Reviews classification – Kinga Głąbińska

Abstract

- The main aim of this project was creating classification models which would be able to decide if the review is positive or negative based on an opinion's text only (neutral opinions were rejected – reason explained in

Data analysis section). RoBERTa (A Robustly Optimized BERT Pretraining Approach) model was used to create vector representation of the opinions. Next step is creating modelsdeciding about expression (positive or negative) of an opinion. Three different approaches were implemented for comparison of the results – Random Forest Classifier, Neural Network and DBSCAN. Please noticethat the reviews used for this project have been written in polish language.

Introduction

For humans, it is rather easy to decide if an opinion is positive or negative. However, for computers it is a little more complicated. They cannot just read and understand the meaning of a sentence. This is the reason why I consider reviews classification based on text worth attention.

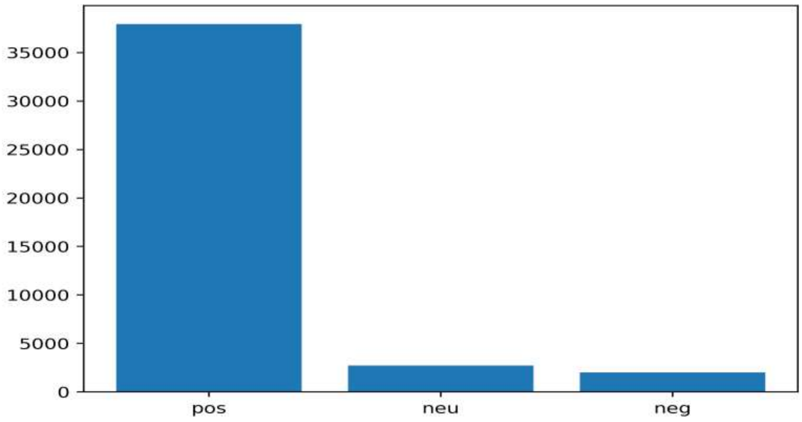
Data

Data used for this project were foundon site opineo.pl and they concern DHL – courier company. Data were collected using web scraping methods. That way36 521opinions have been downloaded.They have been saved in a csv file and contain following information:

- *Star* – number of stars in anopinion;
- *Information* – information if the opinion is positive, negative or neutral (based on *Star*);
- *Opinion* – text of the opinion.

