Emotion-based Hierarchical Clustering of Romanian Poetry

Mihaiela LUPEA[1], Anamaria BRICIU[1]*, Elena BOSTENARU[2]

[1]Faculty of Mathematics and Computer Science, Babeș-Bolyai University,
1 Mihail Kogălniceanu Street, Cluj-Napoca, 400084, Romania
lupea@cs.ubbcluj.ro, anamaria.briciu@cs.ubbcluj.ro (**Corresponding author*)

[2] Faculty of Letters, Babeș-Bolyai University, 31 Horea Street, Cluj-Napoca, 400202, Romania
elenabostenaru@yahoo.com

Abstract: Emotions play a central role in both writing and understanding literary works, and poetry is a genre rich in emotional content, vivid imagery and abstract language. This paper proposes a clustering-based approach to unsupervisedly mine emotional patterns in Mihai Eminescu's poetry. Lexicon-based emotion features are used for the clustering algorithm. Resulting clusters are assessed with regard to manually added characteristics of poems in the form of literary themes. There is a partial overlap between affective and thematic content, consistent with literary evaluations of the same works. Computational approaches have the advantage of being objective and replicable, with unsupervised techniques such as clustering representing a valuable tool in the exploration of literary works. Nonetheless, no specific emotional patterns, as determined by the proposed method, can be fully associated with particular literary themes.

Keywords: Emotion analysis, Unsupervised learning, Hierarchical clustering, Poetry.

1. Introduction

Sentiment analysis has been an important natural language processing task in the last decade, with applications ranging from appraising sentiment in political speeches to identifying consumer attitudes in product reviews. Lately, researchers have focused on analyzing expressions of emotion in text, thus increasing the number of potential applications in fields like psychology and human-computer interaction. Usually, data sets used for the emotion detection task contain tweets or blog entries. However, there is another research avenue that might provide useful insight into emotion expression, and that is the study of emotions in literary texts. In this context, unsupervised learning offers various models which are effective for mining relevant patterns from text, including emotional patterns.

The use of literary texts as data has its advantages, among which the fact that some literary texts are freely available. Moreover, for popular works, there are vast amounts of literary analyses, critic and audience reviews. Therefore, performance of computational models can be assessed using these additional sources, not unimportant in complex tasks such as emotion studies. In addition, the exploration of emotion in literary texts can be relevant to broader studies of emotion, as literature very often encodes representations of complex emotional experiences, very similar to the ones in the real world (Hogan, 2010).

The purpose of this study is to investigate the association between thematic and emotional content in a corpus of Romanian poems authored by Mihai Eminescu using an unsupervised learning-based analysis. The main research question asked in this work is whether the considered literary topics can be characterized by specific emotional patterns in Mihai Eminescu's poetry. The poetic work of this author was selected because it is a point of reference in Romanian literature, and therefore any findings may be of interest to scholars of different domains. *Hierarchical clustering* is used, with lexicon-based emotional features engineered with the help of RoEmoLex– Romanian Emotion Lexicon (Lupea & Briciu, 2019) to uncover emotional patterns in Mihai Eminescu's poetry in an unsupervised manner. Results obtained provide insight into the emotional patterns associated with thematic content. On the one hand, it is found that while the emotional and thematic perspectives are unquestionably tied together, neither allow a clear separation in homogeneous, disjoint groups. On the other hand, change in the poet's outlook on the proposed topics is characterized by clear shifts in emotional content. The evolution of emotions identified through this computational method is also described by literary critics, which makes it the main finding of the present work.

The novelty of this approach resides in using emotion-based features in a clustering approach

https://doi.org/10.24846/v30i1y202110

for Romanian poetry. To our knowledge, no other computational study involving a joint analysis of emotional and thematic content exists for Romanian literature.

In terms of limitations of the proposed approach, the following aspects that should be taken into account in any interpretation of the presented results are mentioned. The use of lexicon-based features assumes that affective content is, at least to some extent, a surface phenomenon - the assumption is that words carry information that allows inference of emotional states (Klinger, Sulyia & Reiter, 2016). Moreover, poetic texts are characterized by a high potential for interpretation and a lack of strict rules for construction of meaning (Mihăilă, 1972). A lexicon-based approach fails to account for modes of affect expression like figurative language, grammar constructions and acoustics (assonance, meter, rhyme).

The structure of the paper is as follows: first, a brief introduction into the field of emotion analysis and its relevance for literary studies is presented. Next, an overview of related work is included. The third section of this paper describes relevant aspects of Mihai Eminescu's poetry, while the fourth discusses the data and methodology used in the study. The fifth part includes the results of the clustering method and subsequent discussions. Finally, some conclusions are drawn and a number of directions for a future work is proposed.

2. Related Work

The key points of the present work with respect to methodology and computational methods used refer to the use of emotional features derived from Romanian texts in a clustering algorithm to assess emotional-thematic associations. To our knowledge, no previous work incorporates all these aspects in a single study. Therefore, the discussion of related work is separated in three categories: first, a brief overview of existing studies involving clustering and literary data, and poetry, in particular, is presented. Second, some insights into the background of emotional feature use in computational literary studies are offered. Lastly, attention is brought upon some stylometric analyses of Romanian literary works and previous computational studies of Mihai Eminescu's poetry.

