Impact of text pre-processing on classification accuracy in Polish

Urszula Gumińska[1][0000−0001−5774−3356], Aneta Poniszewska-Marańda[1][0000−0001−7596−0813], and Joanna Ochelska-Mierzejewskainst1[0000−0002−9295−3962]

[1] Institute of Information Technology, Lodz University of Technology, Lodz, Poland
[2] urszula.krzeszewska@dokt.p.lodz.pl, aneta.poniszewska-maranda,joanna.ochelska-mierzejewska@p.lodz.pl

Abstract. Natural language processing (NLP) like other Machine Learning (ML), Deep Learning (DL) and data processing tasks, requires a large amount of data to be effective. Thus, one of the most significant challenges confronting ML/DL tasks, including NLP, is a lack of data. This is especially noticeable in the case of text data for niche languages like Polish. Manual collection and labelling of text data is the primary method for obtaining language-specific data. However, this is a lengthy and labour-intensive process. As a result, researchers use a variety of other solutions, such as machine translation from another, more developed language in the field, to obtain data more quickly and affordably. For these reasons the suitable experiments, described in the paper, were carried out, resulting in a machine translation model that translates texts from English into Polish. Its results were compared with those of a pre-trained model and translations were subjected to human testing. Moreover, the paper presents the influence of different pre-processing stages on the final result of text classification in Polish in terms of one of six emotions: anger, fear, joy, love, sadness, surprise.

Keywords: Natural language processing · Machine translation task · Polish emotion classification · Text pre-processing.

1 Introduction

Natural Language Processing (NLP) as an important sub-field of Artificial Intelligence (AI) is used to analyse, understand and generate language, it studies the interaction between human and computer through natural language – used by humans in everyday communication [1]. Natural language processing, understood as a field of computer science and linguistics, can encompass areas outside of both machine learning and deep learning, as well as take advantage of the strengths of both.

The concept of noise in Natural Language Processing is remain unclear and frequently described as difficult to define. Text pre-processing is regarded as a standard procedure in automatic text analysis. [2, 3] It is also the first step conducted during most of the language specific tasks. However, during this process

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

both redundant and relevant information can be removed. It is possible to distinguish stages such as: removal of special characters, alignment of case, removal of punctuation marks and inverted commas, removal of possessive pronouns, lemmatization, stop words removal. There are three basic stages in automatic text processing:

1. Text preprocessing – the scope of text preprocessing includes case uniformity, removal of special characters that will not be relevant for further analysis (e.g. newline character).
2. Vectorization – the scope of vectorization includes the presentation of previously prepared texts in a form understandable for a computer, placing in specific data structures.
3. The final processing of the text – in our case, it is text classification.

Natural language processing, as well as other Machine Learning (ML), Deep Learning (DL) and data processing tasks, need a large amount of data to accomplish their task effectively [4]. Thus, one of the biggest challenges facing ML/DL tasks, including NLP, is the lack of data. In language processing, this is particularly evident in the case of text data for niche languages, such as Polish.

The primary method for obtaining language-specific data is manual collection and labelling of text data. However, this is a very time-consuming and labour-intensive process [1]. As a result, researchers use a variety of other solutions to obtain data more quickly and less expensively, including machine translation from another, more developed language in the field [5]. Most often, this is English, as the universal language used by researchers around the world and developed in the field of NLP. Also, all the operations that modify the underlying dataset that are part of pre-processing in inflectional languages, as in Polish, can have a significant impact on the effective use of such difficult-to-acquire data.

The work presented in this paper contributes in creation of machine translation model for English-Polish translation and its comparison with existing pre-trained model. What is more, checking the impact of noise reduction methods from text pre-processing task on classification accuracy for texts with six basic emotions specifically in Polish language is described. Moreover, the conducted experiments provide a basis for further research for better automatic text analysis.

The paper is structured as follows: Section 2 presents the related works in area of text pre-processing on classification accuracy. Section 3 describes the process of machine translation tasks for text pre-processing while section 4 deals with the influence of preprocessing on classification in Polish. Both sections contain described task-specific datasets, experiments and analysis of their results.

2 Related works in text pre-processing on classification accuracy

In Natural Language Processing a concept of noise is not well understood and often described as hard to define. The text pre-processing itself is considered a

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

standard process in automatic text analysis. This often includes such procedures as stop-words removal, lemmatization, character size uniformity [2].