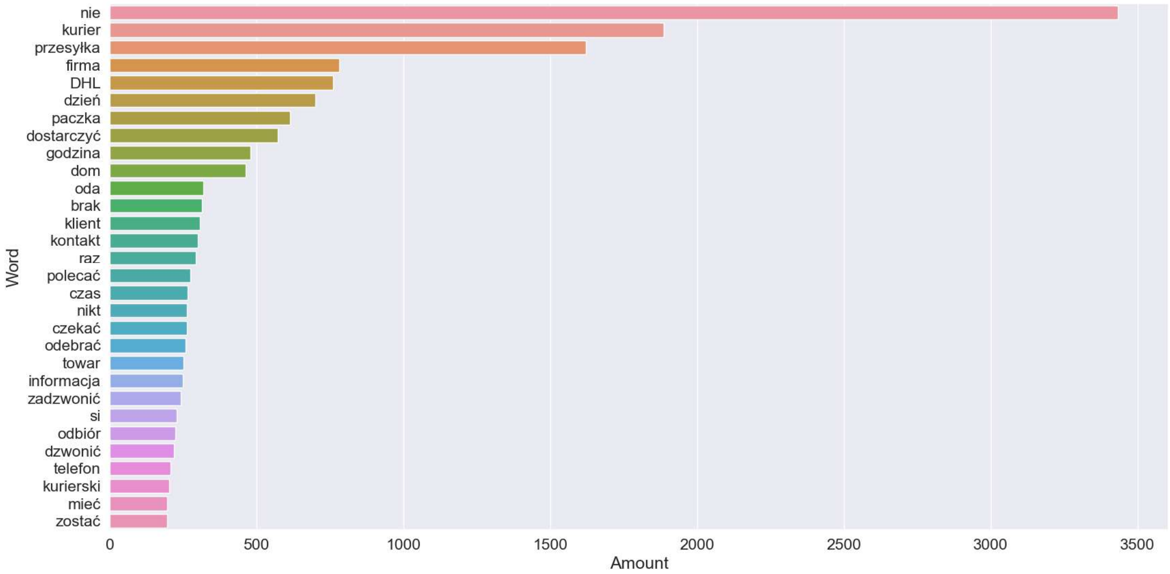
Data analysis

Let us begin by checking proportions of each type of the review. As we can see, the vast majority of opinions are positive. Negatives and neutrals are similarly frequent.

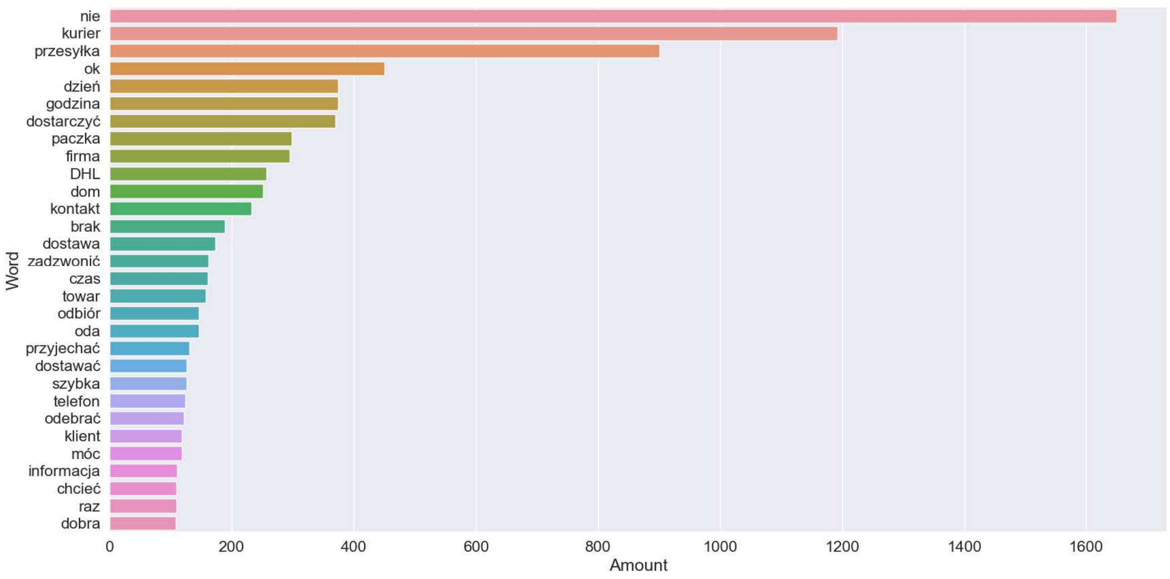


Second step in reviews' analysis was checking a word frequency. It has been done in two ways. The first one is a bar plot containing 30 most popular words and the second – word cloud. The more popular word, the bigger in the plot it is.

