Reviews classification – Kinga Głąbińska

Abstract

The main aim of this project is creating classification models which are able to decide if review is positive or negative based on an opinion's text only (neutral opinion 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 models to decide about expression (positive or negative) of an opinion. Three different approaches were implemented for comparison the results – Random Forest Classifier, Neural Network and DBSCAN. Please notice reviews used for this project are 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 meaning of a sentence. This is reason why I consider reviews classification based on text worth attention.

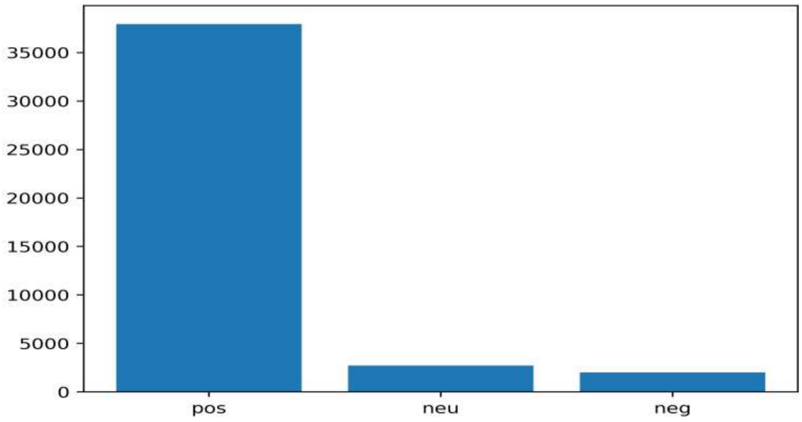
Data

Data used for this project were founded on site opineo.pl and they concern DHL – courier company. Data were collected using web scraping methods. That way 36 521 opinions were downloaded. They were saved in a csv file and contain following informations:

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

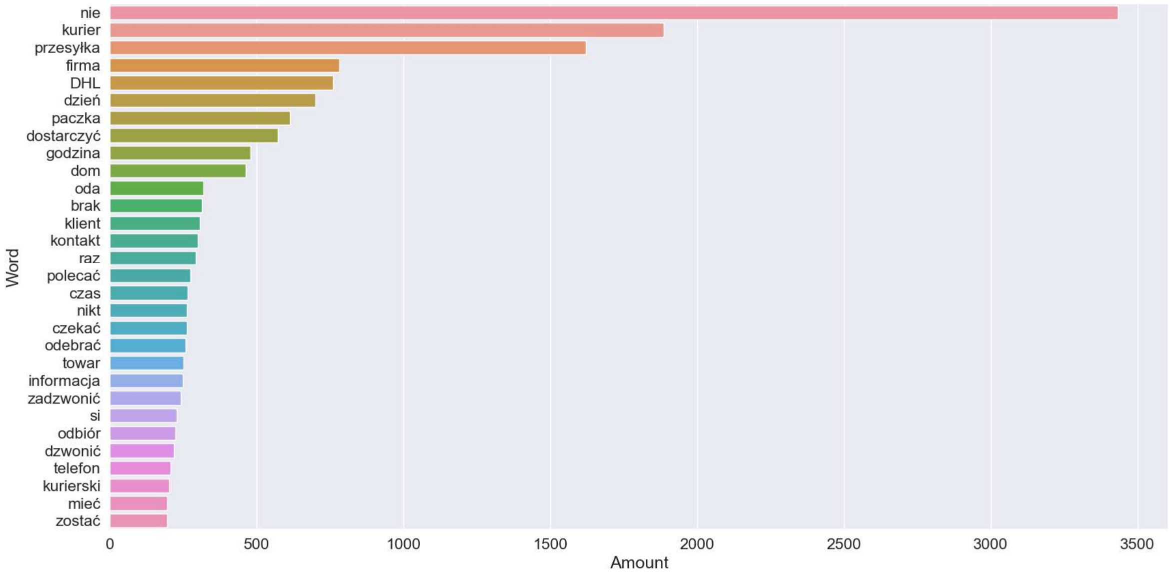
Data analysis

Let's begin with checking proportions of each type of reviews. As we can see, the vast majority of opinions are positive. Negatives and neutrals are similarly frequent.