Research works often focus on the role of using such methods [3], treating this process as necessary for successful text analysis. Given the huge amount of data available for processing text, especially in the earlier stages of NLP development, a reduction of data was necessary to make it more manageable for existing algorithms. It results in the creation of many tools that allow automatic pre-processing of text data [6]. In addition, experiments were performed on texts domain, such as technical texts using common pre-processing techniques [7], or texts in different languages like Arabic [8] or Chinese [9].

However, few of the published works have dealt with the fact that some of the information removed during text preprocessing can be a significant factor affecting the final interpretation of the text. One such paper is [10], which distinguishes between useful and harmful noise. Another paper focusing on categorizing noise [11] divides it into 7 categories: Orthography, Grammatical Errors, Disfluencies in Human Data, Internet Jargon, URLs, Links and Markup, Repetition of Punctuation and Code-switching. Additionally, it states that in various tasks, different elements of texts can be both useful and harmful.

Most work on NLP focuses on English. It is a fairly easy and structured language and, in addition to being acknowledged as a scientific language, is known and understood by the scientific-research community worldwide. Despite the development of NLP research in national languages other than English, there is still a lack of data for specific tasks in these languages [12]. In response to this problem, many researchers are using automatic methods to translate collections of text data from English into national languages – using machine translation.

Currently, most work on machine translation focuses on neural machine translation approaches using deep machine learning. One of the approaches used for this task is the use of Recurrent Neural Networks (RNN). One RNN model is used as a message encoder, while the other is used as a decoder [13]. This is the most intuitive approach within the sequence to sequence translation. Another approach is the use of Convolutional Naural Networks (CNN). Although this solution was not initially popular in machine translation, thanks to the attentional mechanism [15], it has succeeded in being introduced on a large scale. One more approach is the use of the Transformers architecture [15] – a simpler approach than the RNN usage. The Transformer model's self-attention layers learn the dependencies between words in a sequence by examining links between all of the words in the paired sequences and directly modeling those relationships.

3 Machine Translation model for English-Polish translation

One of the steps needed to test the impact of noise on the classification of texts in Polish was to prepare a dataset of texts in Polish. Due to the lack of a large dataset containing all 6 basic emotions in Polish, it was decided to create such a dataset by applying machine translation on such a dataset in English.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

This section outlines the process of creating and verifying an English-to-Polish machine translation model. It describes in detail the prepared network architecture, the data needed to learn and verify the performance of the model, and discusses the results by comparing them with another pretrained model.

3.1 Chosen model architecture

The prepared model for English-Polish translation is based on the Encoder Decoder architecture [14] with an attention layer [15]. An example of how such a model works is shown in figure 1. It is one of the architectures that fits naturally with the machine translation task.

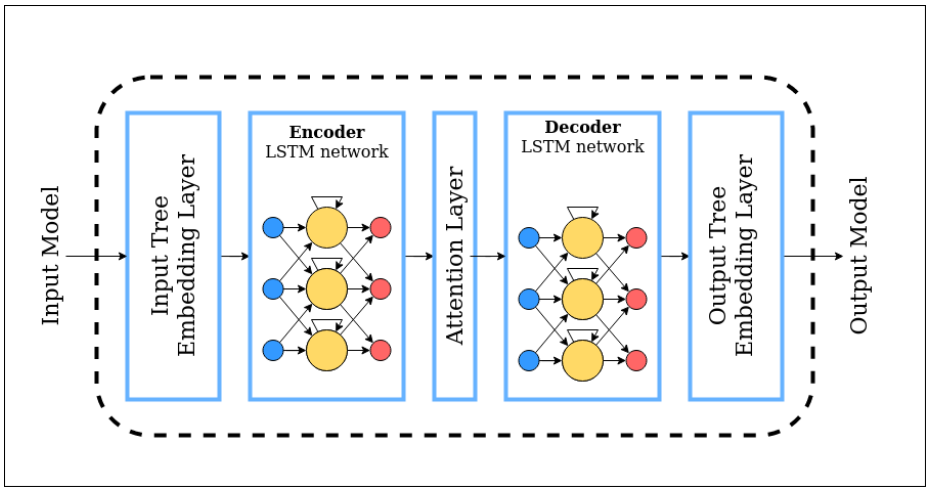

Fig. 1. Encoder Decoder architecture for English-Polish translation.

