Romanian Sentiment Analysis Dataset for Book Reviews

Monica A. Gîrbea
University of Bucharest

Alexandru C. Sasu
University of Bucharest

Abstract

In this paper, we present a custom dataset for the task of sentiment analysis in the Romanian language focusing on book reviews. The purpose of this project was to contribute with a reliable resource to the limited corpus of Romanian-based datasets. The dataset contains positive, negative, and neutral reviews. We perform extensive data analysis on the dataset, as well as train and evaluate a couple of baseline experiments that show promising generalization capabilities.

1 Introduction

Sentiment analysis, or opinion mining, is a natural language processing (NLP) task that determines the emotional tone of a text or message, placing it into one of three categories: positive, negative, or neutral. The data that is usually used for this task consist of emails, customer support chats, social media interactions, and reviews.

Widely spoken languages, such as English, have a lot of resources for NLP tasks, but there is a significant lack of publicly available datasets for languages like Romanian. Like any language, Romanian displays a diverse range of expressions, complex morphology, and internet slang, and thus needs its own datasets that encapsulate these finer details.

In this paper, we introduce a dataset designed to aid sentiment analysis in Romanian. The topic approached is centered on book reviews, an area where people tend to use rich vocabulary and in-depth descriptions of their feelings and opinions.

The data was gathered from Goodreads, an online database of books where users can make their own reading lists, interact with each other, and most importantly in our case, write reviews.

This paper documents the steps that were taken in order to build the dataset.

2 Related Work

2.1 Sentiment Analysis in English

As a universally spoken language, datasets for the English language dominate in areas such as NLP. For sentiment analysis, we can list a few noteworthy ones:

- Social Media Sentiment (Parmar, 2024)

  A dataset made up of user-generated content from multiple online platforms such as X (formerly known as Twitter), Instagram, and Facebook. Besides the classic positive, negative and neutral labels, it contains information about the posts themselves, such as the name of the user, the timestamp, hashtags and interaction counts such as likes and re-posts.

- Multilingual Amazon Reviews Corpus (MARC)(Keung et al., 2020)

  A dataset that focuses on reviews available on the e-commerce platform Amazon and includes the following languages: English, Japanese, German, French, Spanish, and Chinese.

- Yelp Reviews (Zhang et al., 2016)

  Another review based dataset that centers around information from the site Yelp, platform dedicated solely to reviews of businesses. The information present in this dataset is the reviews and their corresponding labels, expressed through the numbers of stars. (one to five).

- Emotion (Saravia et al., 2018) Another dataset that makes use of X (Twitter), but this time there are 6 labels present: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

2.2 Sentiment Analysis in Romanian

- LaRoSeDa - A Large Romanian Sentiment Data Set (Tache et al., 2021)

  A dataset made up of 15000 samples taken from reviews from a Romanian e-commerce platform. The data is divided into two categories: positive and negative. This is achieved by taking the original 1 to 5 star rating from the site and considering only 1 and 2 stars (negative) and 4 and 5 stars (positive).

- SART - Sentiment Analysis from Romanian Tweets (Ciobotaru and Dinu, 2023) This dataset makes use of a preexisting one, made up of 2000 positive and negative samples. The authors have improved the dataset, not only by the addition of new samples, but also by introducing the negative label.

- Movie Reviews in Romanian (Dumitrescu et al., 2020b) The dataset contains movie reviews labeled as positive or negative.

- Sentiment Analysis from Stock Market News in Romanian (Stoean and Lichtblau, 2021)

  This paper provides an alternative to the creation of new datasets specific for Romanian, which consists of translating the original text into English and making use of pre-existing tools for classification.

3 Method

3.1 Creation of the Dataset

The idea for this paper was to create a resource for Romanian like there are for languages like English. As shown prior, the majority of sentiment analysis datasets focus on social media interactions and reviews. X (Twitter), is a popular source of user information since the posts on the site are quite short and expressive. Another popular source is made up of platforms dedicated to products and their reviews.

To our knowledge, there are no sentiment analysis datasets in Romanian for book reviews. We chose to focus on this topic, not only because of the lack of resources, but also due to the rich language that such content can provide. Unlike tweets or product reviews, book reviews tend to use more formal language that could not translate well when used in combination with other datasets.

The platform we chose to source was Goodreads, as it offers a vast variety of books in its database and it is frequented by many users. On this site, users can leave reviews along with a star-based scoring system that goes from one to five (only integer numbers).

In order to obtain reviews from the website, a semi-automatic method was employed. We would manually search for books from different genres and with a variety of ratings, and when we found the desired books, we would then scrape all Romanian reviews from them with the help of an automatic scraper. The scraper was written in Python and used Selenium to interact with the website, as well as Beautiful Soup to realize the effective scraping.

