Department of Mathematics and Computer Science
Department of Education and Psychology
Freie Universität Berlin

Natural Language Processing
Prof. Dr. Arthur M. Jacobs
Summersemester 2021

Evaluating the quality of Wikipedia articles with NLP

Project
by
Raphael Leuner
5094927

Contents

1 Motivation

Since its funding in 2001, Wikipedia has become the dominant source of information on the planet. The website, one of the most visited in the world [12], is hosted and funded by the non-profit Wikimedia foundation and is today available in more than 300 languages. It has been founded as a platform for sharing user-created and -curated knowledge, primarily online-encyclopedia articles, but is today accompanied by different Wiki projects that aim at enriching the text based Wikipedia articles by, for example, data, historic documents or news.

In modern societies, the dependency on open and reliable information access rises constantly. While Wikipedia has democratized information access, as people do not have to buy an expensive printed encyclopedia, it has also democratized the creation and curation of knowledge, as everybody is able to edit the articles on Wikipedia. The content of these articles can be highly influential. In a poll in 2016, 63% of teachers and 67% of students have indicated that they use Wikipedia as a source for the preparation of lessons or studying at least once a week [2]. It has also been shown that Wikipedia does not just reflect and report on the current state of knowledge, but actually shapes new scientific research findings [13]. Another example for the widespread and uncritical use of Wikipedia articles can be found in the fact that, according to news reports, two of the three candidates for chancellor in the German Federal Election 2021 have copied parts of their books from Wikipedia [4][3]. But politicians as well as companies are not just copying from Wikipedia, but also sometimes writing or editing their own articles [1] with the goal of influencing their perception in the public. This not just violates the required "neutral point of view" [14] of articles, one of the five fundamental principals of Wikipedia, it also questions its function as an independent source of information.

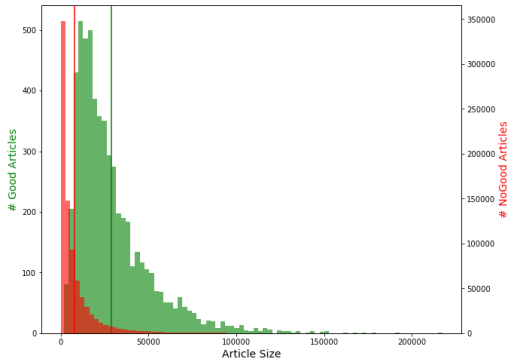
One way, that Wikipedia is trying to deal with its huge responsibility as well as with attempts of deliberately influencing its contents, is by manually marking articles as "good articles" or "featured articles". This annotation is being done manually with a fixed process and set criteria that will be

discussed later. This project attempts to explore the possibility of automating the article evaluation based on methods of Natural Language Processing (NLP).

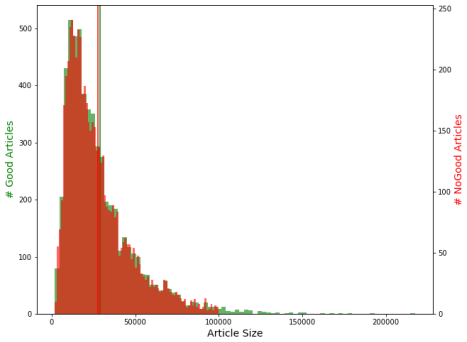
2 What are good Wikipedia articles?

The English speaking Wikipedia is (by September 27, 2021) comprised of 6,383,672 articles. Out of these, 35,106 have been marked as "good articles". Another 5,995 have been marked as "featured articles", for reason of simplicity they will not be attended to in this project [16]. In order for an article to get marked as *good*, it needs to be nominated by an editor. Then, it is being reviewed by another editor against the good article criteria listed below [15]. If the articles are reviewed successfully, they are marked with a green Plus. The criteria for good Wikipedia articles are:

