University of Ljubljana
Faculty of Computer and Information Science

Andrejaana Andova

# Assessment of text readability using statistical and machine learning approaches

DIPLOMA THESIS

UNDERGRADUATE UNIVERSITY STUDY PROGRAM
COMPUTER AND INFORMATION SCIENCE

Mentor: izr. prof. dr. Marko Robnik-Šikonja

Ljubljana, 2017

*This text is formatted with the editor LaTeX*

Faculty of computer and information science issues the following thesis:

Text readability is frequently assessed using statistical measures taking into account length and complexity of sentences, the difficulty of the vocabulary, and the coherence of the text. Machine learning methods can also be used for this task but haven't been tested for Slovene, yet.
Construct two prototype readability scores based on machine learning models and train them on the Šolar corpus containing essays of primary and secondary school students. As the readability score use i) the probability returned by classification model of a text being produced by an older student, and ii) the regression value measuring years of education required to write a given text. Statistically evaluate the produced measures using the Šolar corpus and texts extracted from the ccGigafida corpus.

*I would like to express my sincere gratitude and appreciation to my mentor, prof. Marko Robnik-Šikonja for his guidance, immense help and quick responses.*
*I am grateful to my parents and my brother without whose love, affection and encouragement I would have never made it so far.*
*Special thanks to my dearest friends in Ljubljana who have become my extended family for the last four years and have helped me in many situations.*
*Thank you!*

Dedicated to my parents.

# Contents

# List of acronyms used

| **Acroniym** | **Definition** |
|---|---|
| **SVM** | Support Vector Vachine |
| **ML** | Machine Learning |
| **NB** | Naive Bayes |
| **MNB** | Multinomial Naive Bayes |
| **BNB** | Bernoulli Naive Bayes |
| **SGD** | Stochastic Gradient Descent |
| **NLP** | Natural Language Processing |
| **NN** | Neural Networks |
| **LR** | Linear Regression |
| **BOW** | Bag Of Words |
| ... | ... |

# Abstract

**Title:** Assessment of text readability using statistical and machine learning approaches

**Author:** Andrejaana Andova

This thesis describes a prototype of a system that evaluates the readability of a given text in Slovene. To estimate the readability of a text, we used two methods - regression and classification. The regression method returns a numerical estimation of the readability of a text expressed as years of education, while the classification method tries to classify the input into two classes, where one of the classes is defined as more readable and the other as less readable. We used the corpus Šolar as a training set and first estimated readability using statistical measures. Using features extracted from the texts, we trained different ML algorithms. To assess the quality of our prototypes, we used newspapers and magazines from ccGigafida corpus as a testing set.

**Keywords:** readability, natural language processing, machine learning.

# Povzetek

**Naslov:** Ugotavljanje berljivosti besedil z uporabo statističnih mer in strojnega učenja

**Avtor:** Andrejaana Andova

Ta diplomska naloga opiše prototip sistema, ki oceni berljivost danega besedila v slovenščini. Za oceno berljivosti besedila smo uporabili dve različni metodi – regresijo in klasifikacijo. Regresijska metoda kot oceno berljivosti besedila vrne število ki ustreza število let študija, medtem ko poskusi klasifikacijska metoda besedilo razvrstiti v enega od dveh razredov, kjer je en razred definiran kot bolj berljiv, drugi pa kot manj berljiv. Kot učno množico smo uporabili korpus esejev Šolar. Berljivost smo ocenili z različnimi statističnimi merami in s pomočjo algoritmov strojnega učenja. Kakovost naših prototipov, smo ocenili tudi s pomočjo časopisov in revij iz korpusa ccGigafida.

**Ključne besede:** beljivost, obdelava naravnega jezika, strojno učenje.

# Razširjeni povzetek

Sposobnost branja in razumevanja besedil je pomembna veščina. Izkaže se, da lahko bralec iz besedila razbere največ informacij, če je besedilo primerno bralčevi sposobnosti branja. Za angleški jezik so razvili že več mer berljivosti besedila. Učitelji uporabljajo mere berljivosti, da bi določili primerna besedila za učence. Različni pisci uporabljajo mere berljivosti, da bi svoja besedila čimbolj približali ostalim.

Ker za slovenski jezik mer berljivosti, ki temeljijo na strojnem učenju še nimamo, smo v diplomski nalogi poskusili narediti prototip, ki ocenjuje berljivost slovenskih besedil. Pri tem smo uporabljali dve metodi - regresijsko in klasifikacijsko. Regresijska metoda vrne število, ki ovrednoti berljivost besedila, ki ustreza številu let izobraževanja. Klasifikacijska metoda uporablja dva razreda - eden predstavlja manj berljiva besedila, drugi pa bolj berljiva besedila. Klasifikator pove, s kolikšno verjetnostjo lahko besedilo razvrstimo med bolj berljiva.