Clustering has been an approach used frequently in the study of literary data due to its relative simplicity and interpretability of results. Researchers have used clustering methods with different sets of features for genre classification (Omar, 2020), stylometry and authorship attribution (Luyckx, Daelemans & Vanhoutte, 2006) and thematic analysis (Rahgozar & Inkpen, 2019). Hierarchical clustering was used for Romanian literary data in authorship identification and related tasks (Dinu, Popescu & Dinu, 2008; Dinu, Niculae & Şulea, 2012). Most such clustering approaches use one or more of the following types of features: word or document embeddings, features based on topic modeling, measures of stylistic similarity.

Similarly, the use of emotional features in computational literary studies is widespread, with the early work of (Anderson & McMaster, 1986) which outlines the importance of affect sensing in a text, particularly in a literary context, where results can constitute evidence to be brought to the discussion of standing questions or to formulate new ones. Recently, researchers have used emotion-based features to classify literature pieces into genres (Kim, Padó & Klinger, 2017) and other types of classes. Thematic delineations are less common, presumably because they involve a more challenging multi-label classification setting, and are very rarely approached using emotional features (Barros, Rodriguez & Ortigosa, 2013; Pal & Patel, 2020).

There are various analyses of Romanian literary works (Popescu & Dinu, 2008; Modoc & Gârdan, 2020). For Mihai Eminescu, previous work has focused on other types of features like phonic phenomena and word properties (Popescu et al., 2015) or different perspectives such as differentiation in style in the three phases of Mihai Eminescu's creation (Briciu, 2019).

To summarize, the original aspects of the present work reside in the use of emotional features in conjunction with an unsupervised learning algorithm to identify associations between thematic content and emotional patterns. The current work proposes a new perspective in the interpretation of Mihai Eminescu's poetry based on an objective, replicable computational experiment.

3. Mihai Eminescu's Poetry

Mihai Eminescu is, from many critics' points of view, Romania's national poet. His work draws heavily from Romanian folklore, and the need to create valuable, authentic literature. The innovation brought by Eminescu through poetic language and thematic content makes it difficult to pinpoint the category he belongs to as a poet.

3.1 Topics in Mihai Eminescu's Poetry

In literary terminology, a *theme* or *topic* refers to the general subject of a work. There exists a difference between a literary *theme* and a *motif*, with the latter referring to a smaller visual or contextual unit repeated throughout a work.

While largely belonging to the romantic genre, the multitude of themes and motifs in Mihai Eminescu's work and their interdependence make it difficult to catalogue them in a comprehensive and concise manner. Literary critic G. Călinescu classifies some of the topics under three large contexts (Călinescu & Mihăilă, 1999), namely: *romantic themes* (e.g. the beginning and the end of the world, the moon, the nature, the death as a dream, the madness, the genius' plight, the eternal man, the free love), *psychological subjects* (e.g. the sleep, the dream, the temporal and spatial hallucinations) and *physical worlds* (e.g. the germination, the flora and fauna, the rusticity, the decrepitude, the colossal architectures). Other critics, however, highlight the diversity of symbols in Eminescu's work, outlining the original stylistic valences they acquire, and bring attention to the fact that an attempt to separate a poetry corpus based on thematic content might generate a counterproductive fragmentation of the whole (Manucă, 2008).

In this study, the results obtained in an emotion-based unsupervised learning approach will be compared with thematic groupings proposed by editors and literary critics. The focus is on the separation proposed in literary volumes and thematic anthologies. One such anthology proposes the following (fuzzy) division into three large volumes (thematic threads): *love*, *national* and *philosophical* poems (Zugun, 2002). A subsequent study follows each of these groupings, detailing the evolution of the author's perspective for each thematic thread (Mănucă, 2008).

Within these large groups, tightly interwoven are smaller thematic units, subsumed to the wider concepts of *love, nationality* and *philosophy*. For the present study, the following themes mentioned in (Bărboi, 2008) are chosen:

1. *Time:* an overarching topic, often in the sense of irreversible time.
2. *Cosmogony:* infinity, genesis and end of the universe;
3. *Genius' condition*: the superior being, the impossibility of belonging to the mortal realm;
4. *Nature*: most often intertwined with other themes, depicted by motifs from the flora and fauna category with precise symbolism and emotional connotation; e.g. *cottonwood* - loneliness, *cherry, walnut and apple trees* - childhood, *lilac* - young love;
5. *Death:* death as a dream, a melancholic, pessimistic perspective regarding life in later poems;
6. *Vision about creation and poet's mission:* self-characterizations and meditations regarding the role of a poet and lyrical essays about the concept and goal of poetry.

