Towards the Development of a Romanian Lexicon for the Analysis of Emotions in the Literary Works of Canonical Authors

Veronica GAVRILĂ, Lidia BĂJENARU*, Ciprian DOBRE, Mihaela TOMESCU

National Institute for Research and Development in Informatics – ICI Bucharest,
8-10 Mareşal Averescu Avenue, 011455, Bucharest, Romania
veronica.gavrila@ici.ro, lidia.bajenaru@ici.ro (**Corresponding author*),
ciprian.dobre@ici.ro, mihaela.tomescu@ici.ro

Abstract: The analysis of emotions is still considered a rather difficult procedure and it is constrained by the limitations related to the languages available for this analysis. The Romanian language is one of the richest and most complex languages with different dialects, regionalisms, archaisms and a powerful and expressive poetic language. Starting from the original vision of the poets, the expressiveness of the Romanian language has given birth to special metaphors that often lose their meaning if one tries to translate them. For this reason, the need arose to develop a lexicon specific to the Romanian language that would allow the analysis of emotions from different poems and to successfully identify the emotions presented by an author through the artistic images created in his work. This paper presents the first steps in the development of a Romanian lexicon based on methodologies employed in other languages, mainly the English language. The data set was subject to a filtering and refining process after which it was enriched by using specific tailored web crawlers for adding all the forms and conjugations of the respective words.

Keywords: Lexicon, Natural language processing, Emotion analysis, Crawlers, INTELLIT, Romanian language.

1. Introduction

To a certain extent, all people are guided by emotions, guided by what they feel and depending on the personality of each individual people display these emotions differently. Assessing emotions is important and quite easy for a majority of individuals, that is, when someone is face to face with different persons, that person can "read" the others and see how they feel up to a certain point, but this "reading" is getting complicated when one should interpret a text. Because every person processes emotions differently, it is perfectly normal to interpret them differently.

The emotion analysis in this paper is based on the interpretation of emotion starting from the 8 primary emotions described by Robert Plutchik (1980a) up to classifying the type of each emotion (positive, negative, neutral) within literary texts by using techniques specially created for text analysis, different machine learning algorithms or the detection of emotions based on lexicons (list of words and the associated emotion).

Currently, the analysis of emotions in texts is still considered a difficult procedure, as this analysis is very subjective and differs greatly depending on the context. For example, words may have different connotations depending on the language / context in which they are used (words may have a different meaning if they are used in literary works or speeches etc.). Also, not all words have the same translation from one language into another one and certain words cannot be translated at all. One of the most important factors, especially if one employs methods based on lexicons, is that people perceive and express emotions in different ways and thus one word can express many emotions or no emotion at all, depending on the context and how it was used.

This paper focuses on extracting emotions from works published by Romanian canonical authors. The information about canonical authors represents a part of the data provided by the INTELLIT platform (Gavrilă, Băjenaru & Dobre, 2019a). The INTELLIT platform aims to be the largest digital archive in Romania containing data regarding our national literary heritage. All the information is provided by the Romanian Academy, the Institute of Literary History and Theory and the National Foundation for Science and Art in collaboration with the Publishing House of the Museum of Romanian Literature (Simion, 2005; Simion, 2006). The platform provides information about the life and work of the Romanian authors, about important events in our culture (complete chronology of events from 1994 to 2000), information about all the authors from A to Z, different works written

by the Romanian canonical authors and even an English version of the general dictionary of Romanian literature. To preserve the literary heritage, the INTELLIT platform was developed using the latest technologies. A more detailed presentation of this work can be found in (Gavrilă, Băjenaru & Dobre, 2019b).

The Romanian language is considered one of the richest and most complex languages and because of this it is much more difficult to identify the specific feelings expressed in the context in which they are applied. As the current context is the literary one, this makes the analysis of sentiments even more difficult, because in the works of canonical authors words may be used with their metaphorical meaning and not with the original (basic) one, or a specific regionalism / archaism can be used depending on the author's area. Some words are specific to the Romanian language and have no translation into other languages.