Kot učno množico smo uporabili zbirko esejev iz javno dostopnega korpusa Šolar, ki vsebuje eseje učencev osnovne in srednje šole. Pri klasifikaciji smo eseje učencev osnovne šole ocenili kot manj zahtevne, eseje iz srednje šole pa kot bolj zahtevna besedila. Pri regresiji smo kot oceno berljivosti vsakemu besedilu dodelili število let formalnega izobraževanja učenca.

Iz besedil smo izluščili različne statistične lastnosti. Opazili smo, da nekateri eseji vsebujejo samo en stavek. Da bi imeli boljše predstavnike besedil, smo eseje z manj kot petnajstimi stavki izbrisali. Pri ocenjevanju berljivosti besedila so se za najbolj pomembne izkazale mere kot so Dale-Chall formula, avtomatiziran indeks berljivosti, enostavna mera Gobbledygook, Flesch–Kincaid raven berljivosti ipd. Prav tako smo iz besedila izluščili različne statistike na podlagi števila pojavitev besed v besedilu. Ocenili smo tudi pomembnost besed v učni množici, uteženih s tf-idf utežmi. Veliko besed, uporabljenih v šolskih esejih je nakazovalo na kakšno določeno knjigo. Primer tega so vsa imena, mesta, števila kot tudi nekatere besede, kot so vitez, don, kralj ipd. Da bi boljše ocenili berljivost besedila, smo iz učne množice izbrisali vse besede, ki nakazujejo na neko knjigo. Poleg pojavitev posameznih besed smo naredili statistiko tudi za zaporedja dveh ali treh besed. Ker se je pri tem povečalo število značilk, smo izbrisali vse značilke, ki so se v učni množici pojavile le enkrat. S tem se je število značilk dvakrat zmanjšalo.

Korpus Šolar smo razdelili na testno in učno množico, pri čemer je učna množica obsegala 40% celotnega korpusa. Z uporabo naštetih značilk smo razvili več modelov strojnega učenja. Kot najbolj učinkovita pri klasifikacijskih problemih se je izkazala metoda podpornih vektorjev, ki je razvrstila besedilo v pravilen razred v 96% primerov. Pri regresiji je dala najboljši rezultat linearna regresija, katere povprečna absolutna napaka je bila le 0,57.

Iz ccGigafide smo ocenjevali različne časopise, revije, stripe ipd. Pri klasifikaciji z uporabo statističnih lastnosti so bila besedila iz Cicibana in Alana Forda ocenjena kot manj zahtevna, besedila iz Dela, Mladine ter Dnevnika pa kot bolj zahtevna. Klasifikator je razvrstil 78% besedil iz interneta in 88% besedil iz avtomobilskih revij med zahtevnejša besedila.

Pri ccGigafida smo z regresijo ocenili število let formalnega izobraževanja. Z uporabo statističnih lastnosti je linearna regresija pri Cicibanu ocenila 13 let formalnega izobraževanja, Mladina je bila ocenjena s štiridesetimi leti formalnega izobraževanja, Delo pa s petintridesetimi. Ker je imela naša učna množica majhen razpon (od šest do trinajst let izobraževanja), je regresijski prototip slabo ocenil berljivost besedil korpusa ccGigafida.

Pri analizi rezultatov se je izkazalo, da dobimo ob uporabi izključno statističnih lastnosti besedila boljše rezultate za ccGigafido, medtem ko dobimo najboljše rezultate za zbirko Šolar, če uporabljamo samo vektorje utežene s tf-idf.

# Chapter 1

# Introduction

The ability to read and understand texts is an important skill. A lot of research has been done to measure the readability of texts. In 1920s university-based psychologists established that in order to improve the reading skills of an individual and for better understanding of the text, we need reading material that closely matches the reader's ability [6].

Teachers use different readability measures to decide which text should their students read, to match their reading ability. Furthermore, lawyers, doctors, marketers, writers etc. use different tools for text analysis in order to check the readability of their texts. Their goal is to measure how easily can the general public or the colleagues understand their reports and to get an objective evaluation of their writings.

Although the English language has several tools that measure the readability of a text, there is none for technically less developed languages such as Slovene or Macedonian. In this thesis we developed a prototype of a text readability tool for Slovene.

Measuring the readability of a text is not an easy task. Depending on the reader's background knowledge and the subject of the text, the readability level excessively differs. There are different measures to calculate the readability of a given text. Most of the measures are merely statistical calculations of average sentence length, average word length and so on. In 1948

Dale and Chall came up with a formula for estimating the readability of a text. Afterwards a lot of similar formulas have been introduced.

The goal of the thesis is to automate calculation of readability for Slovene texts using natural language processing (NLP). To do this, we first extract numerical features from the Šolar corpus which consists of essays written by primary and secondary school students and forms a good base to form a readability score. To construct this score, we use classification and regression models. The classsification based method divides the data into less and more difficult texts and trains a classifier to differentiate between them. When a new text is given, the approach estimates the probability of text belonging to the more difficult class. Thus, the user gets an estimate whether his text is more or less readable. The regression method return a numeric estimation of the readability of a text.

As our final test, we estimate the readability on different publications from the ccGigafida corpus.