1. Well written:
   (a) the prose is clear, concise, and understandable to an appropriately broad audience; spelling and grammar are correct; and
   (b) it complies with the manual of style guidelines for lead sections, layout, words to watch, fiction, and list incorporation.
2. Verifiable with no original research:
   (a) it contains a list of all references (sources of information), presented in accordance with the layout style guideline;
   (b) all inline citations are from reliable sources, including those for direct quotations, statistics, published opinion, counter-intuitive or controversial statements that are challenged or likely to be challenged, and contentious material relating to living persons—science-based articles should follow the scientific citation guidelines;
   (c) it contains no original research; and
   (d) it contains no copyright violations nor plagiarism.
3. Broad in its coverage:
   (a) it addresses the main aspects of the topic; and
   (b) it stays focused on the topic without going into unnecessary detail.
4. Neutral: it represents viewpoints fairly and without editorial bias, giving due weight to each.
5. Stable: it does not change significantly from day to day because of an ongoing edit war or content dispute.
6. Illustrated, if possible, by media such as images, video, or audio:
   (a) media are tagged with their copyright statuses, and valid fair use rationales are provided for non-free content; and
   (b) media are relevant to the topic, and have suitable captions.



(a) Length distribution of unprocessed Wikipedia data



(b) Length distribution of sampled, balanced data

Figure 1: Length distributions (in characters) of Wikipedia article dataset

3 Data Preprocessing

Wikipedia is a huge text corpus that has been used for numerous NLP projects and applications before. However, no curated dataset of Wikipedia articles exists with the focus on *good* and $no - good$ articles. Therefore, one of the most important parts of this project was the creation of a dataset focusing on these properties.

Wikipedia regularly creates Data dumps of the article texts. The latest available data dump by the time of this project was from September 1st, 2021. It comprises the current text versions of all English language Wikipedia articles as well as metadata, among others whether the article has been classified as *good*. This dump of more than 6 Million articles has a compressed size of more than 18 GB [5].

There is no way of filtering out only *good* articles from this dataset, therefore every article in the dataset needed to be processed and classified as *good* or $no - good$. Due to memory and computing power constraints both for preprocessing and later analysis of the data, this was not done with the entire dataset, but a smaller subset of articles.

During this first step of article selection, list articles (articles that only contain links to other articles), redirect articles (articles that only contain one link to another article) and stub articles (articles that were considered too short by Wikipedia editors) were disregarded. A total of $1,345,180$ articles out of the dataset were processed, out of these $831,367$ were included in the $no - good$ dataset and 5689 were included in the *good* dataset.

In the next step, the length of articles in the *good* and the $no - good$ dataset were compared. As expected, the *good* articles were significantly longer than the $no - good$ articles, with mean number of characters of $7,962$ and $28,865$ respectively. The distribution of article size in character length are presented in figure 1a.

This created a highly imbalanced dataset both in terms of the number of articles in the dataset for each category as well as the distribution of article length. Therefore, the $no - good$ articles were sampled in order to create an article length distribution that resembles that of the *good* articles. Different $no - good$ datasets were created, apart from a balanced one (same number of *good* and $no - good$ articles) also datasets with 2, 5 and 10 times more $no - good$ than *good* articles were sampled. The distribution of article length in the balanced dataset is presented in figure 1b.

In the next step, the texts were preprocessed and cleaned. Wikipedia links were removed as well as headlines and punctuation. Stopwords were removed using the english nltk stopword list and all words were converted to lower case. The words were not lemmatized and stemmed since

in initial trials the prediction models showed a slightly inferior performance on lemmatized and stemmed Wikipedia texts. In a last step, the data was split at random into 80% training and 20% testing data.

4 Methods

4.1 TF-IDF

The primary method to vectorize the texts for classification using different Machine Learning methods was $tf - idf$ [10] [11]. This is one application of a Bag of Words model, in which the frequency of each term in each document is normalized by the inverse document frequency of the number of documents containing the term. For term t, document d and a total number of documents N:

$$tf(t) = (\#t \text{ in } d)/len(d)$$

$$idf(t) = log(N/\#d \text{ in } D \text{ if } t \text{ in } d)$$

$$tfidf(t) = tf(t) * idf(t)$$

Words were omitted if they appeared in more than half of the articles or less than 10 articles. This resulted in a feature matrix containing $68,061$ features. In the feature matrix of the lemmatized and stemmed dataset the feature matrix contained $52,042$ features.