This paper is organized as follows. Section 2 presents the original theory of emotions described by R. Plutchik, the creator of the wheel of emotions. Section 3 sets forth an overview of related work and lexical databases. Section 4 describes the steps taken in order to develop the Romanian lexicon and proposes possible future work for improving the developed solution. Section 5 outlines the conclusion and the plans for the future starting from the present stage.

2. The Theory of Emotions

Starting from Robert Plutchik's theory of emotions (Plutchik, 1980a) there are 8 primary feelings, namely anger, fear, sadness, disgust, surprise, anticipation, trust and joy. Moreover, he created a wheel of emotions in which he describes the 8 primary emotions and the different relationships between them so that they lead to the creation of new emotions.

Plutchik's wheel of emotions can be seen in Figure 1. Each main emotion described in this work is put in opposition with another one, for example joy is the opposite of sadness, fear is the opposite of anger, anticipation is the opposite of surprise, disgust is the opposite of trust and so on.
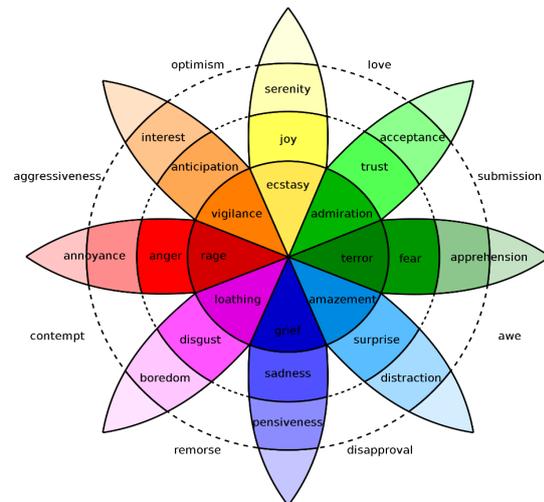
Figure 1. Plutchik's wheel of emotions (Plutchik, 1980b)

Also, all these emotions are arranged on the wheel by intensity (high-intensity emotions are considered those inside and as one moves to the outer layer the intensity of emotions decreases). This is also represented by the color scheme. The colors are darker near the center of the wheel and decrease in intensity as one moves away from the center.

Plutchik's work also states that the combination of two primary feelings creates another feeling, this is expressed in the wheel by colorless emotions. For example, joy and trust create the emotion love and joy and anticipation create the emotion optimism (Plutchik, 1988).

Because of this, the analysis of emotions is considered an extremely difficult process because there are many emotions (starting from the 8 primary and their combinations) with different valences (an emotion can be positive or negative) and different intensities.

3. Literature Review

In the international context there are several lexical databases, especially for the English language but over time they have been adapted for other several languages. One of the first databases, which is also the largest, is Princeton WordNet (Fellbaum, 1998) in which nouns, verbs, adjectives and adverbs are grouped into cognitive synonyms (synsets) each of them expressing a

different concept. This database is public and over the years it has been used and adapted for different applications, for different semantic networks and accessible for several programming languages (.NET, C#, Java, Python, Perl, MySQL etc.). A full list of related projects can be found in (Princeton University, 2021).

Later on, other databases have emerged, using a structure similar to the English version described by Princeton WordNet, but extended to several languages. Examples include EuroWordNet for several European languages like Dutch, Italian, Spanish, German, French, Czech and Estonian (University of Amsterdam, 2021) and BalkaNet (Stamou et al., 2002), for several Central and Eastern European languages like Greek, Turkish, Romanian, Bulgarian, Czech and Serbian.

One of the first and the largest lexicons up to this day is NRC Emotion Lexicon (Mohammad & Turney, 2013) with support for a large variety of languages. This lexicon contains the 8 emotions described by Plutchik and 2 additional attributes (negative and positive). All the annotations were manually done by crowdsourcing. The NRC Emotion Lexicon has the words associated with the emotion they represent and although it was initially developed specifically for the English language, today it includes around 14.000 words and 25.000 associated senses in a multilingual context. Its latest version was translated into 40 different languages but the translations were done automatically and as such there are many errors and words without translation.

Currently, in the Romanian national context, there are not many options that deal with the correlation between words and emotions.