The content of the thesis is outlined below. To get a general overview of the problem, we introduce statistical readability measures in Chapter 2. We give a brief definition of each measure used. In Chapter 3 we explain machine learning methods like SVM, naive Bayes and neural networks. Chapter 4 analyses the data sets used and gives some statistics for them. In Chapter 5 we explain the methodology to get readability scores. In Chapter 6 we classify different texts from Šolar and ccGigafida corpora. In Chapter 7 we present conclusions and ideas for further work.

# Chapter 2

# Readability measures

Several readability measures have been composed in order to calculate the text difficulty. They are mostly adjusted to the English language. However, since they are based on statistics of the text, we use them for the Slovene language.

- **Gunning fog index** [18] estimates the years of formal education a person needs to understand the text on the first reading. It's result can be calculated by the following formula:
  $$0.4 * (\frac{\text{words}}{\text{sentences}} + 100 * \frac{\text{complex words}}{\text{words}})$$
  where the complex words are those composed of 7 or more syllables.

- **Flesh reading ease** [17] high scores indicate that the material is easier to read and lower scores indicate more difficult texts. The formula for the Flesch reading ease score is:
  $$206.835 - 1.015 * \frac{\text{words}}{\text{sentences}} - 84.6 * \frac{\text{syllables}}{\text{words}}.$$

- **Flesch-Kincaid grade level** [17] assesses the number of years required to understand a given text. The formula for the Flesch-Kincaid grade level is:
  $$0.39 * \frac{\text{words}}{\text{sentences}} + 11.8 * \frac{\text{syllables}}{\text{words}} - 15.59.$$

- The **Dale-Chall readability formula** [16] uses a list of 3000 words that groups of 4th grade American students could reliably understand. Words that are not on that list are considering to be difficult. Since the Slovene language does not have a list of 3000 words a person with 4th grade of education would understand, we use the most frequently used 3000 words extracted from 4 corpora: Kres, Jazen, Gos and Šolar [1].
$$0.1579 * \frac{\text{difficult words}}{\text{words}} * 100 + 0.0496 * \frac{\text{words}}{\text{sentences}}.$$
- The **automated readability index** [15] is calculated as:
$$4.71 * \frac{\text{characters}}{\text{words}} + 0.5 * \frac{\text{words}}{\text{sentences}} - 21.43.$$
- **Simple Measure of Gobbledygook(SMOG)** [22] estimates the years of education needed to understand a piece of writing. It was developed as a more accurate and more easily calculated substitute for the Gunning fog index. In its formula, words of 3 or more syllables are referred to as polysyllables.
$$1.0430 * \sqrt{\text{number of polysyllables} * \frac{30}{\text{number of sentences}}} + 3.1291.$$
- **OVIX(word variation index)** [13] is a readability formula designed for the Swedish language. It is calculated as:
$$\frac{\text{number of words}}{log(2 - \frac{\text{number of unique words}}{\text{number of words}})}.$$

# Chapter 3

# Machine learning models

In order to analyse the readability of a given text, we used different machine learning models. We briefly explain the models that performed best on our data sets. In Chapter 6, we present the results.

## 3.1 Support vector machines

A data set samples are presented as points in n-dimensional space, where each feature presents it's own dimension. Therefore, if we have 10 features, each sample from the data set is presented in a 10-dimensional space.
SVM designs a hyper-line that best separates the classes. The basic SVM classifies using linear functions, but using the kernel trick [19], a non-linear classification is obtained.
The SVM time complexity spans from $O(n_{features} \cdot n^2_{samples})$ to $O(n_{features} \cdot n^3_{samples})$ [9]. For large data sets, the SVM method consumes a lot of time to construct the model. It is sensible to noise in the data set and if the number of features is much larger than the number of samples, overfitting might occur. Despite the shortcomings, SVM is one of the most popular and effective machine learning algorithms.

## 3.2 Naive Bayes

Naive Bayes is a classifier based on the Bayes theorem, but assumes that all the attributes are independent [2].

In NLP, attributes often present frequencies of words in the documents. Words that occured in the testing set, might not be present in the training set and we use a smoothing factor that prevents assigning zero probability to features that have not been used in the training set.

Naive Bayes is fast as it uses only basic operations to calculate prior and class conditional probabilities and is appropriate for large data sets. Given that frequently a word depends on other words, the naive Bayes assumption of independence between the attributes is not fulfilled for texts. Despite this, the naive Bayes classifiers shows good results.

## 3.3 Neural networks

Basic building blocks of neural networks are neurons. Divided into layers, neurons between two adjacent layers are connected with each other. They take data from the previous layer and use weights on connections to compute a result [3]. Between the input and output layer, the neural network contains hidden-layers. The hidden-layers contribute to the non-linearity of the algorithm, possibly providing solutions to nonlinear problems.

## 3.4 Stochastic Gradient Descent

Many ML algorithms have convex loss functions. This characteristic is used by the stochastic gradient descent [10] algorithm, which searches through their parameter space to find good parameter values for linear model. SGD is an optimization technique and it doesn't provide optimal parameter values. However, it retrieves good enough parameters to constructs a model. Scikit-learn provides many loss functions based on SVM, logistic regression etc.