4.2 Machine Learning Methods

Classifications on the $tf - idf$ data were done with the most common machine learning classification tools in order to compare their performances: Bayesian classification, Support Vector Machines, Random Forest, ADABOOST, KNN-classification and Logistic Regression. Out of these, the results for Logistic Regression and Random Forest were explored further in order to analyze the importance of individual features in the model. This was done both to ensure there were no terms left in the articles that would easily make the articles identifiable as *good* (such as additional tags) as well as to see if there were any terms that strongly influence the classification as *good* or *bad*.

For Logistic Regression, this was done using the coefficients of the regression model, for the Random Forest model this was done using the Mean Impurity Decrease (MID) that each feature contributes to the model. Permutation feature importance, which could provide even more exact feature importance scores, was not used as it would have been too computationally expensive for the high number of features [9].

4.3 Glove Model

For a more complex classification model, a Glove model [6] for word embedding was combined with a Deep Neural Network for text classification following the approach suggested here [8] and a workflow adapted from [7]. A Glove network pretrained on Wikipedia data provided with the original paper [6] was used to create embeddings of length 50 for every word. Due to memory and computational constrains, the word vectors used for classification were padded/truncated to a length of 1000. The created list of word vector embeddings is then fed to a Deep Neural Network build out of two bidirectional LSTM layers followed by Fully-Connected layers of size 64 and 2. The network architecture was also taken from [7]. The network was trained using sparse categorical cross-entropy as a loss function for a total of 10 epochs.

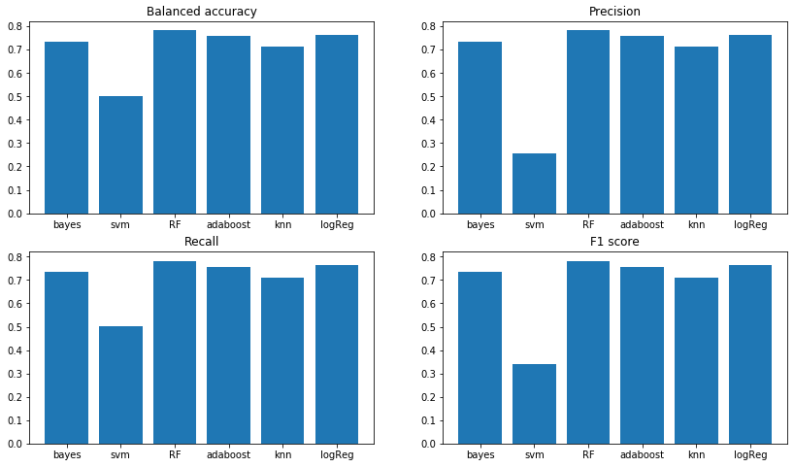


Figure 2: Comparison of classifier performance on the balanced dataset, RF: Random Forest

5 Results

5.1 TF-IDF Model

The combination of $TF - IDF$ text vectorization and the before mentioned Machine Learning methods showed a good performance in classifying the dataset into *good* and $no - good$ articles. The best performance was reached on the balanced dataset by Random Forest and Logistic Regression with accuracies of 0.78 and 0.76 respectively. They are followed by ADABOOST, Bayes Classifier and KNN-classification with 0.75, 0.73 and 0.71. Only Support Vector Machines did not provide comparable classification results. Figure 2 shows a metric comparison of the classification results.

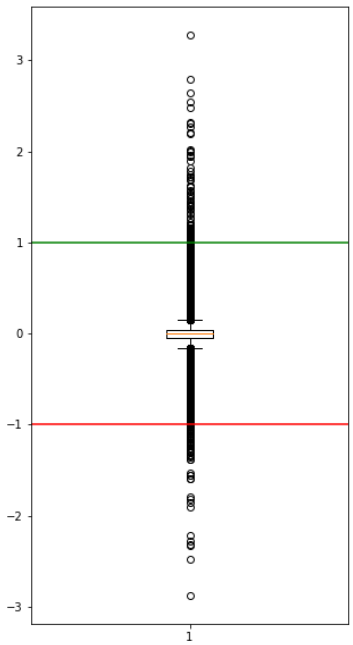
Random Forest and Logistic Regression were also performed on the imbalanced dataset. The results are shown in table 1. Overall, more articles were classified correctly when the classifiers were trained with the more imbalanced datasets, indicated by increasing accuracy. However, less and less "good" articles were classified as "good", indicated by a strong decrease in sensitivity. Therefore, future trials were only made with the balanced dataset.