The first approach that tackled the Romanian language was developed by a team of researchers in 2004 and today it is known as RoWordNet (Tufiș et al., 2004), a semantic network which mirrors Princeton WordNet but for the Romanian language.

Another approach is RoEmoLex (Briciu & Lupea, 2017) that uses the Romanian version of the English words from EmoLex (the words are automatically translated) with a series of additional processing to clean the data (deleting words without valences, deleting duplicates, adding tags, parts of speech, translating missing words / manual validation of the translation etc.). In this work, RoWordNet was also employed in order to assign the corresponding synset along with other resources (SentiWordNet scores (Esuli & Fabrizio, 2006), SUMO categories etc.) to create a strong and clean foundation for the Romanian version of EmoLex.

The TEPROLIN platform (Ion, 2018) is based on text preprocessing that incorporates several NLP applications for which it provides a unified access interface as a Python 3 object. TEPROLIN currently offers 15 text preprocessing operations for Romanian, including: text normalization, diacritics restoration, word hyphenation, word stress identification, word phonetic transcription, numeral rewriting.

Ionescu, Demian & Czibula (2017) propose an automated solution for estimating software development effort based on text descriptions of tasks and activities. The case study is based on data from a software company. Text vectorization methods, and also machine learning elements such as term frequency-inverse document frequency (TF-IDF) and distributed representations of documents (doc2vec) were employed, along with regression algorithms. TF-IDF represents the multiplication between the term frequency statistics, which counts how many times a term appears in a document and the frequency statistics of the reverse document.

4. Lexicon Development Stages

Sentiment analysis uses different Natural Language Processing (NLP) methods and algorithms, the most popular ones are:

- Rule-based: it relies on a set of rules defined previously by someone (lexicons, parts of speech, tokenization etc.). These rules are manually done by a person.

- Automatic: systems rely on machine learning methods that can individually learn without human interaction based on a training set and on the generated models. This problem is usually modeled as a classification problem (a sentiment is positive, negative or neutral).

- Hybrid: these applications usually combine the previous two algorithms.

As the proposed work includes different lexicon data sources, it will focus on the rule-based method. By using this method, words and the eight primary emotions are directly associated. The aforementioned method is further parsed and refined, thereby creating a universal JSON format that will be used in all future refinements.

The upcoming versions of this application are meant to bring this implementation to a hybrid method, starting from the elements processed and implemented in this phase (based on the predefined rules and lists) and leading to an automatic system that associates paragraphs with related emotions which will be based on the refined data set in this phase.

4.1 Preprocessing Steps

This application is based on several data sets which contain the associations between words and emotions from several sources, most of them being in the English language. On this basis a series of custom processing algorithms were applied in order to unify and normalize the data sets. A universal data format (JSON) was chosen for the final data set. This type of format combined with the proposed easy-to-understand structure to create an entity which is easy to process and as a result working with data in this format is much faster. This aspect is favorable and necessary because after the unification of the data sets in a single source a series of processing algorithms and methods were applied to refine the data.

Some of these processing steps consisted in:

- a comparison between data sources to check for inconsistencies and differences;

- deleting duplicates and merging similar data;

- checking indices related to emotions and their analysis (what indices are found in each source and in what way - percentage, number, valence etc.);

- combining all the indices of emotions for each word by using the most precise approximation of them (preferably in percentages);

- identifying incomplete entries (words without translation, without valences or associated feelings);

- highlighting words that are not translated so they could be manually processed;

- translation of words that can be translated and elimination of those that cannot be translated into Romanian;

- checking if at this point the resulting data source includes terms / words that have not been associated with any emotion and their elimination / correction;

- elimination and normalization of diacritics and special characters;

- identification of erroneous entries (words that are poorly translated / do not have an exact translation into Romanian);

- expanding the data set (adding specific tags / parts of speech / verb tenses / word derivations etc.).

Based on the operations described above, applied to the initial data sources, the first draft of the initial data set of emotions was obtained.

In this process one can distinguish two major steps in which the employed data set changes radically:

- Data Merge: Merging similar entries and deleting the duplicates that remain – a mandatory step since using multiple data sources as an input meant that there can be duplicated entries/words with different metadata associated with them.

