Computational Approaches to the Analysis of Human Creativity

Fabio Celli

Abstract In this paper we address the issue of creativity and style computation from a Natural Language Processing perspective. We introduce a computational framework for creativity analysis with two approaches, one agnostic, based on clustering, and one knowlegde-based, that exploits supervised learning and feature selection. While in the agnostic approach can reveal the uniqueness of authors in a meaningful context, the knowledge-based approach can be exploited to extract the culturally relevant features of works and to predict social acceptance. In both the approaches it is required a great effort to define symbols to represent meaningful cues in creativity and style.

1 Introduction and Related Work

Understanding human creativity is a long-standing issue that poses great challenges to many disciplines, humanities above all, but also psychology and computer science. On the humanities side, the flow theory [5] [6] defined creativity as the process by which a symbolic domain in the culture is changed. This theory emphasizes three elements:

1) a culture that defines rules;
2) an entity, person or event, who brings novelty into the rules;
3) a set of people that recognize, accept and validate the innovation.

From the psychological side, the attempts to develop a creativity quotient test similar to the intelligence quotient (IQ) have been unsuccessful [13]. Guilford [8] and Torrance [25] developed tests of Creative Thinking from the 1960s to the 1980s, scoring the uniqueness of subjects in different tasks. These tests have been strongly criticized in recent years [18], with the argument that a person's creativity can only be assessed indirectly, for example with observer ratings [11]. This method has the

Fabio Celli
University of Trento, e-mail: fabio.celli@unitn.it

great problem of the agreement between raters and the limited context that can be analyzed. One of the self assessments available is the Creativity achievement questionaire [4], that is based on the idea that a person's creativity can be measured by the person's achievements recognized by the society in different fields of creativity, from art and music to business. In computer science, most of the effort in this field has been done by artificial creativity, that focuses on social aspects. Based on the dual generate-and-test model of creativity [14], that sees the creative process as the interaction of individuals with their social environment, artificial creativity promotes the study of the creative behaviour of individuals in societies by means of networks of agents, whose parameters and interactions can be observed in a controlled setting [23]. Other computational approaches that focus more on the creative process include logic and machine learning. Wiggins [26] proposed a finite-state language to formalize creativity as a set of functions that generate new elements from the existing ones, while Barbieri [2] successfully exploited a constrained markov process to generate lyrics in the style of an existing author.

We summarize the different approaches and views of creativity in four types:

1) **nomothetic versus idiographic view**. The nomothetic approach focuses on universals of creativity while the idiographic focuses on individual differences and uniqueness of creative individuals [8], [21].

2) **Mental versus social phenomenon**. Creativity is seen as a mental or psychological phenomenon for the generation of new ideas or as the result of an audience's appreciation of a novel idea [23].

3) **Improbabilist versus impossibilist approach**. The improbabilist approaches see creativity as a novel combination of familiar ideas while the impossibilists consider creativity as new ideas never appeared before. This is possible thanks to a transformation of the conceptual space [3].

4) **Rational versus irrational process**. The rational approaches include logic [19] and evolution. Logic sees creativity in processes like abduction (guessing a conclusion given some premises), eduction (relate properties of different objects), deduction and induction (relate properties from general to specific and viceversa). Evolutionary approaches instead see creativity as an adaptation to the environment motivated by problem solving [15], [20] while the irrational approach focuses on the emotional part of creativity [1].

In this paper we address the issue of computational Human Creativity Analysis (HCA) from a Natural Language Processing (NLP) perspective. We define the framework for exploiting NLP techniques for HCA, focusing on styles and their relationships with creativity. In particular, we identify two methods to tackle the issue of HCA from two different sides: the first one is social, idiographic, rational, and improbabilist (we will call this one "agnostic"); the second one is mental, nomothetic, and irrational (we will call the approach to this point of view "knowledge-based"). Based on two different NLP techniques, we develop a framework for HCA, describing the challenges and potential problems, both from a theorethical and practical point of view. The main contributions of this paper to the research community are: 1) the definition of a framework for computational HCA from a quite new perspective and 2) the application of NLP techniques suitable to address HCA tasks.