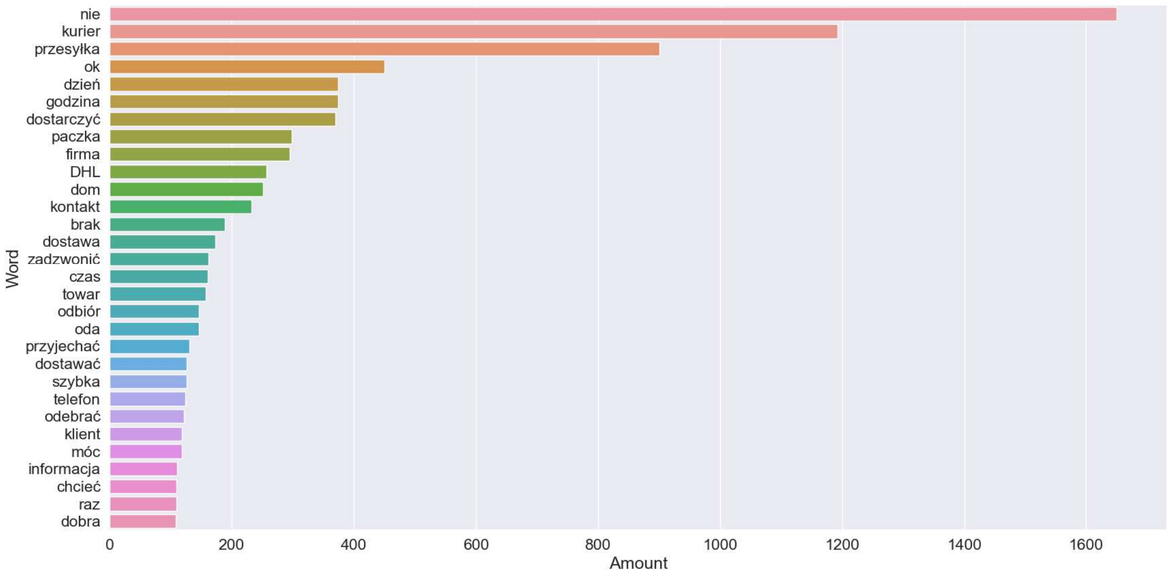


Second step in reviews' analysis was checking a word frequency. It was done in two ways. First of them is a bar plot containing 30 the most popular words and second – word cloud. The more popular word is, the bigger in the plot it is.

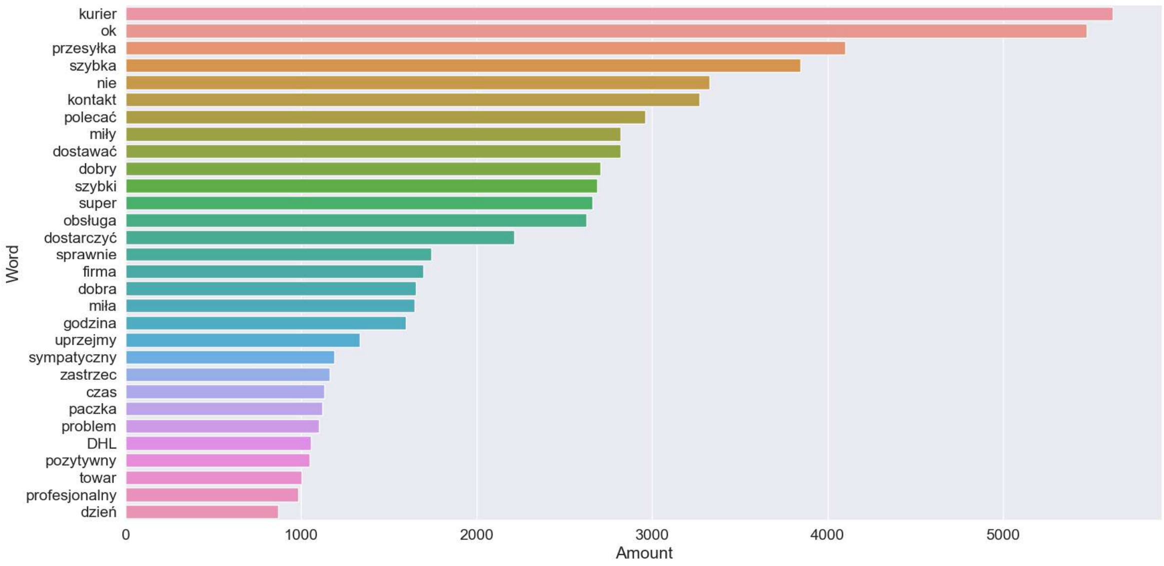
For this task lemmatized words were used. It helped avoid treat the same words (but in different inflections) as separate ones. The stop words were also omitted – they don't show a real intension and they could be such frequent so they could be shown as one of the most popular. Please notice that word "nie" must be removed from the stop words. It is caused by being a homonym so depending on the sentence it could be used as a stop word or as a negation. For that case it is very important to doesn't treat "nie" as a stop word because we analyze reviews. Obviously in negative review "nie" is mostly used as a negation.