- Invalid Entries Filter – Another important step in which the invalid or incomplete entries were identified and removed from the dataset (one considered invalid entries that did not have a Romanian translation, did not have valid metadata or associated feelings.

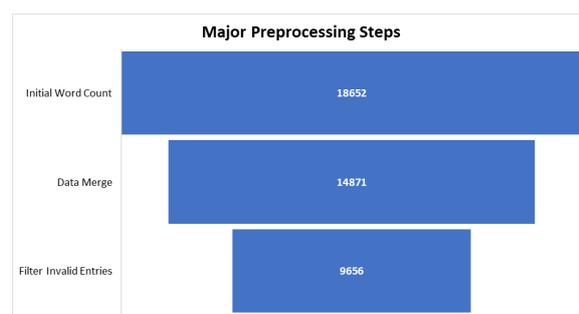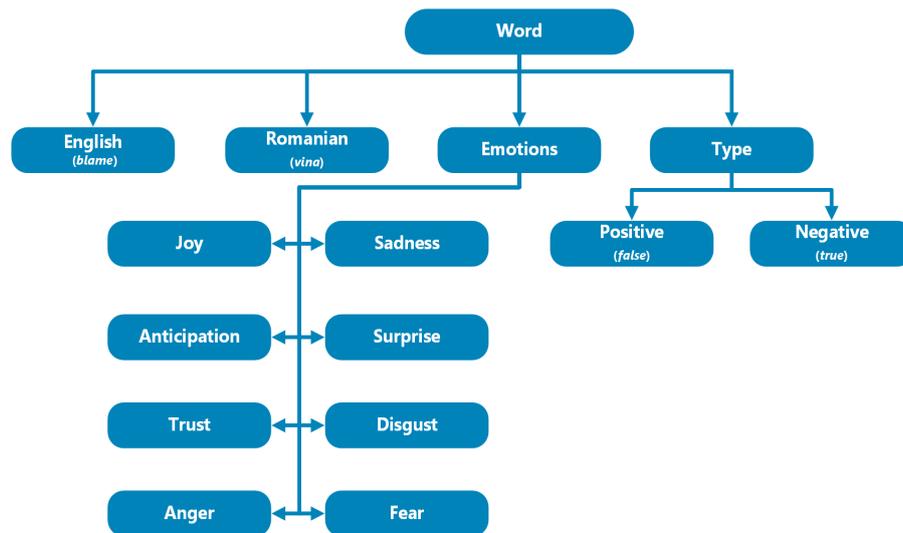The result of the pre-processing operation featuring the steps above is illustrated in Figure 2.

Figure 2. Major Preprocessing Steps

Figure 3. Preprocessing steps

The resulting JSON data is composed of multiple individual structures similar to the one illustrated in Figure 3. The JSON result for the word/entry "blame / vina" was employed as it is shown in the Figure 3.

Since this algorithm is meant to be applied for analyzing emotions in the context of works and literary texts of Romanian canonical authors, it is possible to encounter a series of difficulties in finding a perfect match between the words in those literary works and the employed data set. As they were influenced by the literary currents of those times or even by the geographical region in which they grew up / wrote, many authors used word derivations (different tenses of verbs, adjectives, adverbs, nouns etc.), so there is less chance for finding words in literary works that are exactly in the format stored in the proposed data set.

Due to this aspect, one chose to record for each word different forms of it. For example, in the case of verbs, one took all their forms for all tenses (present, past, future etc.), for all modes (personal / predicative, impersonal / non-predicative), for all conjugations (regular / irregular) etc. This increases the chance of finding that word in the analyzed texts.

This will indeed reduce the employed data set even further but it will improve its utility

as one will have unique terms and the forms mentioned above associated with those unique terms. In the case of the verb "to blame", one will have the unique term in the infinitive form but one will also record the other forms that are encountered under the same entry (for example the present participle – "blaming" or the past tense – "blamed").