The paper is structured as follows: in section 2 we define the notions of style and creativity, and we introduce the framework for HCA. In section 3 we will run some experiments with the discussion of the results and in section 4 we will draw some conclusions.

2 The HCA Framework

Our framework for HCA has two different approaches: one agnostic, that measures creativity as uniqueness and one knowledge-based, that measures social acceptance. The agnostic approach adopts unsupervised learning [7], in particular simple K-

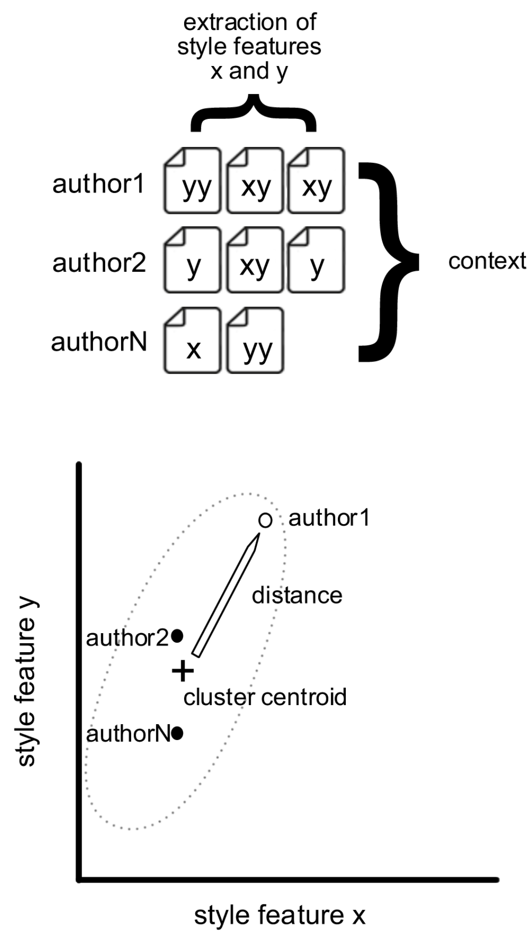

Fig. 1 Representation of the agnostic approach to HCA. Style/feature extraction and comparison between authors are turned into a clustering task. Style is extracted as a set of features from many documents of the same author, then a feature space is used for the inter-author comparison in a clustering task. We define the distance between a target author and the cluster centroid as a measure of creativity.

means algorithms [10]; while the knowledge-based approach exploits supervised learning [12], feature selection methods [17] [22] and annotation, that can be manual or automated by means of knowledge bases.

The basic assumption of the agnostic approach is that styles and creativity can be defined by the contexts where they appear. In order to turn HCA into a learning problem, we define **style** as a set of features that can be extracted from documents of an author, and represented as sequences of symbols, such as bare words, rhyme schemas, notes etc. Creativity is strongly related to style: we define **creativity (uniqueness)** as the distance between a target author and its context in a style feature space, as in figure 1.

In this way, creativity can be computed with simple k-means clustering, that is generally a very fast technique and do not require annotation, but just a post-hoc evaluation. Turning to Csikszentmihaly's theoretical framework mentioned in section 1, what we call context is a representation of the culture that defines rules. In this agnostic approach, creativity is seen as a measure of the novelty or uniqueness of the style of an author in a context. The agnostic approach to HCA addresses points 1 and 2 of the Flow theory:

1) a culture that defines rules, represented as the centroid of the cluster;

2) a person or event who brings novelty into the rules, represented as the points in the space (authors). The distance between each point and the centroid (accepted culture) is the measure of the novelty of that author with respect to the contextual culture.

