crea
Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria

Centro di ricerca
Politiche e Bioeconomia

Sentiment analysis and text mining: new opportunities for FADN sample

28th PACIOLI workshop
Ptuj, Slovenia, 1st of October – 4th of October 2023

Concetta Cardillo, Giuliano Gabrieli, Marco Vassallo

https://www.crea.gov.it/web/politiche-e-bioeconomia

Some definitions: Sentiment Analysis

- **Sentiment Analysis (SA) is the task of Natural Language Processing (NLP) with the aim of classifying sentences, opinions, attitudes from natural language expressions by assigning a positive, negative, or neutral, polarization to each expression (Liu, 2012; Sharma et al., 2014).**
- **Sentiment Analyses were performed using a dictionary-based approach with the novel computational linguistic resource named MAL (Morphologically-inflected Affective Lexicon; Vassallo et al., 2019)**

Morphologically-inflected **A**ffective **L**exicon

Some definitions: Text Mining

- **Text Mining-Clustering (TM-C) is a task of Text Mining (TM) that combines data mining techniques, statistics, and computational linguistics to uncovering relationships and patterns in unstructured textual data resources (Gupta & Lehal, 2009; Younis, 2015).**
- **The TM-C is therefore a cluster analysis technique conducted on textual data processed by choosing the software IRaMuTeQ version 0.7 alpha 2 (Ratinaud, 2014) that embeds the hierarchical descending classification (HDC) algorithm known as co-occurrence text analysis (Illia et al., 2014).**

IRaMuTeQ

R Interface for multidimensional analysis of texts and questionnaires

R & python

http://www.iramuteq.org/

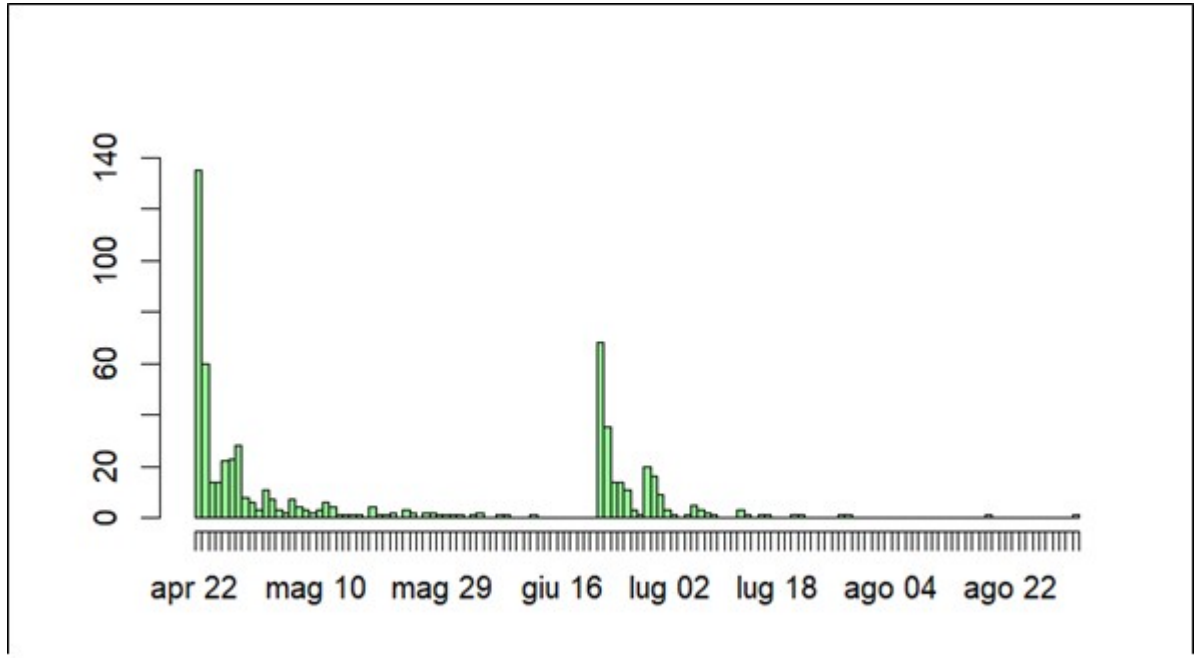
Why RICAsentiment?