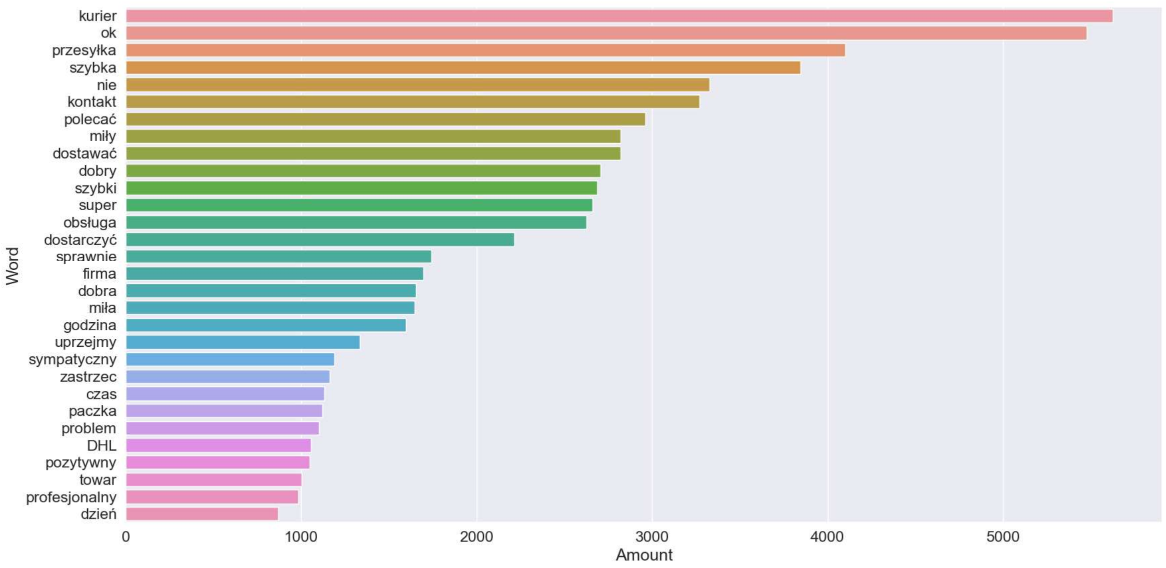
For this task lemmatized words have been used. It helped to avoid treating the same words (but in different inflections) as separate ones. The stop words have also been omitted – they do not show a real intension and they could be so frequent that they could be shown as one of the most popular. Please notice that word "nie" must be removed from the stop words. It is caused by it being ahomonym meaning that depending on the sentence it could be used as a stop word or as a negation. For this case it is crucial not to treat "nie" as a stop word because we are analyzing the reviews. Obviously in a negative review "nie" it is used mostly as a negation.