Given the rating system of the website, we decided to divide our sentiment analysis classes as such:

- 1/2 stars => negative (label 0)
- 3 stars => neutral (label 1)
- 4/5 stars => positive (label 2)

In order to minimize the bias in our dataset, we have limited the data to books belonging to 10 general categories, or genres:

- Fiction
- Romanian Literature
- Classics
- Romance
- Horror
- Young Adult
- Thriller
- Historical Fiction
- Fantasy
- Psychology

Moreover, also with the purpose of minimizing bias, we strove to collect data from as many authors as possible, as well as avoid including more than one book from any one series.

The dataset has 4 columns: "id" (integer numbers created by us to identify entries), "genre", "review", and "label".

For the creation of the dataset 118 books were used, resulting in 6661 samples.

3.2 Dataset Cleanup

The next step was cleaning the dataset of reviews that contained erroneous or inconsistently formatted data such as reviews that consisted only of a link to an outside source where the review was presented. These entries could be considered outliers due to their inconsistent format in comparison to other descriptive reviews, and therefore needed to be taken out.

Other actions included in the dataset cleaning step were: dropping entries that were missing the value in the review column, getting rid of leading and trailing whitespaces, dropping duplicate reviews, and a more in-depth outliers removal.

For the outliers removal process, a multilingual pretrained model was employed in order to obtain the embeddings of reviews, namely the sentence transformer paraphrase-multilingual-MiniLM-L12-v2 from Hugging Face. The choice of using a multilingual model instead of one specialized on Romanian was made due to the fact that some reviews contained English words in-between Romanian ones (this practice is called "romgleză"). Subsequently, an Isolation Forest model with a contamination factor of 5% was used to determine the outliers. Based on the output of the Isolation Forest model (label -1 for outliers and label 1 for non-outliers), the corresponding samples were removed from the dataset.

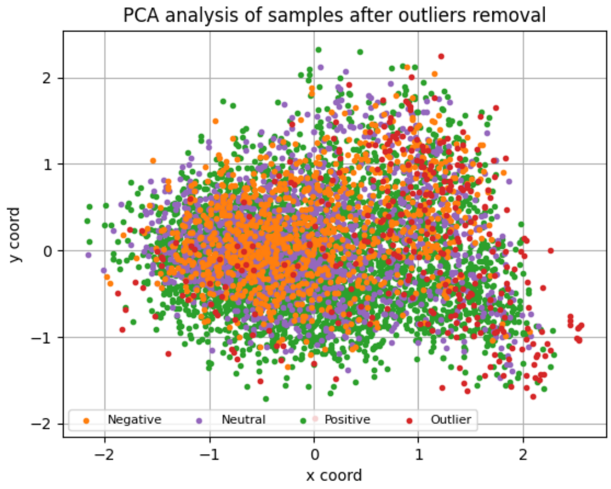

Figure 1: Visual representation of samples during dataset cleanup.

After this step, the dataset contained 6283 samples, with the following label distribution:

Label	No. samples
0	939
1	1445
2	3899

Table 1: Label distribution after dataset cleanup.

and the following genre distribution:

Genre	No. samples
Romanian Literature	1131
Fiction	1082
Classics	964
Romance	749
Fantasy	506
Psychology	492
Horror	411
Young Adult	337
Thriller	309
Historical Fiction	302

Table 2: Genre distribution after dataset cleanup.

3.3 Data Preprocessing

This step involved preparing the data before passing it to ML models and further analyzing it, and it included:

- Normalizing the text font of reviews
- Replacing inconsistent characters. Diacritics and quotes are sometimes different from the standard, and thus we had to replace ş, Ş, ţ, Ţ, “, ” with ș, Ș, ț, Ț, "
- Removal of links starting with http(s)
- Removal of English reviews. Some people write their reviews in two versions: Romanian and English; so we had to keep only the Romanian review

We have refrained from adding more preprocessing methods, such as text to lowercase, stop words removal, stemming, etc., due to fearing that some models might have performed worse (such as cased models).

4 Exploratory Data Analysis

On average, each type of review had the following length in terms of characters:

- negative reviews: 625
- neutral reviews: 741
- positive reviews: 829

As can be seen, the more positive the review, the lengthier it is.

Upon converting the samples into embeddings (with the help of the model from 3.2) and plotting them into a 2D space with the help of the PCA algorithm (see Figure 2), it was observed that semantically, a vast majority of reviews were really similar, regardless of their label.