- **Objective of RICAsentiment survey: to enlarge the FADN information base by exploiting new analysis techniques to understand the opinion of Italian farmers on specific topics, in this case the CAP.**
- **Starting point: the Italian FADN (RICA), a stratified random sample representative of the Italian agriculture.**
- **Advantages:**

1. **The availability of 10,386 farms, directly contactable and from whom opinions can be asked regarding specific topics.**
2. **For each farm, a lot of structural, economic and social information is also available which allows to complete the framework for interpreting the results.**

The data collection methodology

- RICAsentiment collected 615 questionnaires out of 10,386 sent
- Reference period: April 15 2020 – August 31 2020
- First sending April 15 2020
- Second sending June 15 2020
- Frequency of daily data collection



The 3 questions posed

- What advantages and benefits did the 2014-2020 CAP aid produce for your farm?
- And what are the critical issues that you have encountered?
- Given the new post-2020 CAP reform, what would you expect for your farm in particular from the new programming?

Centro di ricerca
Politiche e Bioeconomia

Results: Distribution of questionnaires per Region

- Different coverage among the regions (average 5,4%)
- All the interviewed answered to first question, less to the other questions

FADN code	REGION	Original sample	Universe	Collected questionnaires	% incidence on theoric sample	N of collected questionnaires for each question		
						1	2	3
221	Valle D'Aosta	170	1.257	8	4,7	8	5	7
222	Piemonte	594	36.818	41	6,9	41	40	40
230	Lombardia	717	30.679	50	7,0	50	47	48
241	Trentino	282	7.223	11	3,9	11	10	10
242	Alto Adige	338	13.070	9	2,7	9	8	8
243	Veneto	707	43.404	49	6,9	49	43	47
244	Friuli-Venezia Giulia	451	9.536	30	6,7	30	28	30
250	Liguria	431	3.858	21	4,9	21	18	18
260	Emilia-Romagna	873	43.435	69	7,9	69	59	66
270	Toscana	577	23.884	43	7,5	43	39	41
281	Marche	452	18.838	22	4,9	22	19	21
282	Umbria	460	11.754	40	8,7	40	36	38
291	Lazio	587	30.160	31	5,3	31	28	29
292	Abruzzo	572	17.174	25	4,4	25	23	24
301	Molise	342	7.057	22	6,4	22	20	20
302	Campania	667	40.979	14	2,1	14	14	14
303	Calabria	510	41.089	23	4,5	23	22	23
311	Puglia	723	69.356	40	5,5	40	37	37
312	Basilicata	400	14.970	19	4,8	19	18	19
320	Sicilia	706	71.681	29	4,1	29	27	27
330	Sardegna	547	30.116	19	3,5	19	17	18
	Italia	11.106	566.338	615		615	558	585