In the HCA framework, the third point of the Flow theory - the issue of the set of people that recognize, accept and validate the innovation - can be tackled with the knowledge-based approach by means of feature selection and supervised learning. If we link an appreciation score, retrieved from knowledge bases, to an author's

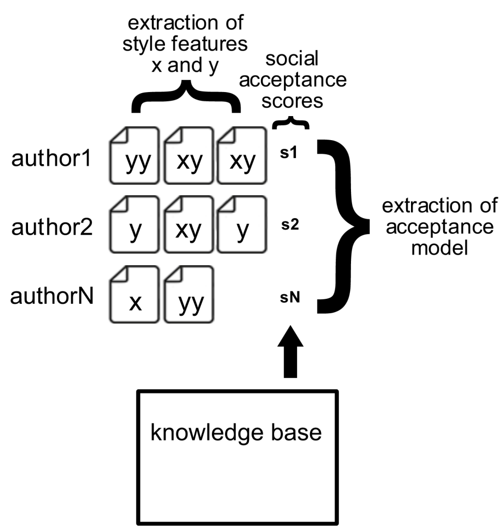

Fig. 2 Representation of knowledge-based approach to HCA. Style features are extracted from documents of one or more author(s), and linked to appreciation scores extracted from knowledge bases. Acceptance models, retrieved by means of feature selection methods and supervised machine learning techniques, can be used to predict appreciation scores.

style in a context, we can use it as the target class and predict appreciation from the space of style features by means of supervised algorithms. These algorithms can provide acceptance models, that link style features to the appreciation scores. In this approach **acceptance** is a function that maps a style feature vector to appreciation scores, that can be extracted from knowledge bases, see figure 2.
The heuristic power of our HCA framework depends on the control of two factors:
1) the definition of a meaningful context;
2) the definition of meaningful symbols to represent styles.
For this reason, corpus collection is very important. The context collected must be homogeneous, in order to represent the rules defined by the culture as the cluster centroid. For example one can compare authors in the same timespan: in that case the cluster centroid represents the general style, or "taste" of the time. Another important aspect is data representation: we can have as input just simple text or more abstract levels of representations, that can be annotated by hand or with ad-hoc algorithms. Language is one clear example of level representation: text can be represented at the level of bare words, part-of-speech, syntactic chunks etc. In the case of style and creativity extraction, formalization is an open issue. For example it is possible to formalize poetry either with rhyme schemas and bare words, or songs as sequences of notes, words, beats-per-minutes, arrangement structures and emotions [16]. In the next section we will present two sample experiments, one of the agnostic approach and one of the knowledge based.

3 Experiments

We ran two experiments in Weka [27] in order to provide examples of creativity analysis. We collected a corpus of titles of albums and songs of 24 bands formed between 1950s ans 1970s (our context). The corpus contains text and dates for a total of about 19500 tokens. We used the occurrence of words associated to each band as features. We removed stop-words, like prepositions and articles, and built a style feature space of more than 1000 dimensions, where each dimension is the frequency of a word. We limited the words used as features to 1000. We selected to sample the data of only 24 bands in order to have a small and controlled experimental setting.
First we tested the knowledge-based approach. Due to the fact that we do not have Knowledge resources designed for creativity analysis, we exploited Wikipedia, extracting the length of the page (in characters) for each band as a measure of social acceptance, based on the idea that the importance of each band corresponds to the quantity of text stored in the collective memory, represented by Wikipedia. In this experiment we tested how social acceptance can be predicted from the style feature space, made of words of album and song titles. To predict social acceptance scores, we used a regression algorithm based on Support Vector Machines (SVM-reg) [24]. With this algorithm, we can compute a formula that predicts social acceptance scores from the style feature space and evaluates the error in the prediction. In practice, the algorithm assign a numerical weight to the words, based on their abil-

Table 1 Results of the experiments. We used a regression based on Support Vector Machines (SVMreg) as classification algorithm, and we tested two different settings: with and without feature selection. As feature selection we used a correlation-based algorithm that evaluates the worth each attribute considering its individual predictive ability along with its degree of redundancy between the selected ones.

algorithm	**features**	**RMSE**
Baseline	1000	0.192
SVMreg	1000	0.211
algorithm	**features**	**RMSE**
Baseline	26	0.192
SVMreg	26	0.087

ity to predict social acceptance score, and compare the prediction with actual social acceptance score values, reporting the error. To evaluate the error, we used Root Mean Squared Error (also called Root Mean Standard Deviation) that represents the sample standard deviation of the differences between predicted values and observed values.