## 3.5 Linear regression

Linear regression builds a model by solving:

$$\alpha X + \beta = y$$

where X presents the features' values in the training set and y presents the outcome variable that we want to predict. $\alpha$ and $\beta$ are the parameters of the model.

## 3.6 Combining the models

Another useful method for classifying data is to combine the results from various classifiers. In this section, we explain different approaches to combine predictors.

- The **majority class voting** system selects the class that most of the classifiers predicted. Assuming that SVM predicted *class1*, naive Bayes predicted *class2* and SGD predicted *class1*, *class1* would get 2 votes and *class2* would get only 1 vote. The result of the majority voting would be *class1.*

- The **weighted voting** takes predictions from different classifiers and weights their votes according to some pre-computed weights. Empirically, we calculated the following weights for each classifier.

| SVM | Naive Bayes | Neural Network | SGD |
|---|---|---|---|
| 0.3 | 0.1 | 0.4 | 0.2 |

So, if SVM predicted *class1*, naive bayes predicted *class2*, neural networks predicted *class2* and SGD predicted *class1*, the result would be:

$$class2 \cdot 0.3 + class1 \cdot 0.1 + class2 \cdot 0.4 + class1 \cdot 0.2 = class1 \cdot 0.3 + class2 \cdot 0.7$$

The model will predict *class2*

# Chapter 4

# Corpora

As a basis for training the readability scores we use Šolar [11] corpus. The corpus contains essays written by high school and elementary school students. To measure the readability of the essays, we assumed that older students produce more difficult texts. We introduced 2 classes - students from elementary school and students from high school. The distribution between the two classes was not proportional. In order to adjust the class distribution, we used the undersampling technique [21] which randomly removes data from the larger class until the proportion between the two classes is equal.

| Before preprocessing | | After preprocessing | |
|---|---|---|---|
| elementary school | high school | elementary school | high school |
| 505 | 2.198 | 373 | 1.915 |

Table 4.1: Distribution of the classes in Šolar. The first part presents the number of texts before we removed the outliers, the second part presents the data after the outlier removal

With the regression method, we predicted the years of education from 1-13. Thus, the essays from the students attending first year elementary school would get a readability score 1, while the students from first year high school would get 10. The distribution of the readability estimations is shown in Table 4.2.

| Readability estimation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of students | 0 | 2 | 0 | 0 | 0 | 26 | 125 | 177 | 175 | 671 | 431 | 521 | 529 |

Table 4.2: Distribution of the readability scores, based on the number of years of education

To get more realistic assessment of readability, we used the ccGigafida data set [7]. The Gigafida corpus is a collection of various types of texts from newspaper articles, magazines, children's books and online texts. The total number of documents is $31,722$, but we used only $2,261$ documents, mainly Slovenian online newspapers like Delo, Mladina etc., but also some car magazines, comic books and children's magazines. We can see the distribution of the texts we analysed in Table 4.3

| Ciciban | Mladina | Delo | Alan Ford | Internet | Auto | Celjan | Dnevnik |
|---|---|---|---|---|---|---|---|
| 17 | 85 | 1255 | 8 | 529 | 19 | 327 | 29 |

Table 4.3: Distribution of used sources from ccGigafida

Both corpora are stored in the XML format. Each entry contains lemma and msd value. The lemma is the base form of the word [20], and the msd value [4] contains morphosyntactis description of the word.
The XML structure in the two corpora is different, therefore we parsed the XML documents and saved the texts as separate files. Each line in our files presents a word or punctuation mark.

# Chapter 5

# Methodology

In this Chapter, we will describe the whole process of building the readability estimators. We extract statistical features from the texts and delete the outlier texts, who have too few sentences to gain some knowledge from them. Afterwards we build the term-document matrices from which we delete words that indicate certain texts. At last, using these features we build 2 different readability estimators - one that uses regression and one that uses classification in his approach.

## 5.1 Statistical features

Using the texts from Šolar, we computed different statistical features from the text. We describe the ones that empirically have shown the best results:

- the **average number of syllables per word**,
- the **percentage of words containing 7 or more syllables**,
- the **percentage of words containing 4 or less syllables**,
- the **percentage of sentences containing 29 or more words**,
- the **percentage of sentences containing 6 or less words**,
- the **average number of words per sentence**,

- the **average word length**,
- **type token ratio** (TTR) refers to the proportion of tokens (individual words) and type (the amount of unique words). A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite.
- **The Guiraud's index** [12] of lexical richness is an alternative to the type-token ratio. It is calculated as:

$$\frac{\text{number of unique words}}{\text{total number of words}}.$$

Besides the statistical features we also used the readability measures described in Chapter 2.