One can notice the sentiment/emotion distribution across the dataset after this step (since one made sure that each term had at least one emotion associated with it), while bearing in mind that a word can have multiple emotions associated with it (Figure 4).
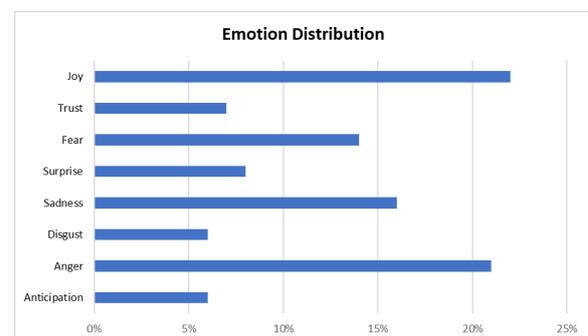
Figure 4. Emotion Distribution

This processing took us to the next step, namely the crawler step.

4.2 Crawler Step

In order to find as many results as possible that would coincide with the proposed data set, it was necessary to enrich it. This was done by creating several crawlers to use on different Romanian web dictionaries in order to expand the data set. Here one can refer to nouns, verbs at different tenses and moods, adjectives, adverbs, different forms and cases etc. because a specific word will rarely be found in the same form in the analyzed texts and in this way, the coverage rate of those words is increased exponentially.

A web crawler is a program that browses the World Wide Web to automatically extract and process certain information according to a specific set of rules (Olston & Najork, 2010). Using the graph structure of the web pages, crawlers act like spiders (as they are also known as web spiders) that cross each node to find information from a start node (a web page originally given). If no data is found on the first page, they automatically access the following links (the children of the current node). In the context of this application, the crawler browses various dictionary sites that were initially declared in order to try and to extract additional information. The starting node for this process is the word itself which means that the program will try to find associations between the words in the proposed dataset and the missing forms from that collection in order to extract specific information (for example grammar, as it is shown in Figure 5).

Each crawler had to be tailored specifically for each dictionary as each of them has specific particularities and a unique format in which it presents data and most of them lack a real API. As a result, raw html is parsed and data is extracted using a specific set of rules. Is the word a noun? What form does it have? Are there any forms of that word in that dictionary that are not in the proposed data set? All of these questions and many more are taken into account when adding logic to the crawlers.

For example, starting from the home page of a dictionary, this will be considered the search entry point for the crawler in that dictionary. The search for a word will take one to a page that will contain in its address the searched word in a similar form like page / search / word.
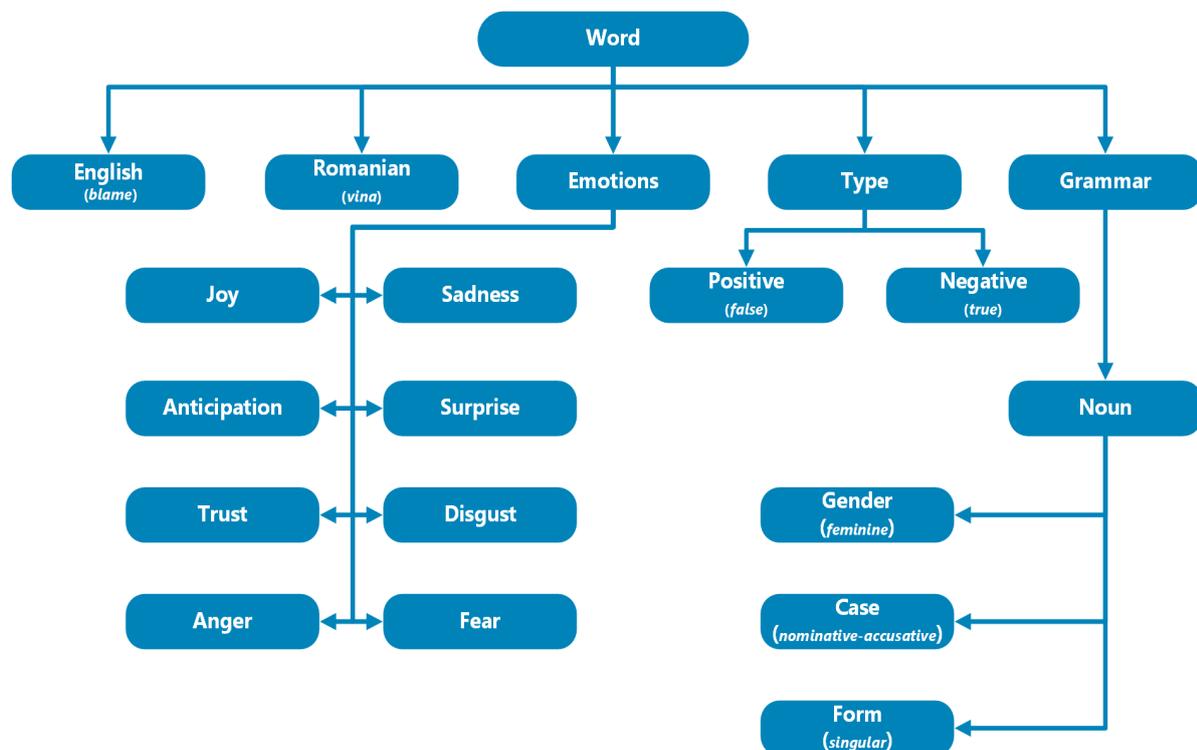