"no-good"/"good" ratio	1	2	5	10
LR - accuracy	0.76	0.79	0.87	0.92
LR - sensitivity for "good" articles	0.76	0.49	0.29	0.18
RF - accuracy	0.78	0.77	0.86	0.92
RF - sensitivity for "good" articles	0.78	0.38	0.21	0.14

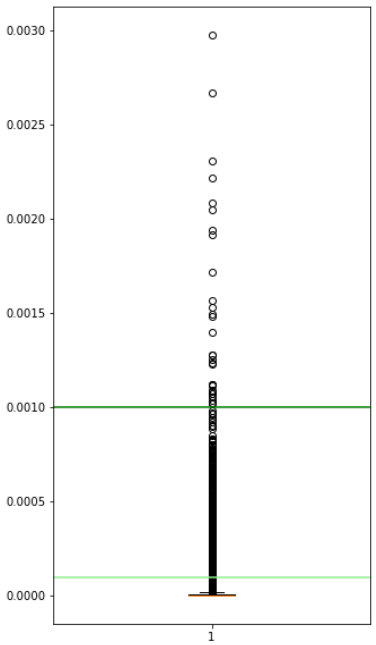
Table 1: Comparison of accuracy and sensitivity for "good" articles between balanced and imbalanced datasets, LR: Logistic Regression, RF: Random Forest

5.1.1 Feature Importance

As described in the Methods section, for both the Random Forest and the Logistic Regression classifier, the feature importance was analyzed. For Logistic Regression, the feature importance is determined by the coefficients of the individual features in the model. They determine, how strongly and in which direction the features (in this case the $tf - idf$ values of individual words) determine the decision. As would be expected, most features are uninformative and therefore have a coefficient of close to 0. The distribution of coefficients is shown in figure 3a. In figure 4 all features are listed, that have a coefficient of larger 1 (87) or smaller -1 (66). The mean coefficent is -0.0006 and the variance is 0.028.



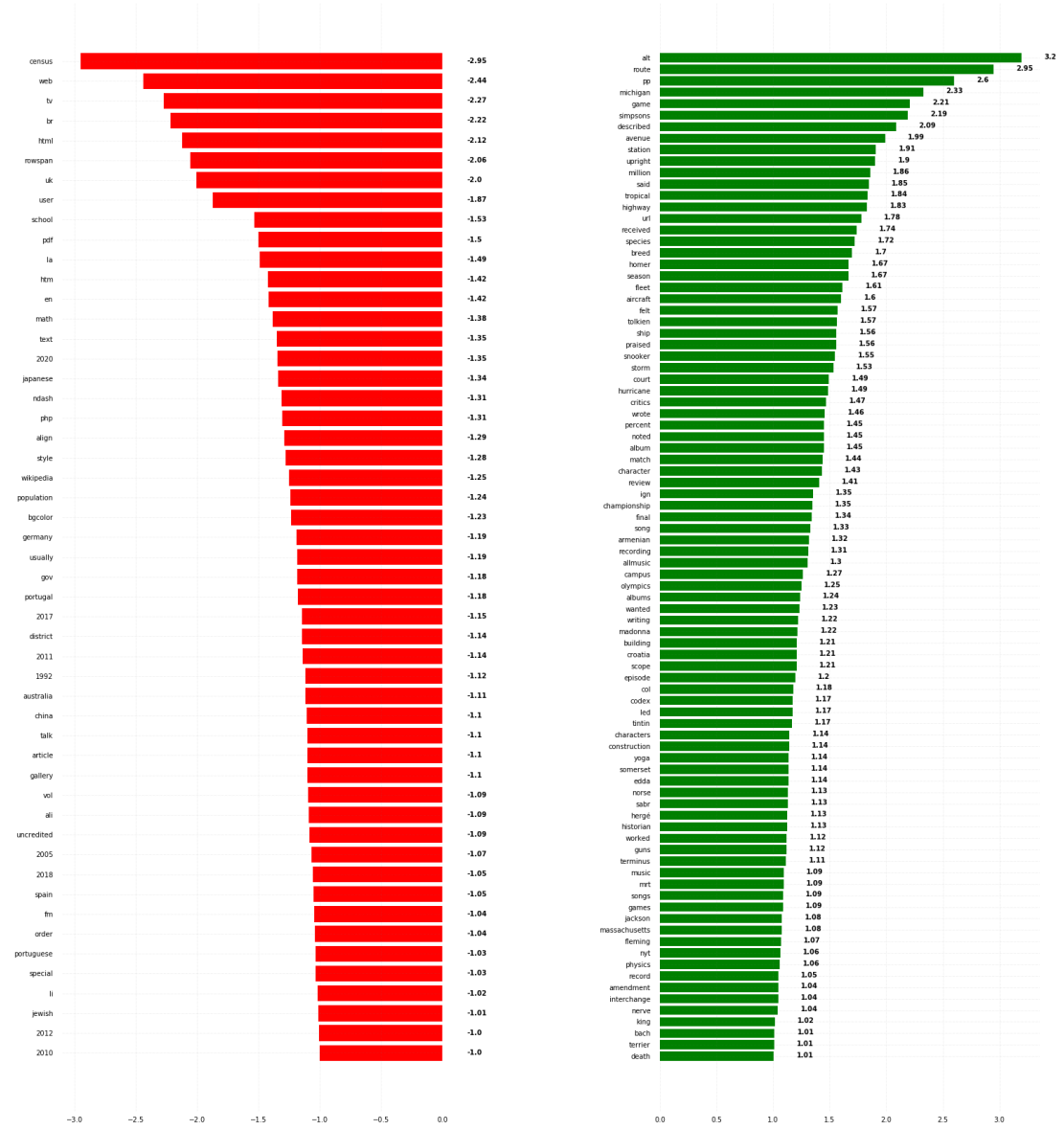
(a) Model coefficients for Logistic Regression. Out of 68,061 features, 87 have coefficients above 1 (green line), 66 below −1 (red line)