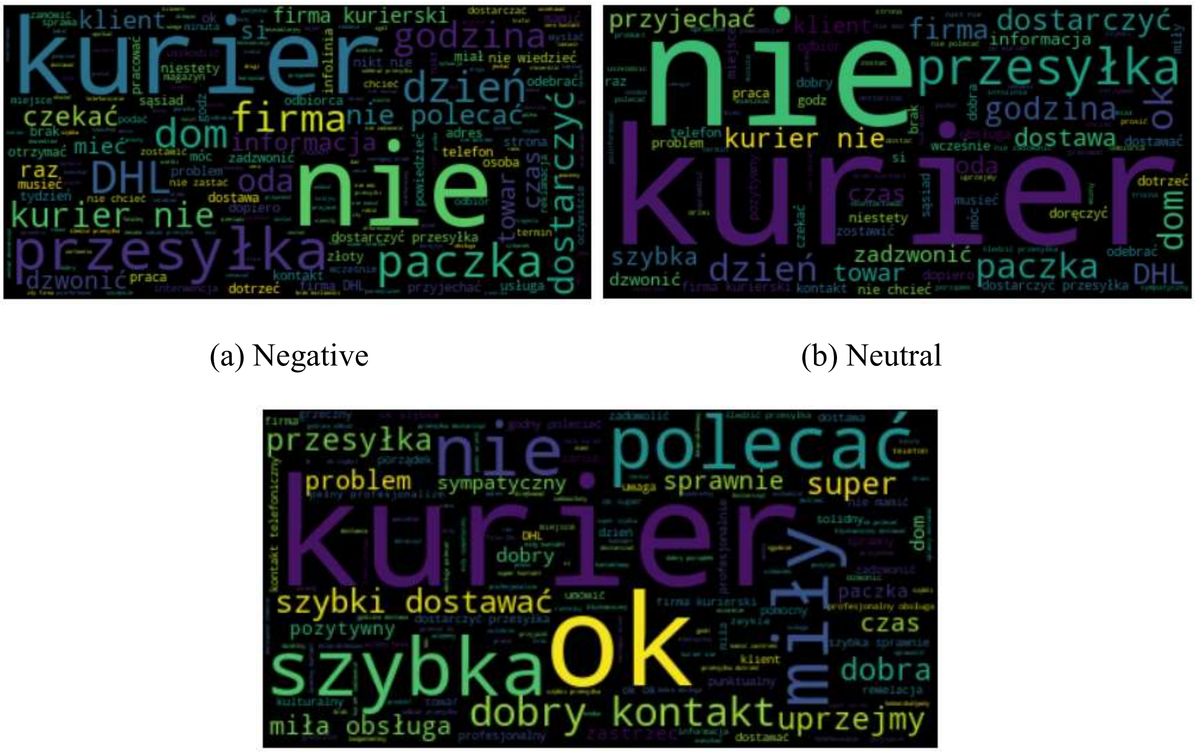
(a) Negative



(b) Neutral



(c) Positive



(a) Negative

(b) Neutral

(c) Positive

Detailed data analysishas shown that the neutral reviews should not be used in an algorithm. They are too similar to both (the positive and the negative) reviews. Therefore, even a human could have a problem finding the difference. Let us see some reviews examples:

- Positive
  - Mily i uprzejmy
  - bardzo dobra
  - szybko
  - kurier miły, pomocny, kontaktuje się przed przyjazdem.
  - Bardzo miła obsługa
- Neutral
  - Bardzo sympatyczny kontakt, punktualność, niestety brak wcześniejszej informacji o dostawie.
  - sprawna dostawa.
  - Ok
  - Dowiozl szybko ale bez kontaktu ze mna zpstawił paczke sasiadowi wiec niewiedzialem gdzie jest paczka.
  - Jestem rozczarowana i zawiedziona.Trzeci raz mnie zawiódł nie dowozi przesyłki do podanego adresu tylko przekazuje przez osoby postronne paczke takie jak np. przypadkowy sąsiad spodkany w innej miejscowosci.Nie po to płacimy wysoki koszt dostawy aby nie dowoził jej na podany adres:(

- Negative
    - Dostawa bez uszkodzeń.
    - Totalna porażka firmy kurisrskiej DHL. Przesyłka powinna byc w poniedziałek a dostałem ja w czwartek.
    - Słabo
    - Poza wszelką krytyką, skrajnie nieuprzejma obsługa, kurier zapomniał chyba że pracuje dla klienta a nie sam dla siebie...
    - Nie wiem jak w innych miastech ale ten olsztyński jest straszny. 'nie ma obowziązku informować o swoim przyjeździe'a jak się umówi na konkretną godzine to i tak jest spoźniony o 2h. TRAGEDIA!

Preprocessing

As data are downloaded directly from thewebsite, they may include information about OpiConnect (internal procedure on this website).There are 2 possible cases. The first case needs to remove an opinion from dataset due to not containing author's words. Instead of that you can see *Opinia jest w trakcieOpiConnect. Stronynawiazałykontakt w celuwyjasnieniasytuacji. Procesdialoguzakonczysie do dniadd.mm.rrrr*. In the second case there are opinions which have been in Opiconnect procedure and it finished without compromise. Then you can see the opinion, as well as the information that this opinion has been in Opiconnect procedure. *Opinia była przedmiotem dialogu w ramach procesu OpiConnect. Strony nie osiągnęły porozumienia.* In that case the reviews have been used but only after removing this information.

The reviews have been transformed to vector using the polish RoBERTa model (see https://github.com/sdadas/polish-roberta). This model is able to deal with punctuation signs, stop-words or letter cases so no more data cleaning was necessary.

Models

There were three types ofmodels that have been used. This approach helped to decide what was the best for the problem you weredealing with. In this section you can find shortly discussed machine learning methods that have been used. Please notice that 5 000 samples have been used for learning and validation. This number was chosen based on the tests results which have shown that increasing this number does not change the models' performance. Other ≈ 25 000 samples have been used for models' analysis.

DBSCAN

First approach is DBSCAN (*Density-based spatial clustering of applications with noise*) which is a method used mostly for outliers detection. It is one of unsupervised ML methods. To explain thisintuitively, we may say that it looks for clusters (we do notspecify the number of clusters on our own) and define where each element belongs. If an element is too far to belong to any cluster, then it is said to be a noise (or an outlier).