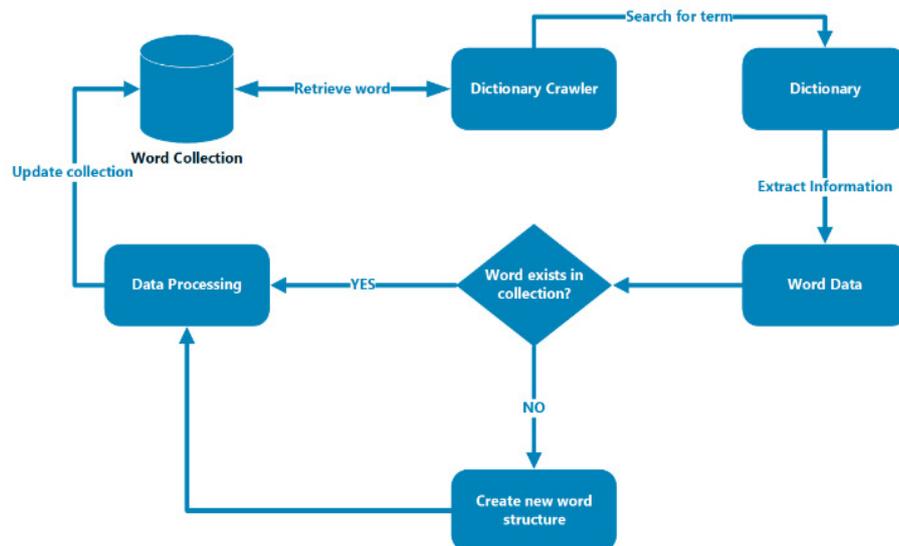**Figure 5**. JSON structure after the crawler step

Figure 6. Web crawling process

As it was mentioned before, as a result of each crawler being tailored specifically for each dictionary, each set of rules was adapted but at the same time it is still based on the same principle. An example of the steps involved in the crawling process is illustrated in Figure 6.

Adding logic and additional rules is an iterative and time-consuming process and as such we expect a great deal of improvements a great deal of improvement should be done in the upcoming versions of this application.

Following the application of the elements described above, the final form for this phase was reached as it is shown in Figure 6 (the case for the word "blame / vina").

The grammar field is dynamically populated depending on the part of speech specific to each word and other attributes.

4.3 Crowdsourcing Data Application

As a future plan, an application could be developed that would help refine the data obtained at this stage and thus increase the accuracy of this data set.

Because the words and the related emotions are extracted from different data sources, the Romanian version of the words is translated automatically, so it raises a series of problems (as it was mentioned in the previous steps). In this context, chances are high that the words in the Romanian language will not express the same feeling (this is also influenced by the fact that the respective word can be a verb, noun, adjective etc. depending on the context).

In the next phase an application will be created which will allow users to view the words in the proposed data set and choose what emotion different words express for him. The system will work similarly to a voting application, thus improving the current data set.

The refinement of the data set obtained in this phase will be done through a crowdsourcing method, in which people will be able to select and associate the words with the respective emotions, an individual and personal response.