This model consists of five basic layers:

1. Input embedding layer – responsible for converting input texts into embeddings. At the input of this layer, natural language texts are provided, in the prepared implementation English texts. Its output, on the other hand, receives vectors of texts of a specified length.
2. Encoder layer – its task is to encode the data using the chosen network architecture. For the prepared implementation, the Long-Short Term Memory (LSTM) model architecture was used.
3. Attention layer – uses the attention mechanism which is a technique that is meant to mimic cognitive attention, helps a neural network in memorizing the large sequences of data. This mechanism in machine translation is responsible for both align and translate. Alignment means finding which parts of

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

the input sequence are relevant to each word in the output, whereas translation is the process of using the relevant information to select the appropriate output.

4. Decoder layer – decodes the information using the selected network architecture. For the prepared implementation, the Long-Short Term Memory (LSTM) model architecture was used.
5. Output embbedding layer – responsible for converting the embeddings obtained from the decoder into natural language texts. At the input of this layer, text vectors of a specific length are given. On its output, there are natural language texts, in the prepared implementation, texts in Polish. The result of this layer is the translated texts.

Machine translation task uses LSTM networks [16] as layers in the Encoder and Decoder. The exact appearance of the prepared model along with the parameters is shown in figure 2. This diagram was generated from a created Python application. It reflects the theoretical Encoder-Decoder architecture presented above together with the connections between specific layers.

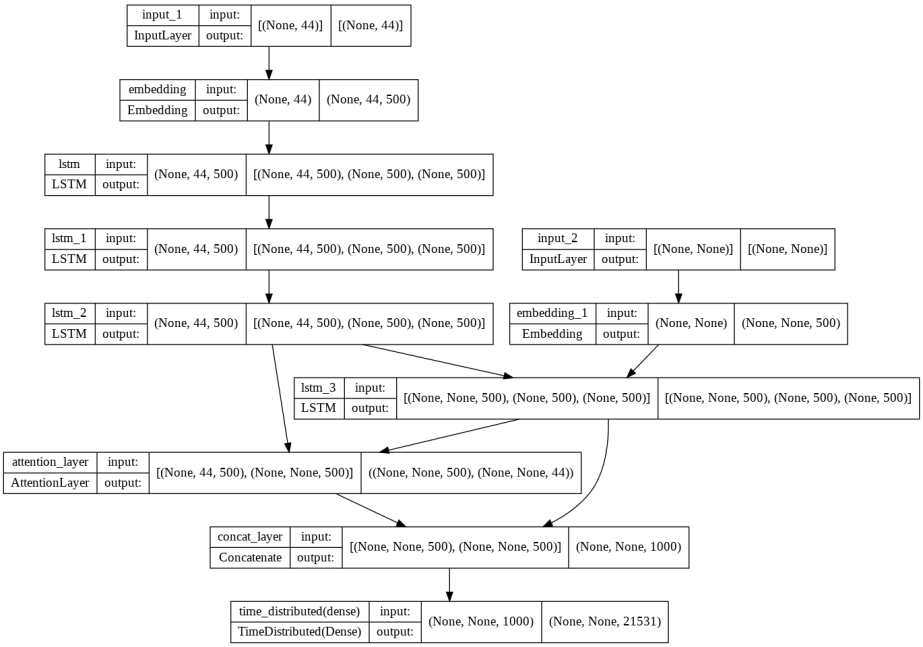

Fig. 2. Prepared model architecture with all parameters.

Attention layer was implemented according to [5]. The Tensorflow library's built-in layers, i.e. LSTM, Input, Dense, Embedding, Concatenate, were used to create the other layers. *Adam optimizer* was applied to the model thus prepared. The model was trained for 25 epochs, with 80 iterations used in each epoch.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

3.2 The dataset used for machine translation task

To conduct the study, it was necessary to select two datasets:

1. The first set serves as a training set – it is on its basis that a DL model was created to allow translations; it is the set that contains texts in both Polish and English.
2. The second set, on the other hand, is the set that we care about translating – it is the collection in English along with the corresponding labels that characterize the emotion assigned to the texts.

Training dataset Is the training dataset, the Tab-delimited Bilingual Sentence Pairs for Polish-English pairs [17] was used. The dataset contains about 46.5k Polish-English pairs in a single text file separated by tabs. Each pair additionally contains copyrights information assigned to it. The sentences in the database are of varying difficulty, ordered from shortest to longest. A sample texts pairs are shown in table 1.