This, combined with the fact that a lot of reviews were also of considerable length, again regardless of the associated label, points to the idea that performing sentiment analysis on this dataset may be difficult, as all three types of reviews are really descriptive and thus have a high degree of similarity.

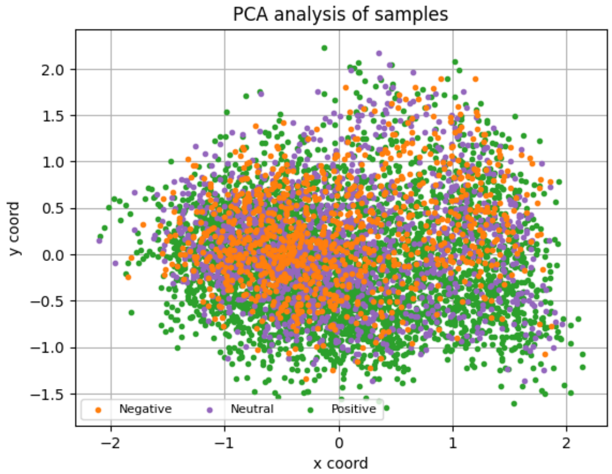

Figure 2: Visual representation of samples.

During the inspection of the distribution of negative and positive words – which were taken from the Romanian lexicon constructed by Chen and Skiena (2014), that was further processed by us so that it did not contain stop words, thus resulting in 2035 negative words and 1283 positive ones – in all reviews of each type, we noticed that the quantity of positive words exceeds that of negative words (see Figure 3), even though the lexicon contained considerably more negative words. The fact that even negative and neutral reviews contained more positive than negative words might also generate difficulties for ML algorithms.

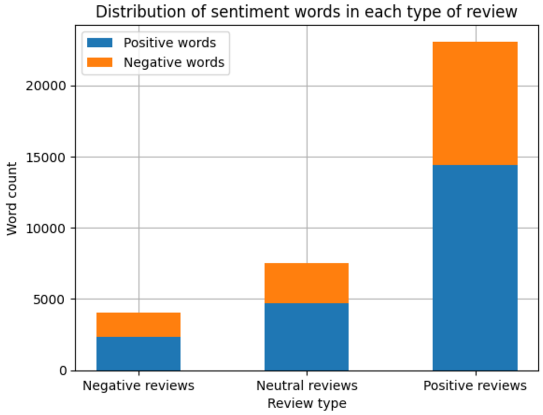

Figure 3: Negative and positive words distribution.

We have also performed a parts-of-speech analysis on our dataset, both intra- and inter-review type.

For the intra-review type analysis, only four parts of speech were chosen, that were deemed the most important for the sentiment analysis task, namely: noun, adverb, verb, and adjective. Given these parts of speech, we checked how much the ratio between PoS 1 and PoS 2 was kept between classes, by further calculating the ratio of those ratios between a class and another class. For most of the ratios, the value was between 0.9 and 1.1, thus not providing too much useful information to distinguish between sentiment classes, except for four values that were slightly out of this range: 0.844, 1.132, 1.164, 1.187; (see Table 3) and may aid in the classification of reviews.

For the inter-review type analysis, we wanted to check whether the ratios between parts of speech remained the same between any two classes. For all parts of speech analyzed, the ratio was largely consistent, except for subordinating conjunctions, where the ratio was slightly higher than others (see Table 4), which may also aid in the classification of reviews.

Class	NOUN:ADV	NOUN:VB	NOUN:ADJ
0 → 1	0.919	1.006	1.041
0 → 2	**0.844**	0.983	1.003
1 → 2	0.919	0.977	0.963
Class	ADV:VB	ADV:ADJ	VB:ADJ
0 → 1	1.095	**1.132**	1.034
0 → 2	**1.164**	**1.187**	1.019
1 → 2	1.063	1.047	0.985

Table 3: Intra-review type analysis.

POS	0:1	0:2	1:2
NOUN	0.541	0.175	0.324
AUX	0.578	0.203	0.351
ADP	0.555	0.180	0.325
PRON	0.568	0.185	0.326
ADV	0.589	0.207	0.353
DET	0.552	0.176	0.320
VERB	0.537	0.178	0.331
ADJ	0.519	0.175	0.336
CCONJ	0.551	0.183	0.332
SCONJ	**0.634**	**0.234**	**0.369**
NUM	0.529	0.186	0.351

Table 4: Inter-review type analysis.

5 Machine Learning Experiments

In this section, we present the approach and the results of training and testing our dataset with machine learning models. The models chosen were pre-trained on Romanian texts.

Our dataset was split into train and test data – we didn't have a validation split, as we only sought a few baseline results – and in order to avoid bias due to class imbalance, we balanced the class weights for our experiments. These details are presented in Table 5.