This collaborative work will help one to outline a lexicon specific to the Romanian language by combining several methods which will ultimately lead to the creation of an automatic system that will be able to make predictions using the newly created lexicon as a basis.

For example, for the word blame, currently it is labeled as a word with a negative meaning which represents anger, sadness, disgust as the main emotion (according to the 8 primary emotions described in the beginning of this work). Suppose the user chooses 2 or even 3 out of the 8 primary emotions, he can go further to express his opinion about what emotion that

word represents. So, after selecting the main emotions, based on the selected choices, in the next step, a secondary panel will be opened where the user can select the intensity of the chosen emotions (for example for anger it can range from rage which is the strongest form of this emotion to annoyance which has the lowest intensity) or other feelings that are obtained by combining two main emotions according to the wheel of emotions described by Robert Plutchik (by combining the emotions anger and disgust one obtains contempt or by combining disgust and sadness one obtains remorse etc.). This action is optional and will not need to be completed in order to avoid burdening the user with too many steps. The main action will be the word-emotion association by selecting from the 8 main ones.

In the last phase, after refining the current data and expanding the data set, in the application it will be possible to view this database with different filtering options with a variety of terms and tags which will allow the user to explore the whole data set in an easy, fast and accurate way, depending on the selected elements.

4.4 Transition to a Hybrid Application

The last step in the development of the current emotion analysis application is the Hybrid implementation. By combining the rule-based method implemented with the automatic method one can achieve a better association of feelings depending on the context. As it was mentioned before, the automatic method relies on machine learning algorithms to make the association between the emotions and the words. Starting from the data set created and enriched by crowdsourcing methods, correct associations can be made between the two components.

The problems encountered in this process are related to the poetic context. Feelings differ depending on the context thus it is a more challenging task to associate feelings with certain words using machine learning. This process will require more data processing and refinement to reach a result close to the context of the words (there is no correct result when we talk about emotions because everyone perceives them differently).

Neural network approaches (Wadawadagi & Pagi, 2020) have been adopted as a possible solution to problems regarding the classification of sentiments. The most popular methods are: Artificial Neural Network (ANN), Recurrent Neural Networks (RNN), Convolutional Neural Network (CNN), Graph Convolutional Network (GCN) and Graph Recurrent Network (GRN), with all the variations and improvements brought to these types of neural networks.

Neural networks were designed to solve problems related to recognizing patterns, learning things and making decisions. The network resembles the human brain and use neurons to learn and make connections.

All networks are based on almost the same operating principle, they take the data from the previous node as input for the current node. They have a start node (input data), a series of hidden nodes (hidden layer where processing is performed) and the end node (result).

As it was mentioned above, ANNs have 3 layers (input, hidden and output layer) and every layer consists in groups of multiple neurons. The input is processed in one direction, forward. These types of networks are used in problems that require one to learn nonlinear data.

RNNs have a recurrent connection on every hidden node, they are basically ANNs with a recurrent process in the hidden network. The output depends on the previous node but also on the current node due to the loop in every node.

CNNs use convolution filters to recognize patterns with minimal processing. A CNN is made up of several type of layers, each layer having a well-defined role in the learning and recognition process. The CNN layers can be split in two groups: feature extraction and classification.

GCNs use the graph data structure to extract and learn patterns. These types of networks are a generalization of the parent network, more precisely GCN is a type of CNN and GNN is a type of RNN that have emerged as a result of the need to work on graph structures.

In the following steps, the aim is to test several variants of neural networks to see which type

of network is best suited for this particular case but also to evaluate the performance of this type of processing in order to find a possible solution to automatically assign the words to the expressed emotion without the need of human intervention.

5. Conclusion

Most Romanian words cannot be translated exactly from English (or other languages) and are influenced both by the context in which they are applied and also by the area in which they are used (the words may have a different meaning depending on the regional area or in the case of literary works they can be altered by the context and by the literary currents of that time).

This complicates the extraction and the interpretation of emotions from different texts, especially in this case which is focused on literature and many words are used for highlighting different ideas, with a metaphorical meaning to enrich the text.