3.2 Phases of Creation

In a vast and thoughtful critique of Mihai Eminescu's works, Petrescu (2009) separates his work in three phases. The first (1866-1870) is characterized by the perfection of the poetic language and the perspective from which the poet sees himself, that of a *bard*. Confidence in the innovative, visionary language begins to emerge, but between 1870-1872 an inevitable crisis appears in terms of creation. Until 1871, his poetic works are based on antithesis, and he creates worlds of contradictory spaces. Feelings of alienation and distrust permeate the layers of poetic language, which loses the power to create new worlds. The third phase (1881-1883) involves distancing as a form of revolt, the attempt to destroy language and cosmos, and a tragic way of thinking which is often found in the nostalgia of death, and the tranquility therein. Consequently, many complex emotions are generated and interwoven in each phase.

4. Methodology

4.1 Data

The current work focuses on a subset of 131 of Mihai Eminescu's poems collected from an available online source (Wikisource, 2021). This poem set is comprised of poems for which thematic content was explicitly stated in corresponding literary analysis books and anthologies (Bărboi, 2008). The fact that literary analyses customarily focus on widely known, contextually relevant works helped build a representative corpus, with poems from all phases of creation and a diverse thematic content.

Table 1. Dataset statistics

Total number of words	**68451**
Total number of unique lemmas	**8143**
Average poem length (in words)	**526.55**
Type-token ratio (words)	**0.16**
Ratio of unique emotional lemmas	**0.16**
Ratio of emotional lemmas used	**0.18**

Table 1 shows a few descriptive statistics for the considered dataset. A distinction between the percentage of unique lemmas found in RoEmoLex out of all unique instances in the provided texts (0.16) and the percentage of lemmas with emotional content out of all in use (i.e. considering repetitions) is made.

The number of poems per topic is as follows: *time (45), cosmogony (13), genius' condition (9), nature (68), death (32), vision about creation (17).* The most frequently co-occurring themes in the present corpus are *love* and *nature* (29 poems), *time* and *nature* (27 poems), and, to a lesser degree, *time* and *love* (15 poems), *love* and *death* (13 poems), *nature* and *death* (14 poems) and *time* and *death* (13 poems). Literary critics usually catalogue *time* and *nature* as overarching topics in Mihai Eminescu's poetry, often providing context for *love* and *death* moments.

4.2 Data Representation

This approach consists of creating vector representations of poems based on emotional content by computing a score for each valence and emotion as shown in Equation 1, using RoEmoLex (Romanian Emotion Lexicon) (Lupea & Briciu, 2019). RoEmoLex is a resource developed for text-based emotion detection in Romanian language and it contains 8486 single-word terms.

These terms are annotated with eight primary emotions E = {*Anger (**Ang**), Anticipation (**Ant**), Disgust, Fear, Joy, Sadness (**Sa**), Surprise, Trust (**T**)*} and two polarities V = {*Positivity (**P**), Negativity (**N**)*}.

$$score(e, poem) = \frac{1}{l_{poem}} \sum_{t_i \in poem} 1_{t_i \in \mathrm{RoEmoLex}_e} \qquad (1)$$

In Equation 1, *e* represents a valence or emotion from the set $S = E \cup V$, l_{poem} represents the length of the *poem* in content tokens, namely nouns, adjectives, adverbs and verbs, and t_i is a content token in the given poem, where i = 1, l_{poem}. The resulting vectors are 10-dimensional. NLP-Cube (Boroș, Dumitrescu & Burtica, 2018) is used to tokenize and lemmatize the text.

For a more precise analysis of the obtained results, a series of secondary emotions are taken into account as defined by (Plutchik, 1982): ***Love**,* a combination of *Joy* and *Trust,* ***Hope***, involving *Anticipation* and *Trust,* ***Despair**,* formed by *Fear* and *Sadness*.

4.3 Clustering

The algorithm used to cluster these documents is hierarchical agglomerative clustering. This unsupervised technique has been chosen because the present study is largely an exploratory one in that it attempts to find associations between emotional patterns and interwoven thematic content. As such, hierarchical clustering was selected since it outputs an informative hierarchy of clusters that shows how the groupings are formed. Moreover, since it does not require a prespecified number of clusters, it is considered more appropriate for this study than other clustering techniques (Manning, Raghavan & Schütze, 2008).

Hierarchical agglomerative clustering (HAC) is applied. It starts with each document forming its own cluster in the beginning, and then successively merges clusters until there is a single cluster that contains all the documents. There are two choices that have a major influence on the results of this type of clustering: the distance function and linkage criterion. After a series of experiments, the Euclidian distance was set for document similarity and the default Scikit-Learn implementation criterion, the Ward criterion, was used, which minimizes the within-cluster variance as each is successively merged.

The result of a HAC algorithm can be visualized as a *dendrogram*, where instances of the considered corpus can be seen on the x-axis, with vertical lines starting from each point. These lines are joined by horizontal ones that represent the merge of two clusters. The y-coordinate of the horizontal line is the similarity between the two merged clusters. A dendrogram allows the viewing of the history of merges that resulted in the obtained clustering (Manning, Raghavan & Schütze, 2008). The algorithm is applied to the set of 131 poems that are tagged with thematic content, and a truncated history of the results starting at a cutoff of 50 clustezed is visualized.