(b) Mean Impurity Decrease (MID) for Random Forest. Out of 68,061 features, 2208 have a MID above 0.0001 (light green line), 30 a MID above 0.001 (dark green line)

Figure 3: Boxplots of feature importance for classification on the balanced dataset

For the Random Forest model, the feature importance is based on the Mean Impurity Decrease (MID) that is contributed by every individual feature. Therefore, it does not determine in which direction features contribute to the decision, but instead how much every feature improves the decision making. The MID is low for every individual feature with a maximum MID of 0.003, only 30 features have a MID of above 0.001. The distribution of MID values for all features is shown in figure 3b. The 30 features with highest MID values can be seen in figure 5.



(a) Features indicating a "no-good" article, ordered from highest to lowest impact below -1

(b) Features indicating a "good" article, ordered from highest to lowest impact above 1

Figure 4: Feature importance for the Logistic Regression classification model

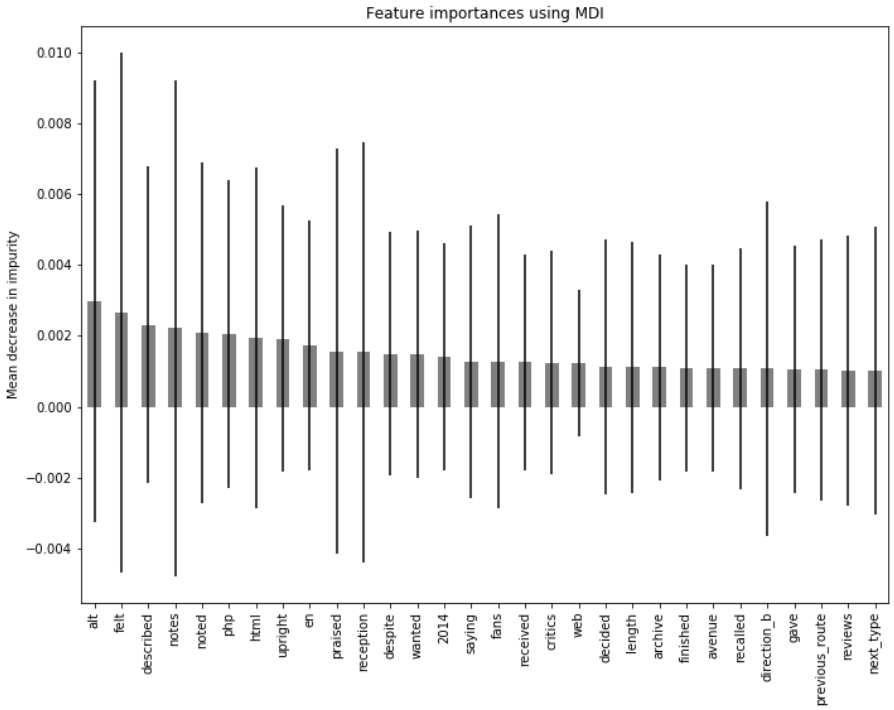


Figure 5: MID (mean and standard deviation) for the 30 most important features in the Random Forest model

5.2 Glove Model

As mentioned before, due to computational constraints, the Glove-DNN classification model was constrained to word embedding vectors of 1000, as with increasing vector length the trainable parameters and therefore the training time of the classification model explodes. Despite this restriction, that for most articles leads to the omission of major parts of the texts, the model reached an accuracy of 0.68. The length of the embedding vectors does play a role in classification performance, in trials with fewer data the classification accuracy increased with increased text embedding length . This shows, that there is future potential for this model, but it requires significantly more computational power than the comparably simpler models based on $TF - IDF$ and "classic" ML-methods.

6 Discussion

6.1 TF-IDF vs. Deep Learning Model

The results between the $TF - IDF$ classification and the Glove-DNN classification model are only partially comparable, as the text vectors for the second model have been restricted. However, the results show that a deterministic model such as $TF - IDF$ can reach quite good performance on the classification tasks and do so with much less computational power than more complex models.

6.2 Feature Importance Analysis

The $TF - IDF$ model also allows for easier analysis of feature importance. Here, two insights are especially interesting. On the one hand, there is not a single feature with very high classification power. That shows that in the preprocessing pipeline all obvious hints (such as tags identifying

