with close length (CL) is 66 while the random with any length (AL) is only 19 links.

Average number of categories for high quality is the highest with 13 categories while the two random classes score smaller results with 11 categories for random with close length and 7 for random with any length.

For external links, the high quality class has an average of 71 links while the random with close length has 33 links and the random with any length has only 6 links.

For internal links, the high quality class has an average of 382 links while the random CL has 271 links and the random with any length has only 94 links.

For Back links, high quality class has an average of 459 links while the random with close length has 394 links and the random with any length has only 118 links.

For multimedia elements, HQ class has in average 25 elements while the random CL has 12 elements and the random with any length has only 4 elements.

For the number of references, we can see that the average number of references forHQ is 126.2 and for random with close length 50.11 while the random with any length is only 7.76.

The average for the Number of HQ Articles Related to an Article in Other Languages for HQ is 2.32 and for random with close length 1.65 while the random with any length is only 0.27.

The average number of shares for high quality articles is 47.24 and for random with close length 71.67 while the random with any length is only 8.32

The average age of article for high quality is 2587 days and for random with close length 2742 while the random with any length it is only 1856 days.

In summary, Internal, External, Back links, multimedia, references, Number of Related High Quality Articles in Other Languages and number of shares are different for high quality class and the two random classes. These features can be used as strong quality indicators.

Results for Edits and Contributors: The last part of Table 2 contains the results of calculating the averages for contributors, edits, edit size and Number of High Quality Articles Related to an Article by the Authors (NHQAA).

From Table 2, we can see that the average number of Bot contributors for high quality is 26 and for random with close length 27 while the random with any length is only 11. The average number of Registered contributors for HQ is 41 and for random with close length is 42 while the random with any length is only 11.

The average number of Anonymous contributors for high quality articles is 46 and for random with close length is 57 while the random with any length is only 12. The number of contributors between the random with close length is very close to the HQ class with a little difference in the number of anonymous contributors. While random with any length has a very different average, which may be directly related to the length of articles.

For Bot edits, HQ class has an average of 82 edits while the random with close length has 83 edits and the random with any length has only 24 edits. For Registered edits, HQ class has in average 296 edits while the random with close length has 188 edits and the random with any length has only 38 edits. Regarding Edit Size, HQ class has in average 761 bytes edit while the random with close length has 781 bytes and the random with any length has only 297 bytes.

For Anonymous users edits, HQ class has in average 72 edits while the random with close length has 94 edits and the random with any length has only 19 edits.

The average number for Number of High Quality Articles Related to an Article by the Authors (NHQAA) for high quality is 3154 and for random with close length 2435 while the random with any length is only 738. This may mean that this feature, as an indicator of writing quality of the authors, can play a big role as a quality indicator of articles.

As seen for the edits between the random with close length and the HQ class, edits from bots are on average the same while edits from registered and anonymous users show differences. We can note that high quality articles have a higher number of edits from registered users compared with the two random classes.
To conclude our analysis, we can see clearly that the number of Anonymous Contributors, the Edit Size and the Number of High Quality Articles Related to an Article by the Authors (NHQAA) are the main features with big differences between High Quality and Random articles and can thus be considered as the stronger quality indicators compared to the other features.

Wikipedia Model ML WEKA Results: We trained the model with extracted features, did many experiments with different feature combinations and different classifiers.
Using 10-fold cross validation to evaluate the performance and Random Forest (RF) Classifier with default options with textual features only, the resulting model for the High Quality and Random with close length gave an accuracy of 68.5%, reflecting neutralization of article length. With the non-textual and the editors features the accuracy increased to 88.3%, a 20% improvement. The number of references played the main role in that (the gain was 13% while for the other features it was 5-9% only).
When using SMO with normalized kernel and all features we reached an accuracy of 89.8%, which slightly better than the result for RF.
After many experiments with many combinations and classifiers, we reached the models with the best performance as shown in Table 3.
In the first Wikipedia model (High Quality with Random) we notice that the values for precision, recall, F-measure and accuracy using SMO classifier (0.9, 0.898, 0.898 and 89.82%, respectively) are higher than those for Random Forest.
In the second Wikipedia model (High Quality with any Length Random) the values for precision, recall, F-measure and accuracy measurements using Random Forest classifier (0.956, 0.955, 0. 955 and 95.50%, respectively) are higher than for the SMO classifier.