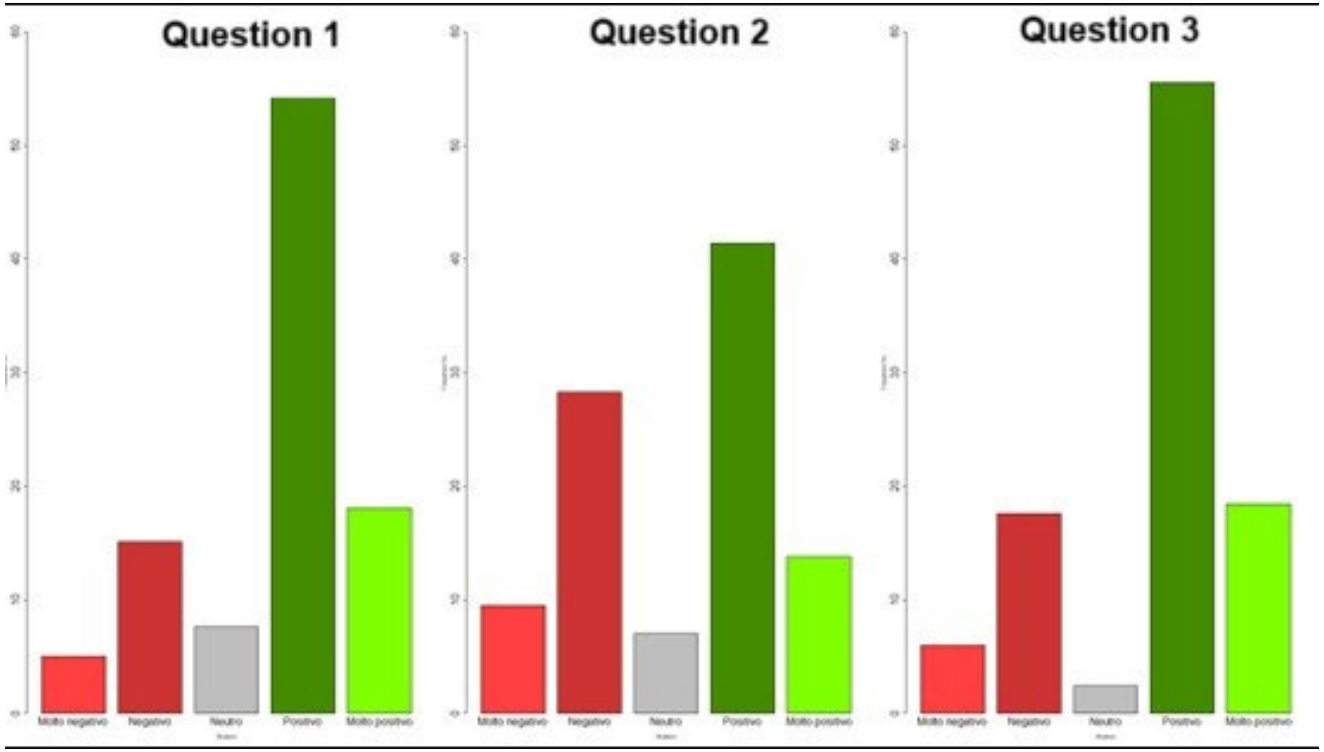
Results: distribution per Type of Farming

- Average value for each ToF: 10,0%
- Min value: 5,2% (olives)
- Max value 17,4% (arable land).

TYPE OF FARMING CODE IN ITALIAN FADN	TYPE OF FARMING	QUESTIONNAIRES	QUESTIONNAIRES (%)
100	Arable land	107	17,4
110	Cereals	93	15,1
200	Horticulture and flowers	59	9,6
310	Grapes	84	13,7
320	Olives	32	5,2
330	Orchards	71	11,5
400	Herbivores	57	9,3
410	Dairy cows	45	7,3
500	Granivores	34	5,5
800	Mixed crops and livestock	33	5,4