## 5.2 Data preparation

During the text analysis of the Šolar data set, we noticed that some of the essays were only few sentences long. While the average number of sentences was 25.6, some of the essays contained merely 1 sentence. Therefore we set a threshold - the text had to be at least 10 sentences long in order to be analysed. With the new criterion we deleted 418 essays. The average number of sentences increased to 29 sentences. Along with the sentence length, the average word length and lexical diversity also increased which indicates that now more complex words are being used. The new statistics of the class distribution is presented in Table 4.1.

## 5.3 Document-term matrix

The statistical features describe the overall text. However, sometimes we want to analyse the text using the words in the documents [14]. We built a matrix whose rows present the documents, while the columns present the lemma values of the words used in the training set. Using this matrix, we analyse whether a word is present in a document or not. We call this matrix binary tf. A modification of this matrix uses frequency of words in a document. We also calculate how important a term is in a given corpus and construct a tf-idf matrix [23]. Each document is presented as a vector, called bag-of-words(BOW) vector.

## 5.4 Word elimination

Since the essays are written by high school and elementary school students, a lot of words used in their texts depend on the books they were analysing. For example, the students from first year high school had to read Sofokles' Antigona. In their essays the name Antigona is mentioned along with words like Kreon, Ismena, king and so on. We eliminate the words specific to a certain book.

Since the msd value of a word is presenting all the grammatical features of the word like sex, tense, case etc. we use it to eliminate the numbers, proper names and their possessiveness. We analysed the terms that tf matrix returned as most frequent and manually, deleted words indicating some book from the most frequent 300 terms. Along with the numbers, proper names and their possesiveness, the following words have been deleted: vitez, don, kihot, baron, povoden, kralj, viteški, viteštvo, tragedija, komedija, starešina, roman, pesem, morski, boginja, drama, grški, princ, morje, vran, epski, puberteta, tuljenje, lirika, kuga, imperij, plod, jabolko, biblija, tujec, grof, grofica, gral, venec, mit, mitološki.

## 5.5 n-grams

Along with single words, we can also count the frequencies of two or three sonsecutive words, called n-grams [8] .The frequencies of single words are referred to as unigrams. The bigrams present the frequencies of two words occurring together in a sequence, while trigrams present the occurrence of three words in a sequence. For the following sentence: "The weather is sunny outside." we get the following n-grams:

| | |
|---|---|
| Unigrams | "The", "weather", "is", "sunny", "outside" |
| Bigrams | "The weather", "weather is", "is sunny", "sunny outside" |
| Trigrams | "The weather is", "weather is sunny", "is sunny outside" |

We add the bigrams and the trigrams as features in the document-term matrices, but since the feature space increased immensely, we decided to delete the terms in the bigrams and trigrams that occure only once.

## 5.6 Predicting

Using the text features previously described, we build 2 different readability estimators. The first one is using classification to estimate the readability of a text. As input, it accepts the statistical features and the term-document matrices previously described and classifies the given text. As a result it returns the probability distribution to both of the classes. For example, one text might belong to the difficult texts with probability 64% and to the less difficult with the rest 36%. We define the classes using the texts from the training sets. The high school essays represent the more difficult class, while the elementary school essays represent the less difficult.

The second readability estimator is using regression methods to estimate the readability. We extract the statistical features and the term-document matrices from the texts and try to give a numerical estimation of the readability of the text. This method returns a single numerical value expressing the readability of the text.

# Chapter 6

# Evaluation of results

We present two different results. In the first part, we train and test our readability score on the Šolar corpus. In the second part we test the produced scores on a corpus of texts extracted from ccGigafida.

## 6.1 Šolar

When using the classification method, we divided the data into training set and testing set, where the testing set contains 40% of the documents. We separated our data to high school and elementary school students. Due to imbalance between high school and elementary school students we used the undersampling technique, which randomly drops examples from the training set until the proportion between the classes is equal. All results were generated by averaging the scores on 30 different training-testing set splits.

We empirically discovered that the linear kernel in SVM gives better results than the RBF kernel. Using BOW vectors, the number of documents is much lower than the number of features, hence, the RBF kernel overfits the model. SVM returned the best when we assigned the parameter C to 100.

In the neural networks, the number of hidden layers strongly affected the results. When using 10 neurons in a hidden-layer on average only 52% of the testing examples were classified correctly, while using 100 neurons, the

| | SVM | NB | NN | SGD | MV | WV |
|---|---|---|---|---|---|---|
| Statistical readability measures | 0.78 | 0.71 | 0.64 | 0.67 | 0.72 | 0.72 |
| All features | 0.79 | 0.71 | 0.71 | 0.60 | 0.7 | 0.73 |

Table 6.1: CA of the model when using only the statistical features for Šolar

| | SVM | NB | NN | SGD | MV | WV |
|---|---|---|---|---|---|---|
| Binary tf | 0.91 | 0.93 | 0.93 | 0.90 | 0.92 | 0.92 |
| tf | 0.91 | 0.94 | 0.92 | 0.92 | 0.92 | 0.92 |
| tf-idf | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 |

Table 6.2: CA of the model when using BOW vectors for Šolar

accuracy increased to 64%. Using 200 neurons, the result didn't improve, but the training time strongly increased. Therefore, we used 100 neurons in the hidden layer in our neural network.