Lastly, as evaluation metric, the mean Silhouette Coefficient (Rousseeuw, 1987) of all samples is used. This coefficient is an internal evaluation measure computed using the mean intra-cluster distance and the mean nearest-cluster distance for each sample in the corpus. Its value ranges between -1 and 1, with negative values indicating that samples have been assigned to wrong clusters (i.e. a different cluster is more similar), values near 0 indicating the overlapping clusters and 1 being the best value.

The hierarchical clustering approach has the advantage of allowing close inspection of the manner in which clusters are formed, which aids in the interpretation of results on highly ambiguous poetic data.

5. Results and Discussion

5.1 Emotional Vocabulary

Table 2 presents some of the most frequent terms expressing *Fear, Sadness* and *Love*, and terms that were identified as not having explicit emotional content (according to RoEmoLex).

Table 2. Most frequent terms for *Fear, Sadness, Love* and non-emotional categories

	Frequent terms
Fear	*durere/pain, moarte/death, umbră/shadow, dur/tough, pierde/to lose, tremura/ tremble, întuneric/darkness*
Sadness	*negru/black, moarte/death, umbră/shadow, plânge/ to cry, trist/sad, stinge/to extinguish, copil/child*
Love	*dulce/sweet, alb/white, soare/sun, amor/amour, cânta/to sing, blând/gentle, fericire/happiness, dor/ longing, iubi/to love*
No emotional content (as identified by RoEmoLex)	*ochi/eye(s), frunte/forehead, păr/hair, lume/world, suflet/ soul, vis/dream, inimă/heart, mare/sea, albastru/blue, roșu/ red, argint/silver, lumină/ light, lună/moon, vânt/wind, rază/ray, vedea/to see, privi/ to watch, auzi/to hear*

Beyond Mihai Eminescu's explicitly emotional vocabulary, the groups of terms that emerge in the non-emotional category is noted, among them dynamical emotional concepts (*soul, dream, heart*) and literary symbols (*sea, moon, blue, silver*). While an investigation into the emotional component of such terms is outside the scope of this paper, it can be a subject for a further study.

5.2 Emotional Content of Poems

In the set of 131 poems, there are 90 poems with a higher percentage of *Positive* words than *Negative* ones. The highest *Positivity* score (0.35) is registered for *Colinde, colinde!*, a short Christmas poem, which also holds the highest percentage of *Joy* (0.35). At the other end of the emotion spectrum, *O, mama…*, an elegy with *death* as the central subject, is the poem with the highest percentages of *Negativity* (0.29), *Fear* (0.19) and *Sadness* (0.26). In the case of derived (secondary) emotions, *Ce-ți doresc eu ție, dulce Românie*, a patriotic poem, contains most *Love* terms, while, surprisingly, the response to polemics of the times regarding the poet's style, *Eu nu cred nici în Iehova*, has the greatest *Hope* (0.13).

Lastly, valence-emotion percentages are correlated as expected: poems with high *Positivity* scores also have high *Anticipation, Joy* and *Trust* scores, while poems with many terms from the *Negativity* category also contain large numbers of terms from

Anger, Disgust, Fear and *Sadness* categories. As for emotions, the strongest correlations (measured using Pearson coefficient) are: *Anger* and *Disgust* (0.82), *Sadness* and *Fear* (0.71), *Anger* and *Fear* (0.6), *Joy* and *Trust* (0.58), *Anticipation* and *Joy* (0.52), *Anticipation* and *Trust* (0.51).

Computing the mean scores for each emotion across poems that are characterized by a given theme reveals that *love* has somewhat higher percentages of *Joy* and *Anticipation*, in contrast to *death*, where the most prevalent emotion is *Sadness*. *Nature* has the lowest mean score for *Suprise*, while *vision about creation* records the highest mean for *Trust*. However, studying the expression of these emotions throughout the years offers more insight into the variation present within each topic.

5.3 Evolution of Emotions

Figure 1 shows that in the first phase of his creation, Mihai Eminescu's *love* poems are characterized by more *Joy*, which decreases as the years pass, as *Sadness* increases. Mănucă (2008) identifies the year 1879 as the start of the third phase of expression concerning love. He states that, from this point on, *love* means detachment and loneliness for the poet, which is reflected in the maximum percentage of *Sadness* around that time frame.

Similarly, meditations start to have more obvious *death* connotations in 1871, the theme being progressively associated with art, and an escape from reality through it. A critical point is reached in 1879, through the poem *Rugăciunea unui dac,* which Mănucă (2008) qualifies as expressing extreme despair (see Figure 2). *History*, on the other hand, seems to be viewed as a similar mix of *Anger, Disgust* and *Fear* throughout the years, with almost identical trajectories of the three emotions, all registering the highest scores around 1878.