We normalized the social acceptance scores and the occurrences of words, in order to obtain values between 0 and 1. Then we applied the "BestFirst" feature selection algorithm [9], that evaluates the correlations between words and social acceptance

```
 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23   <-- assigned to cluster
 0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0 | beatles
 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | rollingstones
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0 | ramones
 0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 | loureed
 0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | velvetunderground
 0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | johnlennon
 0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | ringostarr
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1 | mccartney
 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | joanbaez
 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | janisjoplin
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0 | ledzeppelin
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0 | doors
 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | bobdylan
 0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | elvis
 0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0 | beegees
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0 | stooges
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0 | iggypop
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0 | pinkfloyd
 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | rogerwaters
 0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0 | sydbarrett
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0 | beachboys
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0 | buddyholly
 0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | hendrix
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0 | thewho
```

Fig. 3 Results of the agnostic approach to HCA. We set the number of clusters we want as the same number of instances we have, and test whether authors are clustered alone or not. In this case yes.

scores. This feature selection algorithm is a dimensionality reduction technique, that explores the style feature space and considers the individual predictive ability of each word and its degree of redundancy, selecting only the most relevant ones. In the experiment we tested a setting with all 1000 features and a dimensionality reduced space with 26 features. We split the data in two parts: 66% of the data for the training phase and and 33% for testing and evaluation. We compared our results to a baseline obtained predicting the acceptance scores with the average value computed on the training dataset. The results, obtained with the majority baseline and with the regression algorithm (with and without feature selection) are reported in table 1.
These results and reveal that, despite the baseline obtained a very low error, the regression algorithm can reduce it a lot when we apply feature selection. This experiment reveals that, given words in song titles, we can automatically predict the social acceptance with an error of 0.087. But the most interesting thing is that with feature selection we can find the most relevant words and their weights in the predicton of acceptance scores, reported below and divided between positive and negative.

```
+       0.0396 * Along
+       0.0396 * Comin
+       0.0353 * Coming
+       0.0348 * Got
+       0.0479 * Harder
+       0.0451 * Hey
+       0.0806 * Let
+       0.0182 * Lose
+       0.0343 * Luck
+       0.0928 * Me
+       0.1067 * Must
+       0.0635 * New
+       0.0519 * Sand
+       0.0962 * Yes
+       0.1272 * Yet
+       0.099  * Yourself
-       0.0653 * Death
-       0.0038 * Here
-       0.0295 * Flaming
-       0.0165 * Tight
-       0.0129 * Two
-       0.0743 * Makes
-       0.0872 * Weirdness
```

In a second stage we tested the agnostic approach: We use a simple K-means clustering algorithm with euclidean distance, evaluating the clusters on the band's names. In order to assess the creativity/uniqueness of an author, we can set the number of clusters we want as the same number of instances we have, and test whether authors are clustered alone or not. The authors that are not clustered alone are less unique than others. In our case, as reported in figure 3, all the authors have their own cluster, meaning that they are all unique. If we want to understand which one is the most distant from the centroid, we have to set 2 clusters. In our case it resulted that Joan Baez is the most unique author in terms of words used in song titles and lyrics, with respect to the other bands.