SGD gave the best results using the standard loss function introduced by the SVM.

At first, we experimented with the statistical features. The readability measures in Table 6.1 are computed with features consisting of the readability measures presented in Chapter 2. "All features" row presents features from Chapter 5.1. The results using all the features increased merely 1% compared to classical readability measures. Despite the fact that the readability is not increasing much, we used the features described in Chapter 5.1 because we wanted to gain as much information about the data as possible.

Using the BOW vectors for Šolar, we obtained the best classification accuracies (Table 6.2). Since tf-idf weigihting reflects the importance of words in the whole data set, it is not surprising that it gives the best results from all the BOW vectors.

Besides single terms, we tested the bigrams and trigrams as features. Since the feature space would contain, 322,727 trigrams on average, we removed

| | SVM | NB | NN | SGD | MV | WV |
|---|---|---|---|---|---|---|
| Unigram | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 |
| Bigram | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| Trigram | 0.94 | 0.95 | 0.95 | 0.94 | 0.95 | 0.94 |

Table 6.3: CA of the model when using the unigram, bigram and trigram weighted with tf-idf for Šolar

| SVM | MNB | BNB | NN | SGD |
|---|---|---|---|---|
| 0.85 | 0.72 | 0.85 | 0.71 | 0.47 |

Table 6.4: CA of classification model using statistical features and bigram features weighted with tf-idf for Šolar

the terms occuring only once and gained 28,000 features on average for the trigrams. The results in Table 6.3 show that bigrams perform best. By combining the bigrams with statistical features the accuracy decreased (Table 6.4).

In the regression method, we also divided the data into training set and testing set. To evaluate the results we used mean absolute error computed as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i'|}{n}$$

where $y_i$ is the true value and $y_i'$ is the predicted value. As another metric to estimate the regression error, er used mean squared error:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n}$$

Alternative to the MSE is the root mean square error computed as:

$$MSE = \frac{\sum_{i=1}^{n} \sqrt{(y_i - y_i')^2}}{n}$$

As regression models, we used linear regression, SVM and neural networks.

As shown in Table 6.5, the bigram and the trigram features weighted with tf-idf performed best. However since trigrams are slightly better, we

| | SVM | | | LR | | | NN | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMAE | MAE | MSE | RMAE | MAE | MSE | RMAE |
| Statistical measures | 0.70 | 3.90 | 2 | 1.15 | 2.00 | 1.40 | 7.46 | 114.50 | 8.50 |
| Unigram tf-idf | 0.60 | 0.65 | 0.80 | 0.57 | 0.64 | 0.79 | 1.20 | 2.40 | 1.50 |
| Unigram binary tf-idf | 1.70 | 4.00 | 2.00 | 0.80 | 1.00 | 1.00 | 2.33 | 8.80 | 2.90 |
| Unigram tf | 1.70 | 4.00 | 2.00 | 0.80 | 1.00 | 1.00 | 2.33 | 8.60 | 2.90 |
| Bigram tf-idf | 0.58 | 0.65 | 0.80 | 0.57 | 0.62 | 0.79 | 1.30 | 2.60 | 1.60 |
| Trigram tf-idf | 0.58 | 0.64 | 0.80 | 0.57 | 0.60 | 0.78 | 1.30 | 2.70 | 1.65 |

Table 6.5: Different error estimations for regression in Šolar

| | SVM | Linear Regression | NN |
|---|---|---|---|
| MAE | 1.68 | 0.57 | 0.7 |
| MSE | 3.8 | 0.6 | 0.88 |
| RMSE tf | 1.98 | 0.78 | 0.9 |

Table 6.6: Results of regression model using statistical features and trigram features weighted with tf-idf for Šolar

used the trigrams with tf-idf to build the final result. The final result is calculated by combining the trigram features weighted with tf-idf together with the statistical features (Table 6.6).

## 6.2 ccGigafida

The Gigafida corpus consists of many different texts, from newspaper articles to children's books. We analyse the results using different models trained on the Šolar corpus. Since we don't know the actual readability score of the texts, we present the examples classified into the high school class.
In the regression approach, we present the average readability value that our algorithm returned in each category.

- **Ciciban**
  Ciciban is a magazine written for children older than the age of 6. Mainly elementary school children read this magazine, and therefore, we expect the texts from this magazine to be similarly readable as the elementary school essays.

  The results above show excellent classification using the statistical features. However, we have to note that we used merely 17 examples of Ciciban magazines. The CA might decrease if the testing set was larger.

- **Mladina**

Mladina is a left-wing current affairs magazine. It analyses political problems, and claims to represent the voice that fights the powerful. Many famous philosophers and politicians presented their thoughts in Mladina including Slavoj Žižek and Janez Janša.

- **Delo**
  Delo is a Slovenian daily newspaper. During its 50 years of existence, it covered different topics from politics, economics, sports, culture etc.