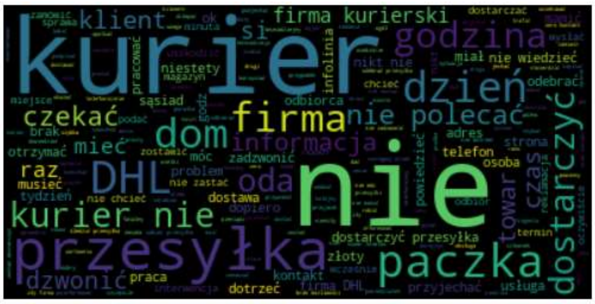
(a) Negative



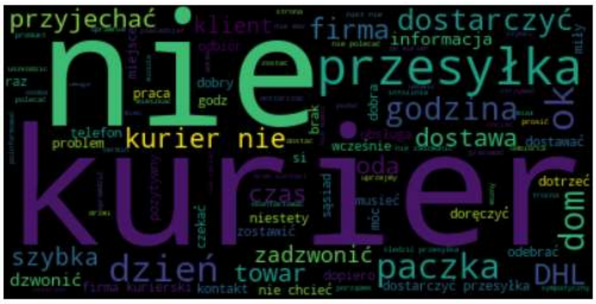
(b) Neutral



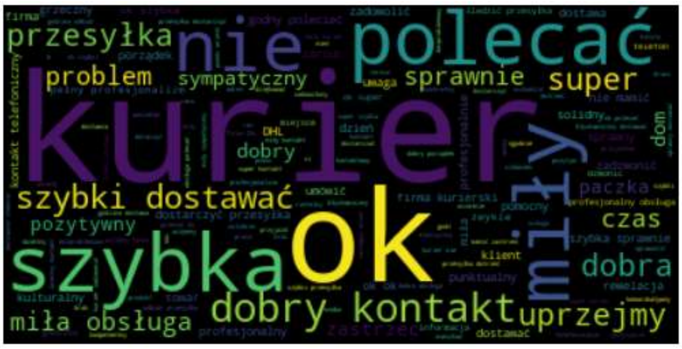
(c) Positive



(a) Negative



(b) Neutral



(c) Positive

Detailed data analysis shown that the neutral reviews should not be used in algorithm. They are too similar to both (positive and negative) reviews. Therefore even human can have a problem to find the difference. Let's 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 a website, they can include information about OpiConnect (internal procedure on this website).There are 2 possible cases. 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 trakcie OpiConnect. Strony nawiazały kontakt w celu wyjasnienia sytuacji. Proces dialogu zakonczy sie do dnia dd.mm.rrrr*. In second case there are opinions which were in Opiconnect procedure and it finished without compromise. Then you can see the opinion but also information that this opinion was in Opiconnect procedure. *Opinia była przedmiotem dialogu w ramach procesu OpiConnect. Strony nie osiągnęły porozumienia.* In that case the reviews were used but after removing this information.

Reviews were 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 were necessary.

Models

There were three types of used model. This approach help to decide what is the best for the problem you try to deal with. In this section you can find shortly discussed machine learning methods which were used. Please notice that 5 000 samples were used for learning and validation. This number were chosen based on tests which showed that increasing this number don't change the models' performance. Others (≈ 25 000) samples were used for models' analysis.

DBSCAN

First approach is DBSCAN (*Density-based spatial clustering of applications with noise*) which is a method mostly used for outliers detection. It's one of an unsupervised ML methods. To intuitively explain this, we can say that it is looking for clusters (we don't specify number of clusters on our own) and define where each element belongs. If an element is too far to belongs to any cluster, then it is said to be noise (or outlier).

DBSCAN is method with two the 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 a core point.

There are three types of points:

- Core – elements which surely belongs to cluster (having at least min. numbers of required neighbors);
- Border – elements which have at least one neighbor and less than required;
- Noise – elements which have no neighbors.