Table 1. Sample Polish-English texts pairs along with the copyright information from used dataset.

English text	Polish text	Copyright info
Wait!	Poczekaj!	CC-BY 2.0 (France) Attribution: tatoeba.org #1744314 (belgavox) & #4476129 (Marek-Mazurkiewicz)
Come quick!	Chodź szybko!	CC-BY 2.0 (France) Attribution: tatoeba.org #274037 (CM) & #354669 (zipangu)
The storm caused a power outage.	Burza spowodowała przerwę w dostawie prądu.	CC-BY 2.0 (France) Attribution: tatoeba.org #1293100 (CK) & #3430132 (konrad509)

The texts prepared in this way underwent additional modification – the word 'start' was added to the Polish text at the beginning of the sentence, and the word 'end' at the end of the sentence. This type of procedure is aimed at easier handling of the texts later on. It is worth noting that at this stage we are only operating on Polish and English texts, omitting copyright.

The dataset was divided into training data and validation data, where 90% of the dataset is the training set and 10% is the validation set. The data was randomly divided between the two sets.

Emotions dataset to translate The Emotion Dataset for Emotion Recognition Tasks [18] was used to create a database under emotion dataset in Polish. It is a dataset containing 2,000 texts in English along with one of 6 emotions [19] assigned to them:

– anger,
– fear,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

– joy,
– love,
– sadness
– and surprise.

A dataset containing tweets exhibiting six different emotions. The dataset comes in the form of a *.csv file containing English text in the first column and a digital designation of one of the 6 emotions in the second column. Each of the emotions in the text is equidistant. The texts are already pre-processed and therefore contain only lowercase letters and no punctuation, allowing the texts to be used directly in ML models. Sample texts along with assigned emotions are shown in table 2.

Table 2. Sample texts along with assigned emotions in tweets dataset.

tweet_id	sentiment	content
1956967666	sadness	Layin n bed with a headache ughhhh...waitin on your call...
1956972270	worry	I ate Something I don't know what it is... Why do I keep Telling things about food
1957088179	happiness	@mrssunshine96 big now!!! Vanessa is going to be 3 in September, its going by so fast! its hard cuz Im workin so much, I miss out on alot

It is this dataset that, once translated, is to serve as a text base to test the impact of noise reduction on classification in Polish. This means that all texts from this dataset were translated using both the pre-trained model and the pre-trained Fairseq model.

3.3 Machine translation task results

The prepared model was initially tested on the validation set. The model was trained for 25 epochs, each of which had 82 iterations. During this time, the values such as loss function and accuracy were measured for both training and validation datasets. As the loss results for the validation set were quite high, especially compared to the test set (Fig. 3). It can be seen that the loss function for the training set decreased its value very rapidly in the first 2 epochs and in the later epochs its decrease was smaller, but it still managed to reach values of around 0.15, while for the test set, the function had a much smaller variation and reached a minimum value of about 0.53.

The obtained results are likely to suggest that the model will not perform particularly well with translations. Therefore it was decided to manually check how the sample translations look like for the prepared model. A few of the translations are shown in table 3. The first column contains three randomly selected sentences in their original version, the second column contains the corresponding Polish translation included in the translation dataset, while the third column contains the translation obtained from the prepared model.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

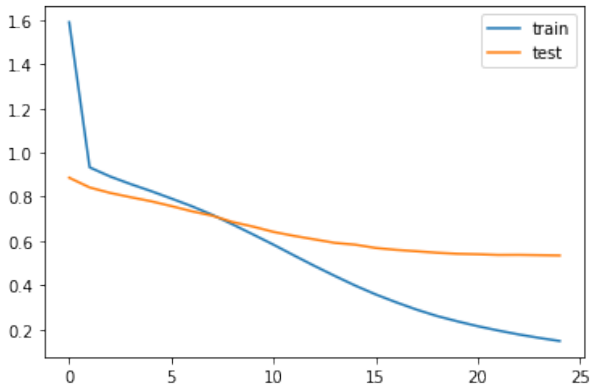

Fig. 3. Loss function over the epochs for train and test dataset.

Table 3. Results of translation in comparison to original text.