- **Alan Ford**
  Alan Ford is an Italian comic book first published in 1969. It gained popularity in Yugoslavia and you can still buy latter editions in Slovenia. Although its basic text structure is simple and most of the time presented in dialogs, it covers progressive topics like racism, capitalism etc. We only have 8 examples of the Alan Ford comic books.

- **Internet**
  We tested texts extracted from the Internet. Because everyone can post on Internet, it is not surprising that SVM classified only 78% of the data into the high school class.

- **Car magazine**
  Car magazines were used as an example of texts with relatively low lexical diversity. It is not surprising that approximately 12% of these texts were classified as elementary school. We only have 19 examples of Car magazine.

- **Celjan**
  Celjan is a local magazine intended for people in the vicinity of Celje.

- **Dnevnik**
  Dnevnik is a daily newspaper published in Ljubljana.

The results show that unlike Šolar corpus, the statistical measures show better results than BOW vectors. A possible explanation would be that the elementary school students start reading books on more serious topics. Therefore, in their essays, they start using more complex words, which confuse the BOW representation. However, their writing skills still haven't improved enough and therefore,they keep the same text structure as before. Thus, the statistical measures still provide good results. This problem could be solved if the students covered more topics, where their actual vocabulary would be expressed.
Another solution would be to gather larger data set because existing 373 essays written by elementary school students, may not be efficient representation of these children's abilities.

Both methods gave poor results with the BOW representation. Using the statistical measures, the classification mehod showed good results by classifying Ciciban as less difficult and texts from Mladina and Delo as difficult. The regression method using LR estimated the average number of years of education of Ciciban readers to be 13, while for Mladina the estimate is 40. Since only 2 essays in Šolar were written by students in their earlier years of education (see Table 4.2), we don't have a decent representation of these readability scores. Also, the journalists writing articles in Mladina and Delo have much better writing abilities than the high school students. Because we don't have any texts representing highly-educated writers in our training set, the algorithm assigned high difficulty to the journalists' texts. Linear regression sorted the sources in the following order: Ciciban, Celjan, Alan Ford, Internet, Dnevnik, Car magazines, Delo, Mladina. Neural networks returned the following order: Ciciban, Alan Ford, Celjan, Delo, Internet, Dnevnik, Mladina, Car magazines.

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 0% | 0.5% | 9% | 25% |
| bigram tf-idf | 50% | 0% | 0% | 71% |

Table 6.7: CA for Ciciban text being on high school level

| | SVM | Linear Regression | NN |
|---|---|---|---|
| Statistical measures | 10.56 | 13.68 | 8.06 |
| Unigram tf-ids | 10 | 10.45 | 2.27 |
| Statistical measures + tf-ids | 10.55 | 9.94 | 9.35 |

Table 6.8: Average number of years of education estimated by the regression method for Ciciban

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 100% | 100% | 99.6% | 75% |
| bigram tf-idf | 67% | 0% | 0% | 90% |

Table 6.9: CA for Mladina text being on high school level

| | SVM | Linear Regression | NN |
|---|---|---|---|
| Statistical measures | 10.57 | 40.9 | 46 |
| Unigram tf-ids | 10 | 10.43 | 2.46 |
| Statistical measures + tf-ids | 10.56 | 18 | 26.25 |

Table 6.10: Average number of years of education estimated by the regression method for Mladina

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 94% | 98% | 90% | 74% |
| bigram tf-idf | 68% | 7% | 19% | 75% |

Table 6.11: CA for Delo text being on high school level

| | SVM | LR | NN |
|---|---|---|---|
| Statistical measures | 10.55 | 34.35 | 29.3 |
| Unigram tf-ids | 10 | 10.44 | 2.2 |
| Statistical measures + tf-ids | 10.55 | 16.1 | 21.1 |

Table 6.12: Average number of years of education estimated by the regression method for Delo

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 0% | 0% | 0% | 9% |
| bigram tf-idf | 37% | 0% | 0% | 53% |

Table 6.13: CA for Alan Ford text being on high school level

| | SVM | LR | NN |
|---|---|---|---|
| Statistical measures | 10.56 | 15.53 | 9.61 |
| Unigram tf-ids | 10 | 10.4 | 2.28 |
| Statistical measures + tf-ids | 10.56 | 10.36 | 11.36 |

Table 6.14: Average number of years of education estimated by the regression method for Alan Ford

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 78% | 77% | 70% | 69% |
| bigram tf-idf | 63% | 6% | 2% | 75% |

Table 6.15: CA for Internet data text being on high school level

| | SVM | LR | NN |
|---|---|---|---|
| Statistical measures | 10.5 | 41.9 | 43.71 |
| Unigram tf-ids | 10 | 10.4 | 2.21 |
| Statistical measures + tf-ids | 10.56 | 18 | 29.4 |

Table 6.16: Average number of years of education estimated by the regression method for Internet data

| | SVM | Naive Bayes | Neural Networks | SGD |
|---|---|---|---|---|
| Statistical measures | 88.7% | 92% | 81% | 78% |
| bigram tf-idf | 62% | 9% | 2% | 63% |