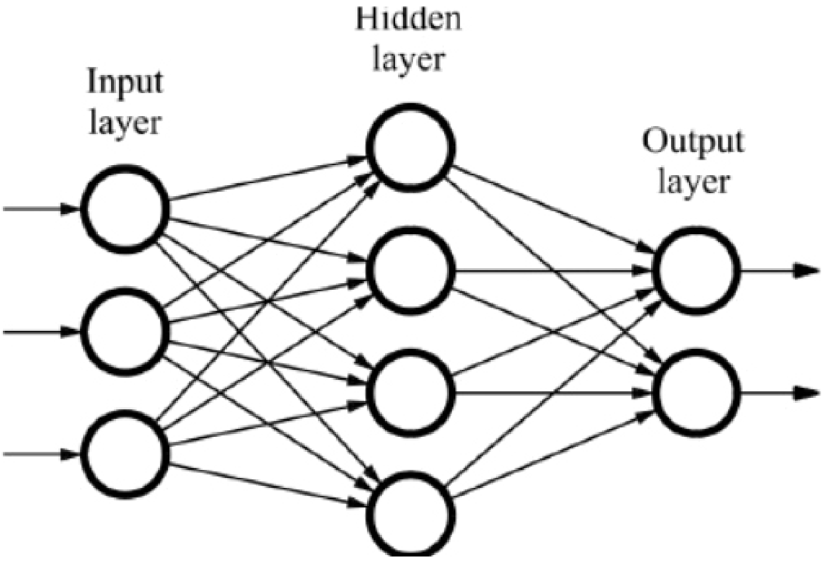
As mentioned before DBSCAN is mostly used for outlier detection. In that case, neutral samples were rejected due to potential problems with recognizing them. Therefore only positive and negative samples were considered. Big difference between amount of positive and negative samples caused using BDSCAN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

Random Forest

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

Neural Network

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



Source: databricks.com

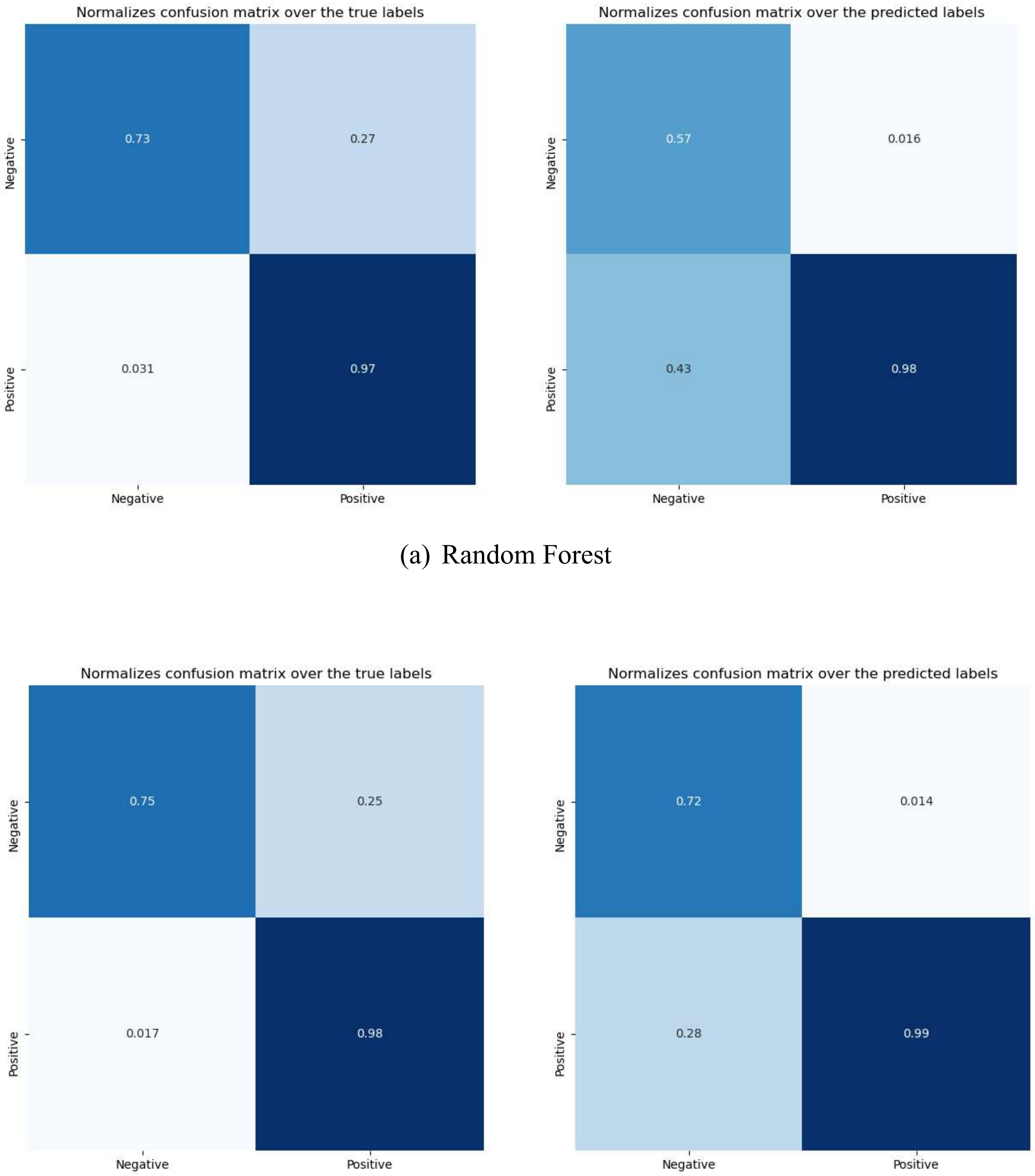
Number of neutrons in the input layer depends on features' quantity. In the output layer it depends on amount of classes (for classification problems). If there is two classes, then there is only one neuron in output layer. Then usually sigmoid function is used as an activation function – it returns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