5.4 Clustering Results

The results at a cutoff of 50 clusters were chosen for evaluation. This choice generated groups small enough to be manually assessed, but large enough to draw meaningful conclusions. The obtained Silhouette Index for this clustering is 0.79.

Figure 3 shows the result of the hierarchical clustering algorithm. In the present discussion of the results, three levels of the hierarchy are taken into account: the three large clusters C-X, C-Y, C-Z, then the smaller clusters X1-X4, Y1-Y3 and Z1-Z4, and finally, some individual clusters from the bottom of the hierarchy. Figure 4 illustrates the mean percentages for each of the large clusters C-X, C-Y, C-Z, for a set of valences (*Positivity, Negativity*), primary emotions (*Anger, Disgust, Fear, Joy, Sadness, Trust*) and secondary emotions (*Hope, Love, Despair*).

Figure 4 shows that poems in C-X are emotional across the board, having high scores in all emotional categories, and, on average, a higher number of *Negative, Anger, Fear, Sadness, Despair* terms than poems in other clusters. An overwhelming majority of the poems are from the later phases of the author's work (95%), which are canonically characterized by darker moods.

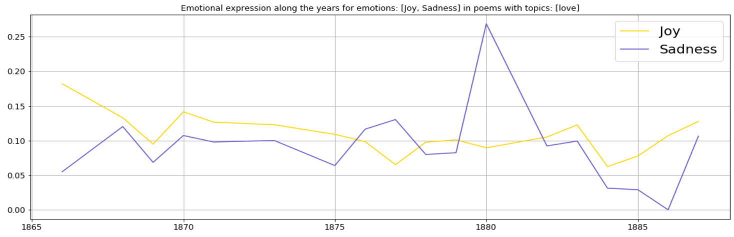

Figure 1. Emotional expression along the years for emotions *Joy* and *Sadness* in poems with topic *love*

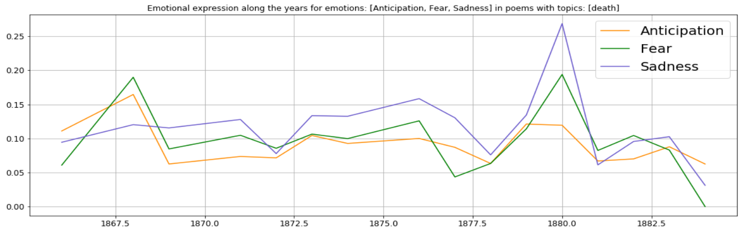

Figure 2. Emotional expression along the years for emotions *Anticipation, Fear* and *Sadness* in poems with topic *death*

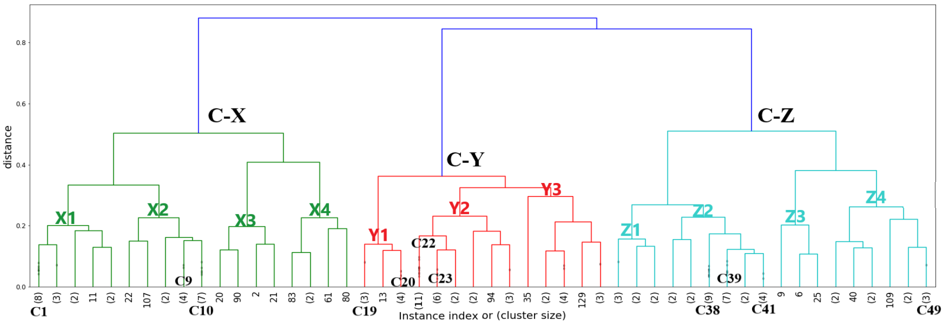

Figure 3. Dendrogram: Hierarchical agglomerative clustering with emotion-based features

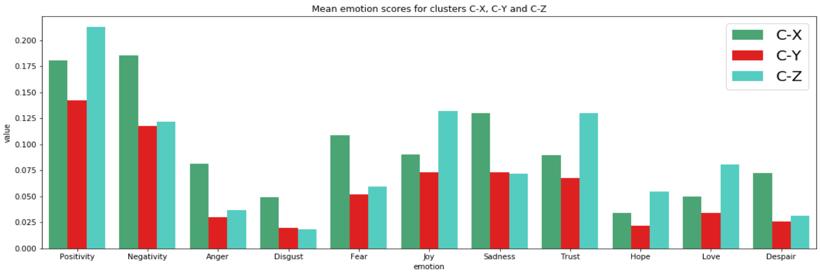

Figure 4. Mean scores for selected emotions for clusters C-X, C-Y, C-Z

The predominant theme in this cluster is *death* (in 42.5% of the poems), followed by *love* and *nature* in equal part (32.5%).

However, not all the clusters contained within C-X are predominantly *Negative* - only X2 and X4. The similarity between X1 and X2 resides in the degree to which the poems are emotional: poems in X1 and X2 are less emotional than those in clusters X3 and X4. As for X1 and X2, they contain longer poems, which also talk about *history* and *cosmogony* in addition to *love* and *death,* as opposed to X3 and X4, which are, for the most part, characterized only by the latter.