Table 6.17: CA for car magazine text being on high school level

| | SVM | LR | NN |
|---|---|---|---|
| Statistical measures | 10.56 | 29.7 | 58.66 |
| Unigram tf-ids | 10 | 10.45 | 2.25 |
| Statistical measures + tf-ids | 10.56 | 15.12 | 20.2 |

Table 6.18: Average number of years of education estimated by the regression method for Car magazines

| | SVM | Neural Networks | SGD |
|---|---|---|---|
| Statistical measures | 10.56 | 28.8 | 40.5 |
| Unigram tf-ids | 10 | 10.43 | 2.32 |
| Statistical measures + tf-ids | 10.56 | 14.78 | 18.63 |

Table 6.19: Average number of years of education estimated by the regression method for Celjan

| | SVM | LR | NN |
|---|---|---|---|
| Statistical measures | 10.56 | 59.34 | 41.5 |
| Unigram tf-ids | 10 | 10.4 | 2.26 |
| Statistical measures + tf-ids | 10.55 | 22.97 | 37.37 |

Table 6.20: Average number of years of education estimated by the regression method for Dnevnik

# Chapter 7

# Conclusion

In this thesis, we measure the readability of texts using two different ML methods - classification and regression. The classification method returns the probability that a text belongs to the class containing more difficult texts, while the regression method returns a numerical estimation of the readability of a text expressed as years of education. As features for ML algorithms we extract different text statistics like average word length, average sentence length, etc. We use bag of words representation to extract information from the term frequencies. We use readability formulas designed for the English and Swedish language.
As the learning data we used the essays from Šolar corpus, while for testing along with Šolar corpus we also used ccGigafida. When testing on Šolar the BOW representation of the text showed best results, with classification accuracy of 96%. The regression method produced MAE of only 0.58. Testing on ccGigafida showed good results without BOW vectors. Texts from Ciciban and Alan Ford were classified as less difficult, while articles from Mladina and Delo were estimated difficult. The regression method classified Ciciban as less difficult. Linear regression sorted the texts according to their difficulty by the following order: Ciciban, Celjan, Alan Ford, Internet, Dnevnik, Car magazines, Delo, Mladina.

Readers understand texts better when there is some logical connection between the ideas in the text(coherence). Some researches have already estimated the coherence of texts using the LSA method [5] . As further work, it would be interesting to analyse the coherence of Slovene textse.

# Bibliography

[1] Pollak, Senja in Arhar Holdt, Špela (2017). seznam pogostih besed KAUČ (v0.1).

[2] Wikipedia contributors. Naive bayes classifier, 2017. [Online; accessed 20-December-2017].

[3] Howard B Demuth, Mark H Beale, Orlando De Jess, and Martin T Hagan. *Neural network design.* Martin Hagan, 2014.

[4] Simon Erjavec, Tomaž in Krek. Oblikoskladenjske specifikacije in označeni korpusi jos. In *Šeste konference jezikovne tehnologije*, page 49, 2008.

[5] Peter W Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307, 1998.

[6] Edward B Fry. Readability: Reading Hall of Fame Book. Newark, DE: International Reading Assn., 2006.

[7] Špela Arhar Holdt, Iztok Kosem, and Nataša Logar Berginc. Izdelava korpusa Gigafida in njegovega spletnega vmesnika. In *Proceedings of 8th Eighth Language Technologies Conference IS-LTC*, volume 12, 2012.

[8] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson, London, 2014.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Tadeja Rozman, Mojca Stritar Kučuk, Iztok Kosem, Krek Simon, Irena Krapš Vodopivec, Špela Arhar Holdt, and Marko Satbej. Šolar.[ljubljana]: Ministrstvo za izobraževanje, znanost, kulturo in šport, 2012.

[12] Zdislava Šišková. Lexical richness in efl students' narratives. *Language Studies Working Papers*, 4:26–36, 2012.

[13] Christian Smith and Arne Jönsson. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *In Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010*. Citeseer.

[14] Wikipedia. Document-term matrix, 2015. [Online; accessed 8-December-2017].

[15] Wikipedia. Automated readability index, 2017. [Online; accessed 22-November-2017].

[16] Wikipedia. Dale–chall readability formula, 2017. [Online; accessed 22-November-2017].

[17] Wikipedia. Flesch–kincaid readability tests, 2017. [Online; accessed 22-November-2017].

[18] Wikipedia. Gunning fog index, 2017. [Online; accessed 22-November-2017].

[19] Wikipedia. Kernel method, 2017. [Online; accessed 24-November-2017].

[20] Wikipedia. Lemmatisation, 2017. [Online; accessed 22-November-2017].

[21] Wikipedia. Oversampling and undersampling in data analysis, 2017. [Online; accessed 5-December-2017].

[22] Wikipedia. Smog, 2017. [Online; accessed 22-November-2017].

[23] Wikipedia. Tf–idf, 2017. [Online; accessed 23-November-2017].