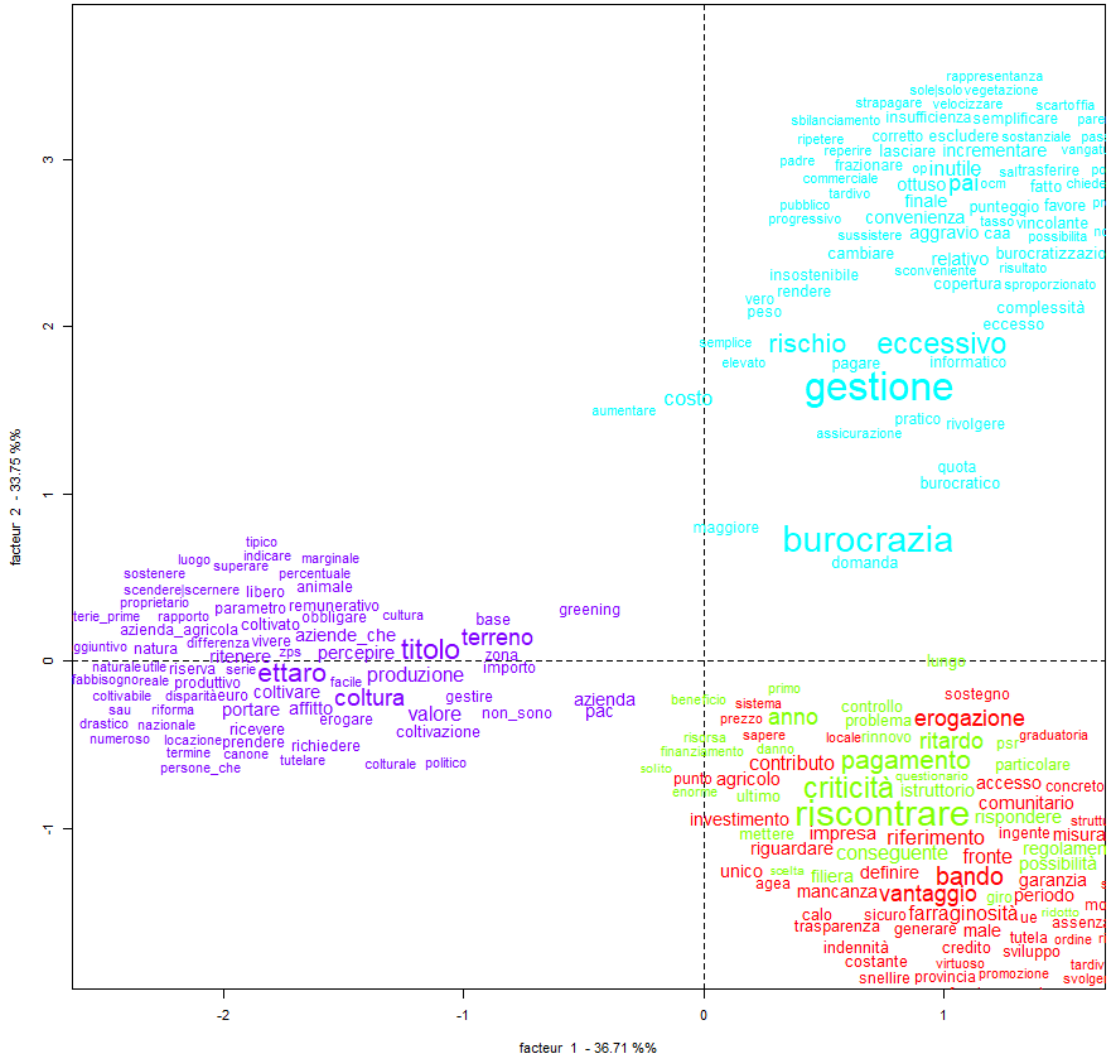
Results: Sentiment analysis

- Sentiment analysis shows how the positive judgments are all greater than the negative ones for all the questions asked.
- However, the result highlights that the polarity of the second question which asks about the critical issues encountered is less concentrated, with a positive polarity just above 50%.
- In the first and third questions high positive polarities are obtained (72.2% and 70.4% respectively).