Sentence	Original translation	Our translation
tom couldnt stand to be in the same room with mary	tom nie mogl zniesc przebywania razem z mary w jednym pokoju	tom nie potrafil sie przyznac ze mary mine mary
this house belongs to my uncle	to jest dom mojego wujka	dom jest tutaj przez dom
is the hotel far from here	czy hotel jest daleko stad	czy jest tu daleko stad

The results of the manual check were not satisfactory. It can be concluded that the model does not manage longer sentences. At this stage, it was known that it would not be possible to apply the model to create a database of emotionally tinged texts in Polish. Nonetheless, it was decided to use the model to translate the target texts with emotional overtones in order to compare the results to those obtained from the Fairseq model. A summary of the translations along with the original text is shown in table 4.

Table 4. Comparison of emotion texts translation between prepared model and pretrained model.

Sentence	Our model translation	Fairseq model translation
i never make her separate from me because i don t ever want her to feel like i m ashamed with her	lubie nauczyc sie jak robilem jak sie nauczyc sie nauczyc jak sie poprawil jak to sie spodoba sie jak ty i uslyszalem jak sie tam	Nigdy nie odrywam jej od siebie, bo nie chce, zeby czula sie zawstydzona.
i start to feel emotional	spalem jak noc	Zaczynam czuc sie emocjonalnie.
i like to have the same breathless feeling as a reader eager to see what will happen next	nigdy nie mialem okazji pomagac ale musze go znalezc od ciebie ale nie przeczytam wszystkiego	Lubie miec to samo uczucie, gdy czytelnik chce zobaczyc, co bedzie dalej.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

It is worth noting that translations using the pre-trained model are inaccurate and seem to be completely unrelated to the original sentence. Translations from Fairseq's pre-trained model are much more accurate, plus the model added capital letters, punctuation, etc. making the sentences grammatically and inflectionally correct.

3.4 Summary of machine translation task

Although no metric was used at this stage to measure the quality of the translations, issue a clear difference in quality between the self-trained model and the pre-trained one. What is worth noting is that the model prepared as part of the task did not produce translations that would allow the preparation of an emotion dataset in Polish. Possible reasons for this result are:

- too small teaching set for the model to master the relevant language relationships,
- the general difficulty of translating from English into inflectional languages, which is the Polish language (many words do not have their direct equivalents, and there is a variety by persons, cases, etc.),
- choosing too simple DL model to create effective translations.

It is worth noting that the pre-trained English-Polish convolutional model for Fairseq predicts translations at a high level. When analysed manually, the translations are of high quality and appear to be sufficient to create a database of emotionally tinged texts in Polish.

In the future, it would be worthwhile to additionally use an independent metric to check the quality of the translation, or to conduct human tests with an evaluation of the quality of resulting translations.

The lack of data in individual languages is still a very big problem, and regardless of the fact that tools are being developed to effectively map a dataset from English to other languages, none of these tools will be as reliable as data manually collected and evaluated by humans. English translations may become the only viable option for researchers working on niche languages, but at this point they do not capture the dynamics associated with a language, and these models need to be refined all the time.

All of this leads to the next part of the paper responsible for testing the impact of noise reduction in the data, using a pre-trained model to obtain data containing emotions.

4 Text pre-processing impact on text

The second process was to evaluate the impact of noise reduction on the final emotion classification result for Polish texts. To accomplish this, a suitable dataset containing texts in Polish with the six basic emotions assigned was prepared, the necessary steps were designed and test cases were proposed using the further pre-processing steps.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

4.1 Polish sentences dataset used in classification

Due to the very poor translation performance of the prepared model, a dataset translated pre-trained Fairseq model was used to test the effect of pre-processing methods on the effectiveness of multi-class classification for emotions in Polish. This dataset was further expanded by 1000 texts in Polish containing some emotional tinge and information about the formality of the text. All six basic emotions mentioned before were available. The texts were collected as part of a project carried out by students of the Lodz University of Technology, majority in computer science. The data was collected from private correspondence of college mailboxes. They were then anonymized.

In the same way, a value indicating the formality of the text was assigned, where "0" meant informal text, "1" partially formal and "2" very formal. However, for the purposes of this study, we will skip the formality aspect of the texts and focus only on the basic emotions.

From the values assigned to the successive emotions, the highest one was selected and the emotion assigned to this highest value became the text label. A sample texts along with assigned values for specific emotions and its label are shown in table 5.