In terms of emotions, the common thread of all the four clusters is the similar *Sadness* score, while the biggest difference is in the *Fear* scores in X3 and X4. Looking at the most frequent words in each of these clusters offers a surprisingly suggestive summary of the contained topics and indications about emotional content: *lume/world, a vedea/to see, ochi/eye, vis/dream, trece/to pass, suflet/soul, stea/star, umbră/shadow, pământ/earth, moarte/death, negru/black* (X1, X2) and *a muri/to die, mereu/always, moarte/death, a crede/to believe, a stinge/to die down, dulce/sweet, durere/pain, a iubi/to love, copil/child, gura/mouth* (X3, X4).

At the bottom level of the hierarchy, the focus is on clusters C1, C9, C10 (see Table 3 for a list of cluster members). First, the emotional and semantic ambivalence of C1 is noted, which has a

Table 3. Mean percentages for emotional feature values in selected clusters

		P	N	Ang	Ant	Fear	Joy	Sa	T
C1	*Demonism, Ghazel, Mortua est!, Rugăciunea unui dac, Te duci..., Venere și Madonă, Împărat și proletar, Înger și demon*	0.18	0.17	0.08	0.09	0.10	0.10	0.12	0.09
C9	*Melancolie, Oricâte stele, Peste vârfuri, În liră-mi geme și suspin-un cânt*	*0.14*	*0.19*	*0.05*	*0.09*	*0.08*	*0.07*	*0.13*	*0.09*
C10	*Despărțire, Gemenii, Pe aceeași ulicioară, Renunțare, Scrisoarea IV, Stelele-n cer, Strigoii*	*0.14*	*0.15*	*0.05*	*0.07*	*0.10*	*0.06*	*0.10*	*0.07*
C19	*La mijloc de codru, Nu mă-nțelegi, Stam în fereastra susă*	*0.17*	*0.04*	*0*	*0.06*	*0.01*	*0.12*	*0.02*	*0.04*
C39	*Atât de fragedă, Făt-Frumos din tei, La Bucovina, O călărire în zori, O, rămâi..., Povestea teiului, Sonete*	*0.19*	*0.11*	*0.03*	*0.11*	*0.06*	*0.12*	*0.07*	*0.12*
C41	*Din străinătate, Freamăt de codru, Lasă-ți lumea ta uitată, Noaptea*	*0.19*	*0.10*	*0.01*	*0.09*	*0.04*	*0.14*	*0.06*	*0.1*
C49	*Când amintirile, De-or trece anii, Pe lângă plopii fără soț*	*0.24*	*0.09*	*0*	*0.15*	*0.04*	*0.14*	*0.06*	*0.16*

similar mean percentage of *Positive* and *Negative* terms, as well as *Joy* and *Sadness* words. From a semantic point of view, many of the poems involve a positive-negative opposition: *Înger și demon* („*Angel and demon*"), *Împărat și proletar* („*King and proletarian*"). Conversely, poems in C9 are more evidently *Negative*, with a high percentage of *Sadness terms*, and all are written between 1876 and 1882. These poems are melancholic, of derealization, detachment and loneliness.

The dramatism of expressed feelings increases in C10, which contains poems where peace is sought in protective deity (*Gemenii*), in idealized historical figures, much like those in *Rugăciunea unui dac* (C1). Most poems in cluster C-X are philosophical in nature, with the themes of *history*, *death* and *love* inextricably intertwined. A better clustering might be obtained if a finer distinction between emotional states is made. As it stands, the *Sadness* generated by the experience of love that is not fulfilled is equated with the *Sadness* caused by the imminence of death.

Cluster C-Y is the least emotional, overall, with only a slightly higher average score for *Disgust* and *Sadness* than C-Z. Thematically, this translates to a dominance of the *nature* (75% of the poems) and *time* (40%) topics. Component clusters Y1, Y2 and Y3 differ in the *nature* context they depict, and consequently in the associations with other themes.

Cluster Y1 mainly contains poems with *nature* and *time* topics, and is, as far as tone is concerned, more *Positive* than *Negative.* The most frequent words in this cluster are *lună/moon, azi/today, trece/to pass, senin/cloudless, noapte/night, întreg/whole.* In contrast, Y2 is made up of poems with a balanced perspective, characterized by *nature, time, history* and *death* topic mixtures. Term presence in this cluster also follows the idea of balance, words with *positive* connotations (*alb/white, soare/sun)* and *negative* nuances (*lung/long, lună/moon, umbră/shadow*) occurring equally frequently.

Poems in Y3 convey a more *negative* tone through the idea of *falling,* and the natural context created by terms such as *munte/mountain, val/wave, suna/to sound, codru/forest*. This cluster is made up of poems about *love* and *history*, both set against a *nature* background. Many of the *nature* poems in this cluster are, in fact, meditations, „dialogues" of inner voices in idyllic settings, compensatory universes sometimes obtained through *love* and sometimes through *historical, legend-like* imaginings, with *death* inevitably looming over both.