Model	Close-Length Model		Any-Length Model	
Classifier →	RF*	SMO**	R F*	SMO**
Precision	0.888	0.900	0.956	0.955
Recall	0.884	0.898	0.955	0.954
F-Measure	0.883	0.898	0.955	0.954
Accuracy %	0.884	0.898	0.955	0.954
*Random Forest		** Normalized Poly Kernel 9 Exponent		

Table 3. Best Models Performance

We ran our model on the new Wikipedia high quality articles listed as *featured* and *good* in the four months following the original dataset collection (we found about 50 such articles). We used these 50 article to test the model and it classified 43of the 50 instances as high quality, consistent with the accuracy using cross validation.
A simple application to classify Wikipedia article based on our model was developed using HTML, PHP and JAVA.

4.2 General Model Features and Classification Results:

Next, we present feature analysis for the general, nonWikipedia model, and its WEKA results. Table 4 contains the results for textual features: word count, sentence count, sentence length, Comma Count, words per comma, shares and title/content relation. As we can see from the averages for each class, the average word count for the HQ class is 949 words, against 1172 for the random, quite a departure from Wikipedia articles dataset relative sizes. The Sentence count

average is 18.7 words for HQ, and 34.2 words for Random, while the sentence length for HQ is 144 words, while it is 42 in Random, a substantial difference on both counts.
Comma count is close between the two classes; 57.6 for HQ and 69.5 for Random.
Average number of words per comma in the HQ class is 52.1, while it is 30.1in Random. The number of shares in the HQ is 232 against 134 in Random.
The relation between title and content is 40.1 for the HQ against 65.6 for random.
From the above results, we can see that all the average numbers of the features are distinguishable between the two classes and can be used as good quality indicators.
The POS tagging results are shown in the Table 5.

Class	**High Quality**	**Random**
Word Count	949.0	1172.0
Sentence Count	18.7	34.2
Sentence Length	144.0	42.0
Comma Count	57.6	69.5
Words per Comma	52.1	30.1
Shares	232.0	134.0
Title/Content	40.1	65.6

Table 4. Textual Features for the General Model

Class	**High Quality**	**Random**
Nouns	0.6147	0.6071
Adjective	0.1286	0.1219
Adverb	0.0052	0.0056
Verb	0.1065	0.1137
Pronouns	0.0224	0.0244
Numbers and Others	0.0304	0.0361
Harf Jar	0.0921	0.0913

Table 5. POS Tags Relative Frequencies.

General Model WEKA Results: When we trained the general model using Random Forest classifier with default options and 10- fold cross validation we noticed that the accuracy for the model with textual features was only 74.5%. When the Title/content attribute was inserted the accuracy improved by 5.5% to reach 80%. When we inserted different combinations of tags we found that the nouns, adjective, pronouns, numbers and others had the most effect. The accuracy for the model with these features reached 84.5%, an improvement of 4.5%. SMO classifier performed worse. Table 6 shows a summary of WEKA results for the General Model.

Classifier	Precision	Recall	F-Measure	Accuracy %
RF	0.845	0.845	0.845	84.5
SMO	0.727	0.725	0.724	72.5

Table 5. WEKA Results for the General Model.

After that, we tested our model with 25 external articles from newspapers and blogs (13 High Quality and 12 Low Quality) the model classified them with 80% accuracy.
The last test we did was to test the general model on Wikipedia articles; we tested it on 100 articles (50 high quality and 50 random). The model had 78% accuracy for this testing case.

5 Conclusions and Future Work

We developed models for Arabic article quality assessment in the presence and absence of extensive Wikipedia-style metadata. The algorithms achieved reasonable results. The basic limitation we had from the beginning of our work was the scarcity of annotated articles for training. The combination of Wikipedia articles and own annotated articles helped solve this problem. We succeeded in building a good model to classify Wikipedia and general articles based on textual properties, Wikipedia metadata when available, and on other general properties. The idea of determining the quality of any Arabic article can help a lot in many fields like education, media, improving Arabic text content and others.