4 Conclusions

In this paper, we presented a computational framework for HCA inspired by computational linguistics and NLP. This framework has two approaches, one agnostic, based on clustering, and one knowlegde-based, based on supervised learning and feature selection. While in the agnostic approach can reveal the uniqueness of authors from their work, the knowledge-based approach can be exploited to extract the most accepted features of works and to predict social acceptance. In both the approaches it is required a great effort to define symbols to represent meaningful cues in creativity and style.
In the future we would like to do new experiments of HCA on a very large scale.

References

1. James R Averill. Creativity in the domain of emotion. *Handbook of cognition and emotion*, pages 765–782, 1999.
2. Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. Markov constraints for generating lyrics with style. In *ECAI*, pages 115–120, 2012.
3. Margaret A Boden. Creativity: a framework for research. *Behavioral and Brain Sciences*, 17(03):558–570, 1994.
4. Shelley H Carson, Jordan B Peterson, and Daniel M Higgins. Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, 17(1):37–50, 2005.
5. Mihaly Csikszentmihalyi. *Society, culture, and person: A systems view of creativity*. Cambridge University Press, 1988.
6. Mihaly Csikszentmihalyi. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 1997.
7. Nizar Grira, Michel Crucianu, and Nozha Boujemaa. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, page 11, 2004.
8. Joy Paul Guilford. *Intelligence, creativity, and their educational implications*. RR Knapp San Diego, 1968.
9. Mark A Hall and Lloyd A Smith. *Practical feature subset selection for machine learning*. Springer, 1998.
10. Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
11. James C Kaufman, Joohyun Lee, John Baer, and Soonmook Lee. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity*, 2(2):96–106, 2007.

12. Sotiris B Kotsiantis. Supervised machine learning: a review of classification techniques. *Informatica (03505596)*, 31(3), 2007.
13. Ulrich Kraft. Unleashing creativity. *Scientific American Mind*, 16(1):16–23, 2005.
14. Yu-Tung Liu. Creativity or novelty?: Cognitive-computational versus social-cultural. *Design Studies*, 21(3):261–276, 2000.
15. Charalampos Mainemelis. Stealing fire: Creative deviance in the evolution of new ideas. *Academy of Management Review*, 35(4):558–578, 2010.
16. Rada Mihalcea and Carlo Strapparava. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599. Association for Computational Linguistics, 2012.
17. Luis Carlos Molina, Lluís Belanche, and Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. In *IEEE International Conference on Data Mining, 2002*, pages 306–313. IEEE, 2002.
18. Davide Piffer. Can creativity be measured? an attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, 7(3):258 – 264, 2012.
19. Karl H Pribram. Brain and the creative act. *Encyclopedia of creativity*, 2:213–217, 1999.
20. Mark A Runco. *Creativity: Theories and themes: Research, development, and practice*. Academic Press, 2010.
21. Mark A. Runco, Jill Nemiro, and Herbert J Walberg. Personal explicit theories of creativity. *The Journal of Creative Behavior*, 32(1):1–17, 1998.
22. A Salappa, M Doumpos, and C Zopounidis. Feature selection algorithms in classification problems: an experimental evaluation. *Optimisation Methods and Software*, 22(1):199–212, 2007.
23. Robert Saunders and John S Gero. Artificial creativity: A synthetic approach to the study of creative behaviour. *Computational and Cognitive Models of Creative Design V, Key Centre of Design Computing and Cognition, University of Sydney, Sydney*, pages 113–139, 2001.
24. Shirish Krishnaj Shevade, S Sathiya Keerthi, Chiranjib Bhattacharyya, and Karaturi Radha Krishna Murthy. Improvements to the smo algorithm for svm regression. *Neural Networks, IEEE Transactions on*, 11(5):1188–1193, 2000.
25. E Paul Torrance. The nature of creativity as manifest in its testing. *The nature of creativity*, pages 43–75, 1988.
26. Geriant Wiggins. Categorising creative systems. In *Proceedings of third (IJCAI) workshop on creative systems: Approaches to creativity in artificial intelligence and cognitive science*. Citeseer, 2003.
27. Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA, 3 edition, 2011.