articles as *good*) were successfully removed.

On the other hand, the explanation of why exactly these features are the most informative is not trivial. For example, the features indicating *good* articles in figure 4b in the Logistic Regression model contain a lot of terms from (pop)culture, such as "simpsons, homer, tolkien, album, championship, recording, song, madonna". Several explanations for this seem possible, for example there could be fans who are writing especially detailed and exhaustive articles about these items. But it could also be that articles with these topics are more likely to be tagged as *good* independently of the actual quality of the article.

Some terms that are more strongly indicating $no - good$ articles in figure 4a seem counter intuitive. For example the fact that the term "census" has the highest association with $no - good$ articles is surprising, because it is a term generally associated with factual information. It is notable, that a lot of terms associated with $no - good$ articles are related to countries or governments. But there are also some generic terms in this list, such as years or terms like "html, pdf, php, rowspan". Some of these terms can also be found in the feature importance analysis from the Random Forest model (figure 5). These could be artefacts from text cleaning (for example not fully removed hyperlinks) and maybe should be removed explicitly in further preprocessing.

6.3 Which criteria for *good* articles can be analyzed by NLP?

In section 2, the official Wikipedia criteria for good articles has been presented. Not all of these criteria can be analyzed by word embeddings or $tf - idf$ scores. For example the readability can only be analyzed on a word level, not based on sentence or paragraph structures. Some other categories, such as the number and quality of citations and illustrations and the stability of the text (how often it is changed significantly) have deliberately not been analyzed in this project and information about them has been removed during preprocessing.

The fact, that a classification on a word-by-word level on a smaller subsample of the dataset can already reach a classification accuracy of close to 80% is quite striking. Especially, since it "only" covers two of the six criteria (and even those only partially), namely "well written" in the selection of words and terms and the broadness of the coverage of a given topic by word choices. There are many other potential NLP and non-NLP methods to classify Wikipedia articles and a combination of these models could potentially boost performance even further.