DBSCAN is a method with two most important parameters:

- Maximum distance between two samples to be considered as the neighbors.
- Minimum number of samples in a neighborhood to be considered as the core point.

There are three types of points:

- Core – elements which surely belong to the cluster (having at least min. number of required neighbors);
- Border – elements which have at least one neighbor or less than required;
- Noise – elements which have no neighbors.

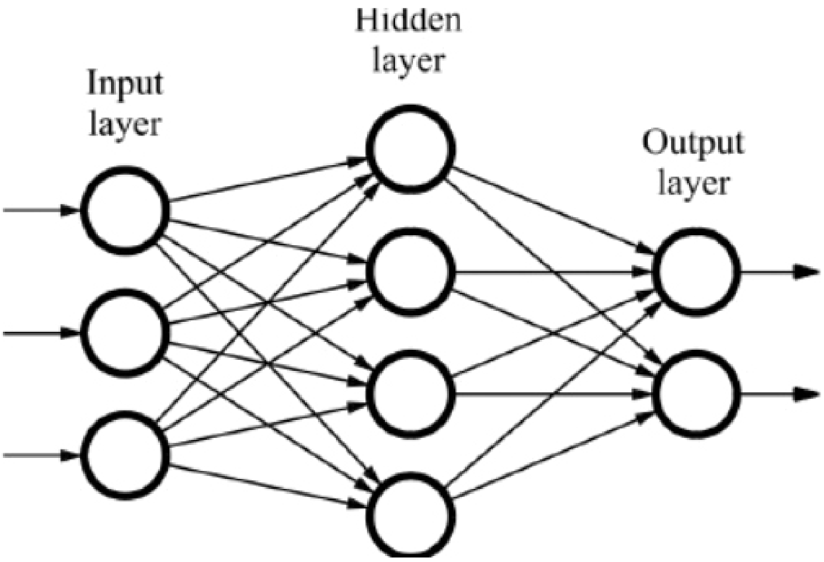
As mentioned before DBSCAN is mostly used for outlier detection. In that case, neutral samples have been rejected due to potential problems with recognizing them. Therefore, only positive and negative samples have been considered. The big difference between the amount of positive and negative samples caused using BDSCAN for this particular problem. It has been assumed that outliers are negative, and others are positive.

Random Forest

Random forest method uses decision trees which are one of the most popular decision algorithms. The decision tree consists of the so-called rootsand branches leading to the following vertices (nodes) where some decisions are made. Last nodes are called leaves and there, instead of making a decision, the algorithm returns a classification information. Random Forest is an ensemble learning method because the decision is made based on voting of the trees. Each of these trees is created on a random subset of the training data. Therefore, each tree makes a decision based on the different data.

Neural Network

Artificial neural networks (ANN) were inspired by biological neural networks. They are models containing layers. The first layer is called the*input layer,* thelast one is called the*output layer* and each one between them is called the *hidden layer*. The layers contain neurons which cannot be connected inside the layer and they are fully connected with neurons from adjacent layers (see the picture below).