Label	Train	Test	Class weight
0	752	187	2.2304
1	1157	288	1.4494
2	3117	784	0.5371

Table 5: Dataset split and class weights.

On the newly formed train and test datasets, we applied the preprocessing described in the previous section, along with the tokenization specific to the models used. The models were trained for 25 epochs, with early stopping applied when they showed no improvement. The metrics followed were training loss, test loss and test f1 score.

5.1 Models

Further ahead, we showcase the results for each of the models used with the help of classification reports and confusion matrices.

- BERT base, cased model for Romanian (Dumitrescu et al., 2020a)

	precision	recall	f1-score	support
0	0.51	0.68	0.59	187
1	0.39	0.40	0.39	288
2	0.84	0.77	0.80	782
accuracy			0.67	1257
macro avg	0.58	0.62	0.59	1257
weighted avg	0.69	0.67	0.68	1257

Table 6: Classification report for BERT base.

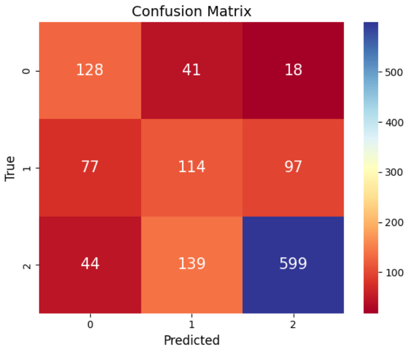

Figure 4: Confusion matrix for BERT base.

- RoBERT-base (Masala et al., 2020)

	precision	recall	f1-score	support
0	0.56	0.68	0.62	187
1	0.40	0.41	0.41	288
2	0.84	0.79	0.81	782
accuracy			0.69	1257
macro avg	0.60	0.63	0.61	1257
weighted avg	0.70	0.69	0.69	1257

Table 7: Classification report for RoBERT-base.

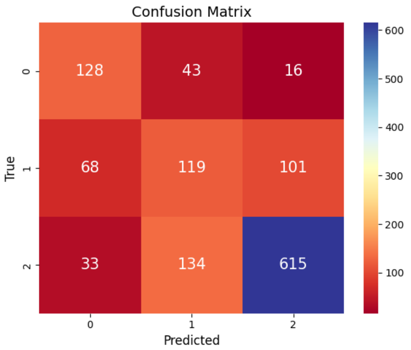

Figure 5: Confusion matrix for RoBERT-base.

As we can see in Table 7, classes 0 and 2 perform better than class 1. All classes tend to "leak" into the neighboring classes, as observed in the confusion matrices, seeing that, for example, most wrongly labeled samples from class 2 are predicted as class 1. This is quite normal, but it becomes an obvious problem for middle classes like our neutral label. Class 1 is confused with both class 0 and class 2, resulting in a lot of wrongly predicted samples.

In an attempt to solve this issue, the weights of the classes were modified, adding slightly more weight to the middle class. This was achieved by multiplying the original value with 1.2, while the other classes were left the same. This was tested on the

RoBERT-base model and the results are displayed in Table 8.

	precision	recall	f1-score	support
0	0.58	0.63	0.61	187
1	0.39	0.49	0.43	288
2	0.85	0.76	0.80	782
accuracy			0.68	1257
macro avg	0.61	0.63	0.61	1257
weighted avg	0.71	0.68	0.69	1257

Table 8: Classification report for RoBERT-base with modified weights.

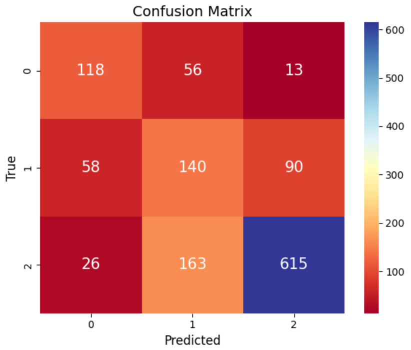

Figure 6: Confusion matrix for RoBERT-base with modified weights.

As it can be seen, while increasing the weight for class 1 resulted in better results for for this class, the other two classes were negatively affected, thus not resulting in an increase of overall performance.

With this same model, an alternative to balanced weights was tried, which consists of undersampling so that all of the classes have a similar amount of labels. We can see that the previous problem with class 1 persists in this case as well (see Table 9 and Figure 7).

	precision	recall	f1-score	support
0	0.66	0.70	0.68	183
1	0.51	0.39	0.44	188
2	0.65	0.75	0.70	193
accuracy			0.62	564
macro avg	0.60	0.61	0.61	564
weighted avg	0.60	0.62	0.61	564

Table 9: Classification report for RoBERT-base with undersampled dataset.