Finally, the third large cluster, C-Z, is less emotional than C-X, and dominantly *Positive*, with the highest average scores for *Anticipation, Joy, Trust, Hope* and *Love* emotional categories. Topic-wise, *nature* and *love* are predominant (47% of the poems).

Time and *vision about creation* also register a notable presence: 30%, and 20%, respectively. From a temporal perspective, C-Z is the cluster that contains most poems written during the first stage of the author's creation (6 poems vs. 7 poems total in C-X and C-Y). All component clusters (Z1, Z2, Z3, Z4) have a higher average percentage of *Positive* terms in comparison with *Negative* ones. The small cluster Z3, made of only 3 poems (*"Ce-ți doresc eu ție, dulce Românie!", "Colinde, colinde"*, *"De-aș avea"*) contains the poems with the highest scores for *Joy* and *Love* emotional categories in the corpus.

Z4 is the subcluster with the closest values to Z3 for these two emotions, though the percentages only reach approximately half of the value obtained for Z3 (e.g. 0.27 (Z3) vs. 0.14 (Z4) for *Joy*). Z4 contains poems about *unrequited* or *unfulfilled love*, also having the highest percentage of *Sadness* and *Despair* terms in C-Z.

The other two subclusters of C-Z, namely Z1 and Z2, are emotionally similar, with close scores in *Joy, Hope* and *Love* categories. Z1 contains poems with more *Trust* terms than Z2. In contrast, Z2 has high *Sadness* and *Despair* emotional percentages. Thematically, Z1 is a cluster that contains many poems tagged with a single theme. The average number of topics per poem in subcluster Z1 is 1.28, while in Z2 poems are a combination of at least 2 topics (average: 2.26). The point of similarity resides in the *love, nature, time* and *vision about creation* topics, with difference in semantic nuances for the concept of time, often associated with *cosmogony* and *genius' plight* topics in Z2.

The overlap between subsets of most frequent terms in each of these subclusters appears to approximate the emotional and thematic overlap between them. The common words between clusters are *lume/world* (Z1, Z2, Z4)*, vis/dream* (Z2, Z3)*, dulce/sweet* (Z3, Z4), *eye/ochi* (Z2, Z4), *a trece/to pass* (Z2, Z4). In particular, the most frequent words in Z1 poems are *a veni/to arrive*, *vânt/wind*, *a cere/to* *ask, poveste/story, fată/girl* and *a atinge/to touch*, while in Z2 we find *a vedea/to see, suflet/soul, frumos/beautiful, inimă/heart*. Z3 has the most specific vocabulary, a heterogeneous semantic mix between frequently occurring words in the 3 contained poems. Arguably, the vocabulary slice obtained for Z4 is the most suggestive: *durere/pain, a ști/to know, a uita/to forget, speranță/hope, dor/longing, noapte/night*.

Lastly, a thematically homogeneous cluster, namely C49, is highlighted. All three works in this group are tagged with the topic *time*. This cluster has the highest mean percentage of *Anticipation* and *Trust* term occurrence.

5.5 Validation from a Literary Perspective

The thematic content identified by the proposed unsupervised approach, based on emotional features, is in agreement with critics' appraisal.

First, the results obtained with respect to tracking expressed emotions across the years are highlighted. Figures 1 and 2 show an evolution of emotion in poems with certain topics that is also described in (Mănucă, 2008). Similarly, the clustering results can also be interpreted in the context of Mihai Eminescu's phases of creation.

It has been shown that first-phase poems cluster together mostly in groups with high *Positivity* and *Joy* percentages and poems from later phases are grouped in clusters with *Negative* tendencies, the emotional content also being signaled in literary discussions (Mănucă, 2008; Petrescu, 2009).

A clear association between specific thematic content and emotional patterns was not found, only a confirmation of general emotional tendencies. For instance, poems with "negative" topics such as *death* have higher percentages of *Negativity* terms and associated emotions *Sadness, Fear, Anger*. "Neutral" topics, on the other hand, like *nature* and *time,* are shown to have less explicit emotional content. Finally, poems that propose semantic oppositions (e.g. angel vs. demon) also have ambivalent emotional content.

However, considering the multitude of emotions and nuances of emotions in a literary text, it is fair to say that a topic can be characterized only by emotional pluralism, frequently involving contradictory content. Consequently, even small differences between valence and emotion scores become crucial for a correct interpretation of the results. The topic of *love*, for example, is multi-faceted, with overwhelming *Negativity* associated with it in later phases of creation, as opposed to the *Positive, Joyous,* optimistic outlook in the first phase. Therefore, the definition of new topics that incorporate both emotional and semantic aspects is proposed. They could be defined by the frequent terms enumerated for the corresponding clusters. For instance, types of such *love* topics could be *unfulfilled love (disheartening)*; *adolescent love and desire (optimistic, hopeful)*; *fantastic, idyllic love (nostalgic)*.