Starting from different data sources, the words together with their related emotions were processed and combined into a single data source that was the basis of the first phase of the development of this algorithm for analysing emotions. Although refining the data was a great effort and an intensive process, it was still not enough. The chances of finding the same word in the same form in texts is quite small and hence the need to develop the data set in terms of words and their different forms had arisen. This was solved

in the first phase with the help of some advanced crawlers, through which the proposed dataset was completed and refined.

As a plan for the future, the aim is to take this application from the rules-based method to a hybrid / automatic method.

Starting from the results of this phase and from the resulting data set, the intention is to expand more and more this database of emotions by two methods: adding new words with attached emotions (by creating an interface for the application in which different people will be able to vote the association between words and emotions with different percentages - similar to a crowd sourcing action) and automating the process by using techniques specific to natural language processing and machine learning methods (starting from the current data set that will represent the training set to the point where the action of association could be automatically done).

Acknowledgements

REFERENCES

Briciu, A. & Lupea, M. (2017). RoEmoLex - A Romanian Emotion Lexicon, *Studia Universitatis Babeș-Bolyai Informatica*, *62*(2), 45-56.

Esuli, A. & Fabrizio, S. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, (pp. 417-422).

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gavrilă, V., Băjenaru, L. & Dobre, D. (2019a). Towards the Development of an Innovative Platform for the Systematization and Preservation of the Romanian Literary Patrimony, *Transilvanian Review*, *XXVIII* (Supplement 1), 231-244.

Gavrilă, V., Băjenaru, L. & Dobre, D. (2019b). Modern Single Page Application Architecture: A Case Study, *Studies in Informatics and Control*, *28*(2), 231-238. DOI: 10.24846/v28i2y201911

Ion, R. (2018). TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018)*, November 22-23, Iași, Romania (pp. 69-76).

Ionescu, V. S., Demian, H. & Czibula, I. G. (2017). Natural Language Processing and Machine Learning Methods for Software Development Effort Estimation, *Studies in Informatics and Control*, *26*(2), 219-228. DOI: 10.24846/v26i2y201710

Mohammad, S. M. & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence*, *29*(3), 436-465.

Olston, C. & Najork, M. (2010). Web Crawling, *Foundations and Trends in Information Retrieval*, *4*(3), 175-246.

Plutchik, R. (1980a). A general psychoevolutionary theory of emotion. In Plutchik, R. & Kellerman, H. (eds.), *Theories of emotion*, *vol. 1*. New York, Academic Press.

Plutchik, R. (1980b). Emotion: Theory, research, and experience. In Plutchik, R. & Kellerman, H. (eds.), *Theories of emotion*, *vol. 1*. New York, Academic Press.

Plutchik, R. (1988). The Nature of Emotions: Clinical Implications. In Clynes, M. & Panksepp, J. (eds), *Emotions and Psychopathology*. Boston, MA, Springer.

Princeton University (2021). *WordNet - A Lexical Database for English - Related projects*. Available at: <https://wordnet.princeton.edu/related-projects, accessed in 2021>.

Simion, E. (ed.) (2005). *General Dictionary of Romanian Literature*, *vol. 4*. Romanian Academy. Univers Enciclopedic Publishing House.

Simion, E. (ed.) (2006). *General Dictionary of Romanian Literature*, *vol. 5*. Romanian Academy. Univers Enciclopedic Publishing House.

Stamou, S., Oflazer, K., Pala, K., Christodulakis, D., Cristea, D., Tufiș, D., Koeva, S., Totkov, G., Dutoit, D. & Grigoriadou, M. (2002). BALKANET A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the International WordNet Conference*, Mysore, India (pp. 21-25).

Tufiș, D., Barbu, E., Mititelu, V. B., Ion, R. & Bozianu, L. (2004). The Romanian Wordnet Romanian, *Journal on Information Science and Technology*, *7*(2-3), 105-122.

University of Amsterdam (2021). *EuroWordNet Objectives*. Available at: <http://projects.illc.uva.nl/EuroWordNet, accessed in 2021>.

Wadawadagi, R. & Pagi, V. (2020). Sentiment analysis with deep neural networks: comparative study and performance assessment, *Artificial Intelligence Review, 53*, 6155-6195.