Source: databricks.com

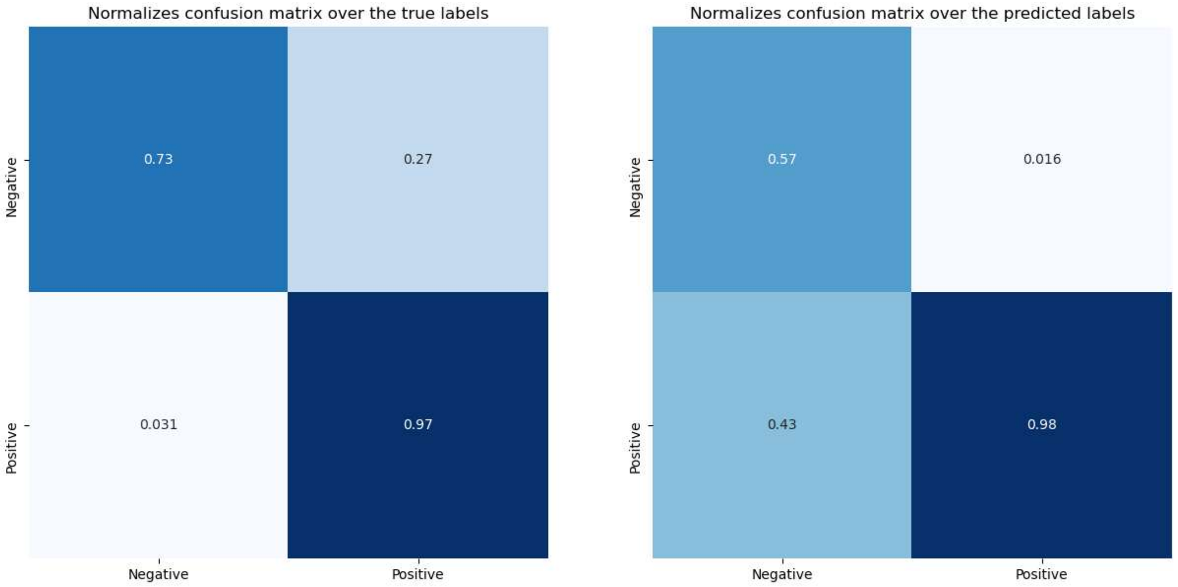
Number of neutrons in the input layer depends on features' quantity. In the output layer it depends on the number of classes (for classification problems). If there are two classes, then there is only one neuron in the output layer. Then the sigmoid function is usually used as an activation function – it returns values between 0 and 1, which can be interpreted as probability of belonging to the positive class.

Results

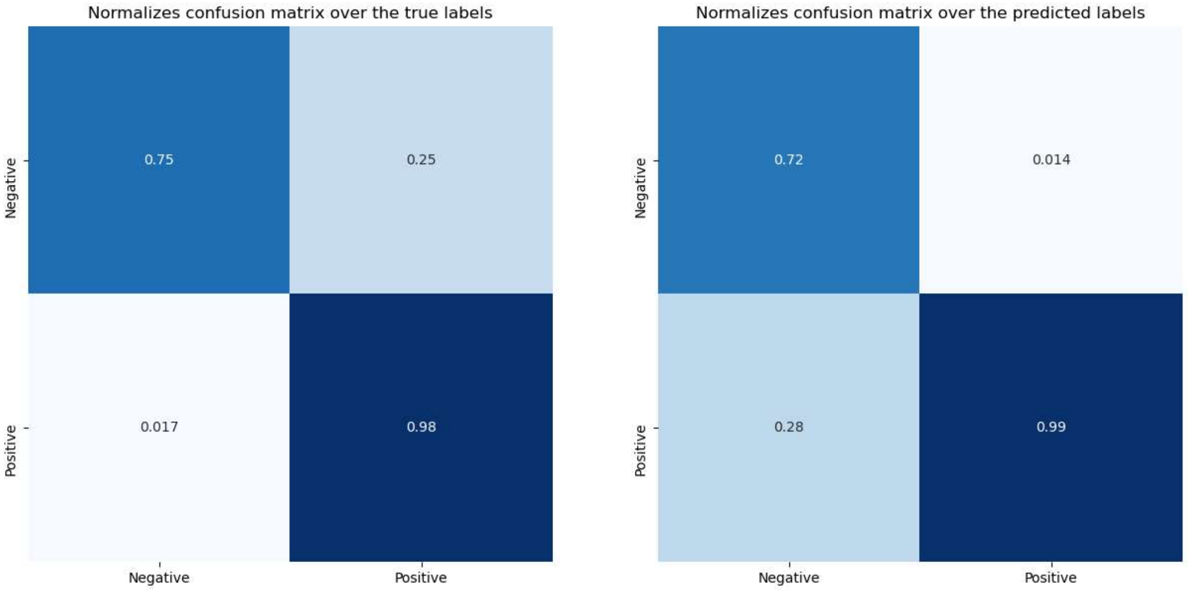
For models' comparison classification reports and confusion matrices have been used. Neural network is the best model and DBSCAN is the worst for the analyzed problem.