Table 5. Sample text along with assigned values for specific emotions and its label.

id	text	formal	joy	sadness	fear	disgust	surprise	love	label
1	W zalaczeniu pismo od dyrekcji w sprawie dni wolnych od pracy w 2021 r (i jeden dzien juz podany na 2022). Pozdrawiam Jan Kowalski (ang. Attached is a letter from the management regarding public holidays in 2021 (and one day already given for 2022). Best regards Jan Kowalski)	2	0	1	0	0	0	0	sadness
2	Dzien dobry, bardzo, bardzo, bardzo dziekuje. Wszystko jest juz w jak najlepszym porzadku.Pozdrawiam, Ania. (ang. Good morning, thank you very, very, very much. Everything is now in the best possible order.Regards, Ania.)	0.5	1.5	1	0	0	0.5	1.5	happy

4.2 Development Tools used for performing experiments

The Jupiter Notebook was prepared as part of the experiments. It contains all the experiments as well as downloading the necessary libraries, database reading, data preparation and summary of experiment results.

To facilitate the development process the Google Colab environment was used, as it already contains the basic Python libraries installed, and also allows the use of computing resources necessary for the machine learning processes.

The following libraries were used in this phase:

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

- *stop-words* [20] – library was used to obtain a publicly available list of stop words for the Polish language (for this purpose the list available within the *nltk* library is often used, unfortunately it does not contain a list of words for Polish language),
- *morfeusz2* [21] – library contains functions based on Polish language dictionary that allow basic operations related to vocabulary and sentence syntax; it has been used for lemmatization of single words,
- *pandas* [22] – the basic library used to work on the database,
- *gensim* [23] – library containing the used *doc2vec* text vectorization method.

4.3 Experiments verifying the impact of noise removal

Within the framework of noise removal task, the six experiments corresponding to different types of pre-processing were carried out. The basic steps for which the experiments were conducted include:

1. texts without any processing (1),
2. texts with commonly available stop-words removed (2),
3. texts with deleted stop-words selected manually (3),
4. texts after stemming (4),
5. texts after deleting commonly available stop words and stemming (5),
6. texts after deleting selected manually stop words and stemming (6).

To conduct the experiments on such a small dataset, the 5-fold cross validation [24] on datasets divided 4:1 was decided to used, where 4 is the training set and 1 is the validation set. In this way, the best result obtained on the validation set was considered as the result of the experiment.

The following activities-functions were performed to carry out the following steps:

1. Removal of publicly available stop words.
2. Removal of stop words selected manually (The stop words list differs only from the publicly available one by removing the word "no"). The list of used stop words for the Polish language is shown in figure 4
3. Lemmatization.
4. Validation.

For all texts in the performed experiments, the appropriate combination of presented functions was chosen, and before validation, the model was trained on each of the pre-processed datasets. The doc2vec model architecture was used.

The validation process implementation is shown in figure 5.

This process takes into account the comparison of expected attributed emotion with result of doc2vec model [25]. All positive and negative results were counted and then efficacy was calculated from these, comparing with the whole data set. No additional analyses were performed to check whether any of the six basic emotions are classified better or worse. This was not necessary to obtain the results of effect of noise removal on the classification result, which should be independent of the specific emotion.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

```
my_stop_words = ['ach', 'aj', 'albo', 'bardzo', 'bez', 'bo', 'być', 'ci', 'cię',
'ciebie', 'co', 'czy', 'daleko', 'dla', 'dlaczego', 'dlatego', 'do', 'dobrze', 'dokąd',
'dość', 'dużo', 'dwa', 'dwaj', 'dwie', 'dwoje', 'dziś', 'dzisiaj', 'gdyby', 'gdzie', 'go',
'ich', 'ile', 'im', 'inny', 'ja', 'ją', 'jak', 'jakby', 'jaki', 'je', 'jeden', 'jedna', 'jedno',
'jego', 'jej', 'jemu', 'jeśli', 'jest', 'jestem', 'jeżeli', 'już', 'każdy', 'kiedy',
'kierunku', 'kto', 'ku', 'lub', 'ma', 'mają', 'mam', 'mi', 'mną', 'mnie', 'moi',
'mój', 'moja', 'moje', 'może', 'mu', 'my', 'na', 'nam', 'nami', 'nas', 'nasi', 'nasz',
'nasza', 'nasze', 'natychmiast', 'nią', 'nic', 'nich', 'niego', 'niej', 'niemu', 'nigdy',
'nim', 'nimi', 'niż', 'obok', 'od', 'około', 'on', 'ona', 'one', 'oni', 'ono', 'owszem',
'po', 'pod', 'ponieważ', 'przed', 'przedtem', 'są', 'sam', 'sama', 'się', 'skąd', 'tak',
'taki', 'tam', 'ten', 'to', 'tobą', 'tobie', 'tu', 'tutaj', 'twoi', 'twój', 'twoja', 'twoje',
'ty', 'wam', 'wami', 'was', 'wasi', 'wasz', 'wasza', 'wasze', 'we', 'więc', 'wszystko',
'wtedy', 'wy', 'żaden', 'zawsze', 'że']
```