7 Further research

The first step for future research would be to validate the results of these smaller trials using the entire set of *good* articles from Wikipedia. For the Glove-DNN model, one potential approach of boosting performance could be to train a Glove or Word2Vec model on the dataset from scratch instead of using a pretrained model and increase the size of the word vectors to be used in the model. Both of these steps will require more computational power.

In order to create a really good predictor for the quality of Wikipedia articles, the word-wise classification model could be combined with an automated analysis of the quality of sources of an article, an analysis of the sentiment of the article in order to check if its written from a neutral point of view as well as an analysis of the number of major changes the article has undergone recently. These models could be combined to a quality score that can give some indication on the trustworthiness of an article. However, even a combined model like this would struggle with identifying wrong facts in an article or, even worse, deliberate manipulation. Therefore, a certain amount of scepticism towards information solely from Wikipedia is still in order and can not be

replaced by automatic article assessment.

8 Code and Data

The Wikipedia dump used for this project is available on the Wikipedia websites [5]. The Python code for this project is available at https://github.com/rlnrbio/WikipediaClassification.

References

[1] Anna Biselli. "Mit freundlichen Edits aus dem Bundestag". In: *Netzpolitik.org* (Sept. 3, 2021). URL: https://netzpolitik.org/2021/wikipedia-edits-aus-dem-bundestag-abgeordnete-wahlkampf/ (visited on 09/13/2021).

[2] Initiative D21. "Lehrwelt, Lernwelt, Lebenswelt: Digitale Bildung im Dreieck SchülerInnen-Eltern-Lehrkräfte". In: (2016). URL: https://initiatived21.de/app/uploads/2017/01/d21_schule_digital2016.pdf (visited on 09/13/2021).

[3] Philippe Debionne. "Plagiatsvorwürfe gegen Annalena Baerbock: Hat sie in ihrem Buch abgeschrieben?" In: *Berliner Zeitung* (June 29, 2021). URL: https://www.berliner-zeitung.de/news/in-neuem-buch-plagiatsvorwuerfe-gegen-annalena-baerbock-li.168185 (visited on 09/13/2021).

[4] dpa. "Abgeschrieben bei Wikipedia: Plagiatsvorwürfe gegen Armin Laschet erhärten sich". In: *Berliner Zeitung* (Aug. 6, 2021). URL: https://www.berliner-zeitung.de/news/abgeschrieben-bei-wikipedia-plagiatsvorwuerfe-gegen-armin-laschet-erhaerten-sich-li.175493 (visited on 09/13/2021).

[5] *English Wikipedia Dump*. URL: https://dumps.wikimedia.org/enwiki/20210901/ (visited on 09/01/2021).

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[7] Mauro di Pietro. *Natural Language Processing - Text Classification example*. URL: https://github.com/mdipietro09/DataScience_ArtificialIntelligence_Utils/blob/master/natural_language_processing/example_text_classification.ipynb (visited on 09/20/2021).

[8] Mauro di Pietro. "Text Classification with NLP: Tf-Idf vs Word2Vec vs BERT". In: *Towards Datascience* (July 18, 2020). URL: https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794 (visited on 09/20/2021).

[9] *Random Forest Feature Importance*. URL: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (visited on 09/15/2021).

[10] "TF–IDF". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 986–987. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_832. URL: https://doi.org/10.1007/978-0-387-30164-8_832.

[11] *TF-IDF*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (visited on 09/15/2021).

[12] *The top 500 sites on the web*. URL: https://www.alexa.com/topsites (visited on 09/26/2021).

[13] Neil Thompson and Douglas Hanley. "Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial". In: *SSRN Electronic Journal* (2017). DOI: 10.2139/ssrn.3039505. URL: https://doi.org/10.2139/ssrn.3039505.

[14] Wikipedia. "Five pillars". In: *Wikipedia* (Sept. 26, 2021). URL: https://en.wikipedia.org/wiki/Wikipedia:Five_pillars (visited on 09/26/2021).

[15] Wikipedia. "Good article criteria". In: *Wikipedia* (Aug. 16, 2021). URL: https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria (visited on 09/26/2021).

[16] Wikipedia. “Good articles”. In: *Wikipedia* (May 16, 2020). URL: https://en.wikipedia.org/wiki/Wikipedia:Good_articles (visited on 09/26/2021).