Taking into account the number of poems considered, and the comprehensiveness that characterizes RoEmoLex, it can be concluded that from a literary perspective, the study demonstrates that Mihai Eminescu's poetic language follows an internal logic of emotions, and that the mix of emotions in his poetry can be organized to generate an accurate assessment of a predominant state of mind and thematic content.

6. Conclusion

This paper presents an emotion-based analysis of Mihai Eminescu's work, based on a corpus of 131 selected poems. The proposed set is a representative one, including poems from all of Eminescu's phases of creation, and diverse thematic content.

The technique used was hierarchical clustering with lexicon-based emotion features obtained using RoEmoLex. The original contribution of this paper is the application of an unsupervised machine learning technique to a poetry corpus in order to explore associations between thematic content and emotional patterns. The present results show that some of literary scholars' findings can be replicated using a computational technique, which can also provide new perspectives in viewing the proposed subject (e.g. definition of new topics).

The drawbacks of this approach refer to the flat view of emotions, which has a reduced power of representation. In a further work, an emotion intensity scale or a breakdown of current emotion labels into finer-grained ones (e.g. *Sadness*: hurt, suffering, hopelessness, melancholy, grief, regret, displeasure, dejection) should be employed. Similarly, taking into account a trajectory of emotion across a poem may yield better results.

REFERENCES

Anderson, C. W. & McMaster, G. E. (1986). Modeling emotional tone in stories using tension levels and categorical states, *Computers and Humanities, 20*(1), 3-9.

Barros, L., Rodriguez, P. & Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A Study on Quevedo's poetry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 141-146).

Bărboi, C. (2008). *Mihai Eminescu - O lume dăruită nouă: sinteze, comentarii, aprecieri critice, texte adnotate*. Editura Universitară.

Boroș, T., Dumitrescu, Ș. D. & Burtica, R. (2018). NLP-Cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 171-179).

Briciu, A. (2019). Quantitative analysis of style in Mihai Eminescu's poetry, *Studia Universitatis Babeș-Bolyai Informatica*, *64*(2), 80-95.

Călinescu, G. & Mihăilă, I. (1999). *Opera lui Mihai Eminescu*. Editura Academiei Române.

Dinu, L.P., Niculae, V. & Șulea, O. (2012). Pastiche detection based on stopword rankings. Exposing impersonators of a Romanian writer. In *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection* (pp. 72-77).

Dinu, L.P., Popescu, M. & Dinu, A. (2008). Authorship identification of Romanian texts with controversial paternity. In *Proceedings LREC 2008* (pp. 3392-3397).

Hogan, P. C. (2010). Fiction and feelings: On the place of literature in the study of emotion, *Emotion Review*, 2(2), 184-195.

Kim, E., Padó, S. & Klinger, R. (2017). Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 17-26).

Klinger, R., Suliya, S. S. & Reiter, N. (2016). Automatic emotion detection for quantitative literary studies: A case study based on Franz Kafka's "Das Schloss" and "Amerika", *Digital Humanities 2016: Conference Abstracts,* 826-828.

Lupea, M. & Briciu, A. (2019). Studying emotions in Romanian words using Formal Concept Analysis, *Computer Speech & Language, 57*, 128-145.

Luyckx, K., Daelemans, W. & Vanhoutte, E. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06),* (pp. 30-35).

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mănucă, D. (2008). *Oglinzi paralele*. Editura EuroPress.

Mihăilă, E. (1972). Modele matematice de analiză a operei literare, *Analiză și interpretare. Orientări în critica literară contemporană*. Editura Științifică.

Modoc, E. & Gârdan, D. (2020). Style at the scale of the canon. A stylometric analysis of 100 Romanian novels published between 1920 and 1940, *Metacritic Journal for Comparative Studies and Theory, 6,* 48-63.

Omar, A. (2020). Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods, *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(2), 99-107.

Pal, K. & Patel, B.V. (2020). Model for classification of poems in Hindi language based on Ras, *Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, 141,* 655-661.

Petrescu, I. E. (2009). *Studii eminesciene*. Casa Cărții de Știință.

Plutchik, R. (1982). A psychoevolutionary theory of emotions, *Social Science Information*, *21*(4-5), 529-553.

Popescu, M. & Dinu, L.P. (2008). Rank distance as a stylistic similarity. In *Proceedings of the 22nd International Conference on Computational Linguistics COLING 2008* (pp. 91-94).

Popescu, I. I., Lupea, M. A., Tătar, D. & Altmann, G. (2015). *Quantitative analysis of poetic texts.* De Gruyter Mouton.

Rahgozar, A. & Inkpen, D. (2019). Semantics and Homothetic Clustering of Hafez Poetry. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 82-90).

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Computational and Applied Mathematics, 20*, 53-65.

Wikisource (2021). *Autor: Mihai Eminescu*. Last accessed: 21 March 2021, available at: <https://ro.wikisource.org/wiki/Autor: Mihai_Eminescu>.

Zugun, P. (2002). *Mihai Eminescu. Opera poetică (3 volumes*), Editura Tehnopress.