Metric	Model	precision	recall	f1-score
Negativesamples	RF	0,56993	0,729073	0,639753
	NN	0,717431	0,749521	0,733125
	DBSCAN	0,084409	0,964856	0,155238
Positivesamples	RF	0,984369	0,968761	0,976503
	NN	0,985741	0,983238	0,984488
	DBSCAN	0,995106	0,40574	0,576443
accuracy	RF	0,955882858		
	NN	0,970680125		
	DBSCAN	0,435781234		
macro avg	RF	0,777149	0,848917	0,808128
	NN	0,851586	0,866379	0,858806
	DBSCAN	0,539758	0,685298	0,365841
weightedavg	RF	0,962101	0,955883	0,958409
	NN	0,971325	0,97068	0,970982
	DBSCAN	0,946174	0,435781	0,553812

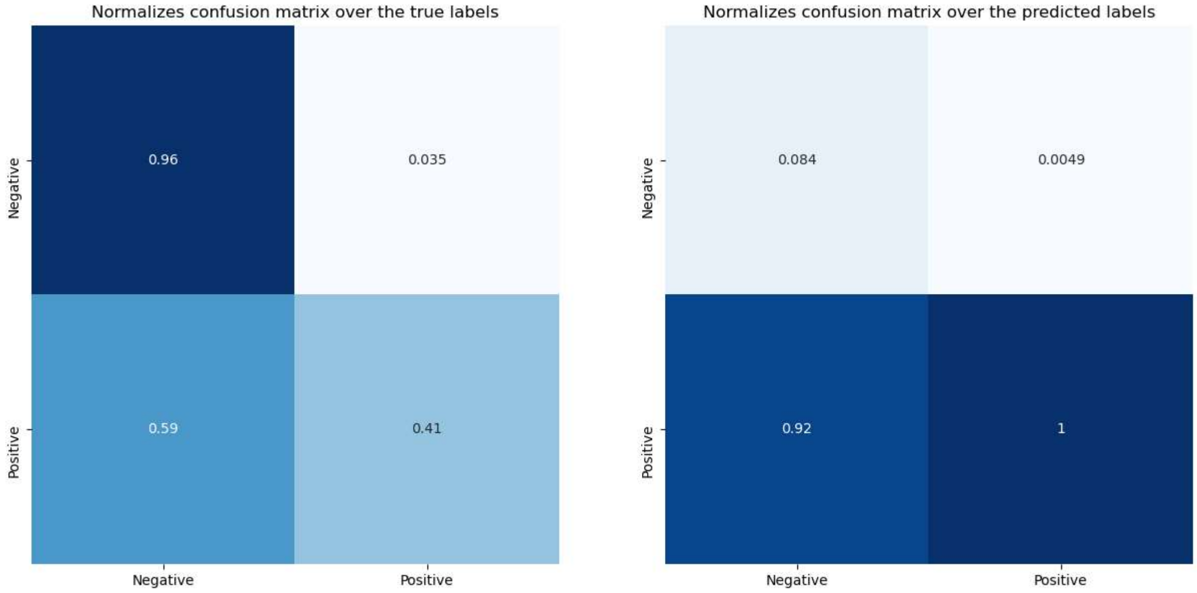
DBSCAN model is able to classify almost all the negative samples correctly (high recall) however, it does not work as goodwhen it comes to other metrics. Neural networks arethe best for most metrics. Random forest does not appear as the best one even once, it actually appears as the worst twice.Confusion matrices are presented below. They are normalized – the left matrices are normalized over the true labels and the right matrices over the predicted labels.



(a) Random Forest



(b) Neural Network



(c) BDSCAN

Summary

Sentiment analysis is an interesting task. The performed analysis has shown that based on the target, different methods should be used. For this projectthe main aim wasto achieve the model which is as good as possible in relation to each considered metrics.The neural Network was found to be the best one.