Fig. 4. List of stop words used in the experiments of noise removal.

In this paper, the main focus of the analyses is on the influence of noise reduction on the final task of prepared text vectors, namely, to understand them. Therefore, it was decided to use the classification as a determinant of the effectiveness of the used methods. What is more, the simplest of the methods was used for classification to avoid the influence of choice of the method itself on the final result.

Each model was learned for 200 epochs, where the text vector consisted of 100 elements.

```
def validate(test_doc, model1):
  true = 0
  false = 0
  total = 0
  for test in test_doc:
    vector = model1.infer_vector(test.words)
    total += 1
    if model1.docvecs.most_similar([vector])[0][0] == test.tags[0]:
      true += 1
    else:
      false += 1

  print("true: " + str(true))
  print("false: " + str(false))
  print("total: " + str(total))
```

Fig. 5. Validation process implementation for noise removal.

To summarize – the process of verification the impact of noise reduction has 3 main steps as in most text analysis tasks:

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

1. Text pre-processing - which includes in separate experiments all noise reduction scenarios.
2. Tokenization and vectorization - using doc2vec model.
3. The final processing - in this case, classification, this step is also the validation process, as the classification accuracy was chosen as the impact determinant for noise reduction.

4.4 Results of the noise reduction experiments

The results of noise reduction experiments on texts in Polish language are shown in table 6. The best results were obtained for texts that had not undergone any pre-processing. The worst results were obtained by texts for which stop words were removed.

Both the result for commonly available and chosen manually stop words is very low. If we classify 6 classes, as in this case (we classify 6 emotions), with random assignment of classes, we will get a result close to 17%. Those obtained by removing the stop word are not much better. Such a low score may result from the fact that the texts are very short, usually containing only a few single sentences. Therefore, when we take away the stop words from the texts, they do not carry much information, which makes it much more difficult to classify them well.

Table 6. Results of noise reduction experiments on texts in Polish.

Type of pre-processing on data	Max accuracy
Raw data	34%
Generally available stop words removed	18%
Selected by myself stop words removed	19%
Lemmatized texts	31%
Lemmatized texts with generally available stop words removed	30%
Lemmatized texts with selected by myself stop words removed	32%

Nevertheless, it is worth noting that the same tendency, which is present in the texts in which only stop words are removed, persists also in the texts subjected to lemmatization with the removal of stop words. Namely, the list of stop words, appropriately selected for the needs of language processing, allows to obtain higher classification efficiency.

As for the results from the model with lemmatized texts, it is not the worst, compared to the others. In this case, words from the same word family will have exactly the same representation, unlike in the case of non-lemmatized texts, where the words are treated as completely different but similar.

5 Conclusions

Text pre-processing has an impact on the final text classification accuracy – regardless of whether the final classification result will be higher or lower than

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

with raw data, one cannot help but notice the difference in performance of models trained on differently processed data. Such an impact is particularly noticeable in Polish, which, as an inflected language, carries a large part of the information specifically in the elements that are lost during preprocessing.

Using deep learning modelling methods, it is worth leaving the whole text unprocessed (more information for model to learn) – the deep learning method used within the study that created text vectors performed better with raw data. This is likely due to the fact that the texts were short and feeding them whole allowed the model to learn additional dependencies available in the texts [25, 26, 4]. If it is possible, it is worth leaving the texts unchanged, especially with usage of deep learning methods. Such texts contain the most information. Unfortunately, this is not always possible due to limitations in memory and computational resources. In addition, for the Polish language, retaining information on variety, punctuation or word order further increases the time cost of the learning process.