Our research can be improved in many ways. Increase the training data to be more general, and incorporate newly evaluated data (Wikipedia and general) which can open the door for researching more features as quality indicators. One can also apply the work to different types of web data: dialectal or mixed texts, shorter posts like tweets and other social media posts. One can look into many other features as potential quality indicators like the spelling errors, Arabic writing patterns, foreign language content, the use of dialect in writing, reference quality, mentions in more academic and general social media applications, document recency, site/author trust, Wikipedia infobox properties and access patterns. Deep learning may also be a promising technology for quality assessment of Arabic articles. The quality of shorter Arabic web content like tweets, comments and social media posts may be another interesting research issue to tackle.

References

1. Featured Articles, Arabic Wikipedia ويكيبيديا:مقالات مختارة, http://ar.wikipedia.org/wiki/ويكيبيديا:مقالات_مختارة, Accessed 1/11/2014, 12/3/2020.
2. Good Articles, Arabic Wikipedia ويكيبيديا:مقالات جيدة , http://ar.wikipedia.org/wiki/ويكيبيديا:مقالات_جيدة, Accessed 1/11/2014, 12/3/2020.
3. API: Main_page ,http://www.mediawiki.org/wiki/API:Main_page, Accessed: 2/5/2014.
4. Blumenstock. J. "Size Matters: Word Count as a Measure of Quality of Wikipedia." *Proceedings of the 17th International Conference on WWW*, pp. 1095-1096. ACM, 2008.
5. Cozza V., Petrocchi M., Spognardi A. (2016) A Matter of Words: NLP for Quality Evaluation of Wikipedia Medical Articles. In: Bozzon A., Cudre-Maroux P., Pautasso C. (eds) Web Engineering. ICWE 2016. Lecture Notes in Computer Science, vol 9671. Springer. 2016.
6. Dang, Q., Ignat C. "Quality Assessment of Wikipedia Articles without Feature Engineering". *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, Jun 2016, Newark, NJ. pp.27-30
7. De La Calzada G., Dekhtyar A. "On measuring the quality of Wikipedia articles." In Proceedings of the 4th workshop on Information credibility, pp. 11-18. ACM, 2010.
8. Lim, E., Vuong B., Lauw H., Sun A. "Measuring Qualities of Articles Contributed by Online Communities." In *Web Intelligence*, pp. 81-87. 2006.
9. Lipka N., Benno S. "Identifying featured articles in Wikipedia: writing style matters." *Proceedings of the 19th international conference on www*, pp. 1147-1148. ACM, 2010.
10. Safa A.F., "Enhancing Quality Arabic Web Content through Cross Lingual Information Retrieval Methods". Master Thesis. Birzeit University, Palestine. 2019.
11. Shen, A., Qi, J., Baldwin, T. "A Hybrid Model for Quality Assessment of Wikipedia Articles". *Proceedings of Australasian Language Technology Association Workshop*, pp. 43-52. 2017
12. Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml,
13. Stvilia B., Twidale, M, Smith L., Gasser, L. "Assessing Information Quality of a Community-Based Encyclopedia." In Proceedings of the ICIQ, pp. 442-454. 2005.
14. M. Warncke-Wang M., Cosley D., Riedl, J. "Tell me More: an Actionable Quality Model for Wikipedia." *Proceedings of the 9th International Symposium on Open Collaboration*, ACM, 2013.
15. Węcel K., Lewoniewski W. (2015) Modelling the Quality of Attributes in Wikipedia Infoboxes. In: Abramowicz W. (eds) Business Information Systems Workshops. BIS 2015. Lecture Notes in Business Information Processing, vol 228. Springer.
16. Yahya A., Salhi, A. "Quality assessment of Arabic Web Content: The Case of Arabic Wikipedia" *10th International Conference on Innovations in Information Technology (INNOVATIONS-2014)*, pp. 36-41. IEEE, 2014.