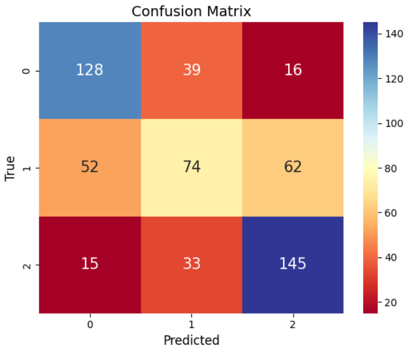

Figure 7: Confusion matrix for RoBERT-base with undersampled dataset.

- RoBERT-small (Masala et al., 2020)

Another version of the previous model was tried, the difference between them being represented in Table 10.

Model	Weights	L	H	A	MLM accuracy
RoBERT-small	19M	12	256	8	0.531
RoBERT-base	114M	12	768	12	0.651

Table 10: Different versions of RoBERT.

RoBERT-small is faster, due to its reduced size, but it accentuates previously mentioned problems: class 1 is predicted as either 0 or 2 and now even class 2 is heavily predicted as 0, which only happened in moderation before.

	precision	recall	f1-score	support
0	0.35	0.64	0.45	187
1	0.33	0.25	0.29	288
2	0.81	0.73	0.77	782
accuracy			0.60	1257
macro avg	0.50	0.54	0.50	1257
weighted avg	0.63	0.60	0.61	1257

Table 11: Classification report for RoBERT-small.

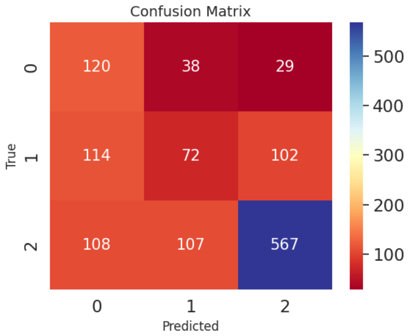

Figure 8: Confusion matrix for RoBERT-small.

- Bert Legal Romanian (Ceausu and Nisioi, 2022)

  This model represents a failed attempt, as it presented unusable results. This could be caused by the nature of this model, as it was trained on legal documents and thus contains a lot of domain-specific lexicon.

	precision	recall	f1-score	support
0	0.15	0.99	0.26	187
1	0.00	0.00	0.00	288
2	0.60	0.00	0.01	782
accuracy			0.15	1257
macro avg	0.25	0.33	0.09	12
weighted avg	0.40	0.15	0.04	1257

Table 12: Classification report for Bert Legal Romanian.

5.2 Results

To summarize the previously presented experiments, we have put together the important results for the models in Table 13.

Model	Macro-Avg. F1 Score
BERT base	0.59
RoBERT base	**0.61**
RoBERT small	0.50
BERT Legal	0.09

Table 13: Performance for all the models.

It appears that RoBERT-base performed the best, with a macro F1 score of 61%. Although not satisfactory for the sentiment analysis task, it shows promising results and may indicate room for improvement with additional preprocessing, feature extraction, and model (hyper) parameters tuning.

6 Conclusion

We have introduced a new Romanian corpus for sentiment analysis, as well as performed data analysis and tested multiple pretrained Romanian models on it in order to get a basic understanding of its capabilities. Our exploratory data analysis showed that all three types of classes are mostly similar, which may create difficulties in training AI models. Despite this, while training a few baseline models on our dataset, we have managed to obtain at most 61% accuracy on the test data, indicating that there could be room for improvement in terms of model strategies, even with the current state of the dataset.

Future works that may be performed on our dataset could take advantage of the intra- and inter-class analysis of multiple parts of speech performed during our exploratory data analysis. Moreover, further investigation of the genre column may also prove useful in the classification task.

References

Corina Ceausu and Sergiu Nisioi. 2022. Identifying draft bills impacting existing legislation: a case study on Romanian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3670–3674, Marseille, France. European Language Resources Association.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 383–389. Association for Computational Linguistics.

Alexandra Ciobotaru and Liviu P. Dinu. 2023. Sart covid-sentiro: Datasets for sentiment analysis applied to analyzing covid-19 vaccination perception in romanian tweets. In *Proceedings of the 27th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2023)*, Athens, Greece.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020a. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Stefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020b. The birth of romanian bert. *arXiv preprint arXiv:2009.08712*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert–a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.

Kashish Parmar. 2024. Social media sentiments analysis dataset.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Catalin Stoean and Daniel Lichtblau. 2021. Sentiment analysis from stock market news in romanian using chaos game representation. In *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 252–258.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 949–956, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.