Choosing the right list of stop words to remove depending on the text classification task has an impact on model accuracy – most of the generally available stop word lists were prepared for classification of the subject of the utterance into not emotion or sentiment of the text, so when we want to get information about the emotion in the text, it is important to choose a stop word list for this need. Simply leaving the word 'not' in the texts allowed us to get better results of classification efficiency in Polish. The reason may be that the Polish language allows for double negation and the structure in this language is not defined so strictly.

In summary, it is worthwhile for the research to consider whether the preprocessing method used affects the final classification result too much. Especially when we perform classification in a language other than English like Polish, for which most of the pre-processing methods have been prepared.

References

1. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios (https://aclanthology.org/2021.naacl-main.201) Hedderich et al., NAACL (2021)
2. Tabassum, A., and Rajendra R. Patil: A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. In: International Research Journal of Engineering and Technology, pp. 4864–4867 (2020)
3. Haddi, E., Xiaohui L., and Yong S.: The Role of Text Pre-processing in Sentiment Analysis. In: International Conference on Information Technology and Quantitative Management Vol. 15, pp. 26–32 (2013)
4. Goyal, P., Pandey, S., and Jain, K.: Deep Learning for Natural Language Processing: Creating Neural Networks with Python, Apress 1st ed. Edition (2018)
5. Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409 (2014)
6. Khyani, D., and Siddhartha, B.S.: An Interpretation of Lemmatization and Stemming in Natural Language Processing. In: Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology, Vol. 22. pp. 350–357 (2021)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14

7. Sarica, S., and Luo, J.: Stopwords in Technical Language Processing, PLoS ONE, Vol. 16(8), e0254937 (2020)
8. Abu, E.K., and Khair, I.: Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study. In: International Journal of Computing & Information Sciences, Vol. 4, pp. 119–133 (2006)
9. Feng, Z., Fu, L.W., Xiaotie, D., and Song H.: Automatic identification of Chinese stop words, In: Research on Comp. Science, Vol. 18, pp. 151–162 (2006)
10. Camacho-Collados, J., and Pilevar, M.T.: On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In: Proceedings of 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 40–46 (2018)
11. Al Sharou, K., Li, Z., and Specia, L.: Towards a Better Understanding of Noise in Natural Language Processing. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 53–62, Held Online. INCOMA Ltd. (2021)
12. Przybyla, O.: Issues of Polish Question Answering. In: Proceedings of 1st Conference Information Technologies: Research and their Interdisciplinary Applications (ITRIA 2012), pp. 96–102 (2012)
13. Dzmitry, B., Cho K., and Bengio, I.: Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of 3rd International Conference on Learning Representations, ICLR 2015, San Diego, USA, (2015)
14. Aitken, K., Ramasesh, V., Cao, Y., and Maheswaranathan, N.: Understanding How Encoder-Decoder Architectures Attend. In: proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021) (2021)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I.: Attention Is All You Need. In: Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017), USA (2017)
16. Hochreiter, S., and Schmidhuber, J.: Long Short-term Memory. In: Neural computation. Vol. 9, pp. 1735–80. 10.1162/neco.1997.9.8.1735 (1998)
17. Tab-delimited Bilingual Sentence Pairs dataset page. http://www.manythings.org/anki/. Last accessed 1 Feb 2023
18. Emotion Dataset for Emotion Recognition Tasks. https://www.kaggle.com/datasets/parulpandey/emotion-dataset. Last accessed 1 Feb 2023
19. Piorkowska, M., and Wrobel, M.: Basic Emotions. In: Encyclopedia of Personality and Individual Differences, Editors: Zeigler-Hill, V. and Shackelford, T.K. 10.1007/978-3-319-28099-8-495-1 (2017)
20. Stop words documentation, https://pypi.org/project/stop-words/. Last accessed 1 Feb 2023
21. Kieras, W., and Wolinski, M.: (In Polish) Morfeusz 2 – analizator i generator fleksyjny dla jezyka polskiego. Jezyk Polski, XCVII(1), pp. 75–83 (2017)
22. Pandas documentation, https://pandas.pydata.org/. Last accessed 1 Feb 2023.
23. Gensim documentation, https://pypi.org/project/gensim/. Accessed 1 Feb 2023.
24. Yadav, S., and Shukla, S.: Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In: Proceedings of IEEE 6th International Conference on Advanced Computing (IACC), pp. 78–83 (2016)
25. Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents. In: Proceedings of 31st International Conference on Machine Learning, ICML (2014)
26. Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations (2013)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_14