Results

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

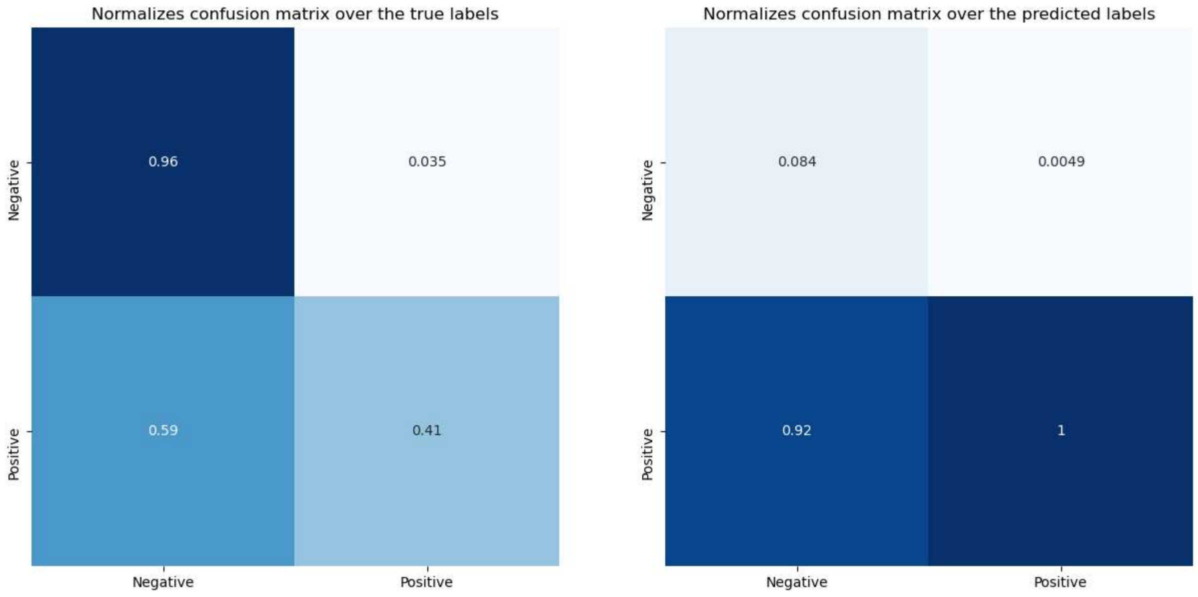
Metric	Model	precision	recall	f1-score
Negative samples	RF	0,56993	0,729073	0,639753
	NN	0,717431	0,749521	0,733125
	DBSCAN	0,084409	0,964856	0,155238
Positive samples	RF	0,984369	0,968761	0,976503
	NN	0,985741	0,983238	0,984488
	DBSCAN	0,995106	0,40574	0,576443
accuracy	RF	0,955882858		
	NN	0,970680125		
	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
weighted avg	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 negative samples correctly (high recall) however it is weak for others metrics. Neural networks is the best for most metrics. Random forest is never the best and only twice the worst. Confusion matrices are presented below. They are normalized – left matrices are normalized over the true labels and right matrices over the predicted labels.



(a) Random Forest

(b) Neural Network



(c) BDSCAN

Summary

Sentiment analysis is interesting task. The analysis performed showed that based on target different methods should be used. For this project main aim was achieve model which is as good as possible in relation to each considered metrics so Neural Network was found as the best one.