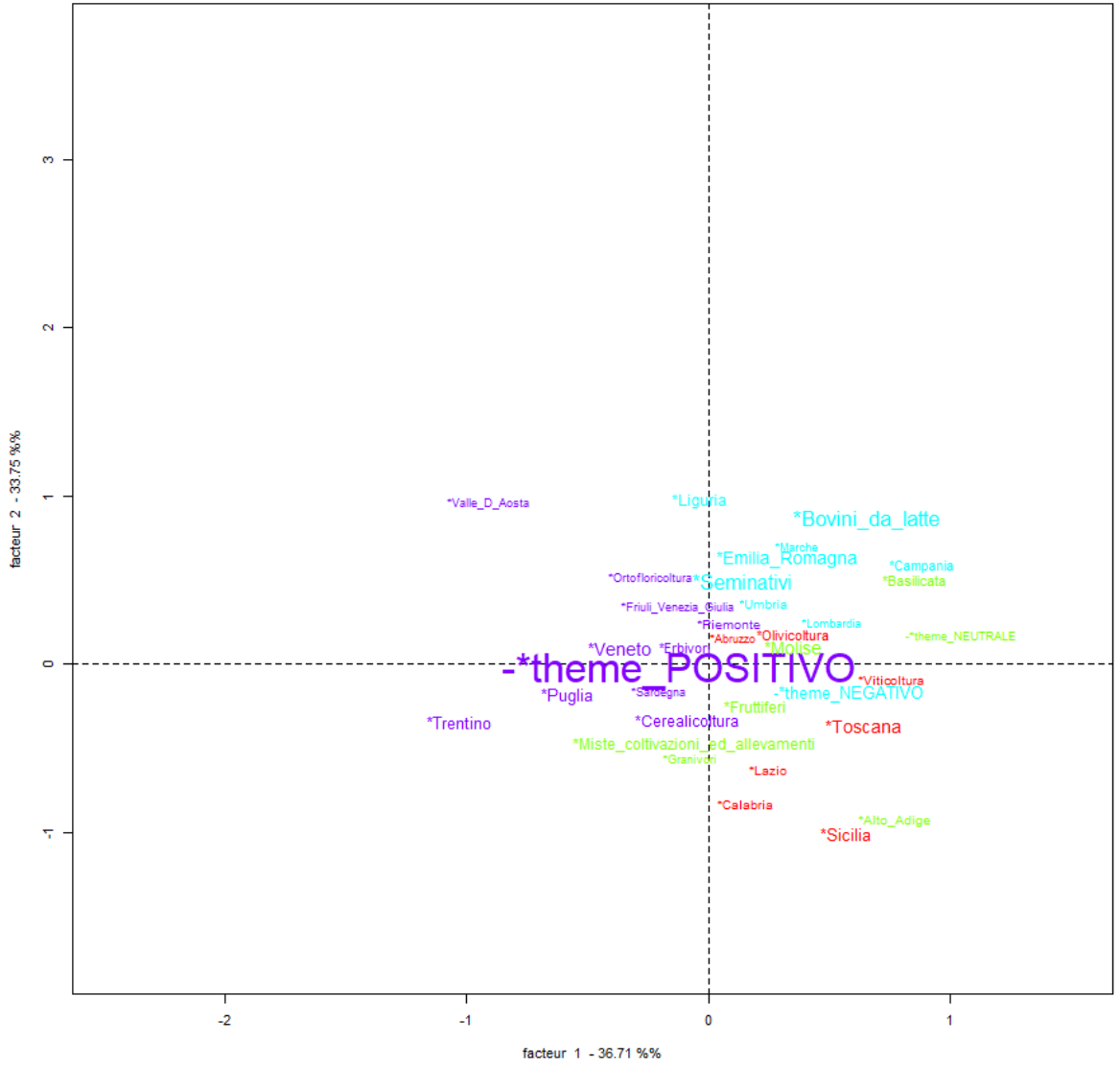
Results: the clusters for question 2

- The method used to identify the clusters is ALCESTE, implemented within the software IRaMuTeQ[3] version 0.7 alpha 2
- It allows to view homogeneous groups of responses by associating them with the types of respondents, in this case agricultural holdings (region, type of farming, polarity).
- The larger character is an indicator of greater frequency on the part of that media in dealing with the topic in question.



Results: the clusters for question 2

- 4 themes representing the critical issues.
- Theme positive indicates that critical issues exist but some farms manage to address them
- Subsequently, the most relevant words were selected on which a similarity analysis was built, with the scores in the branches indicating the strength of the link, useful in interpreting the cluster.

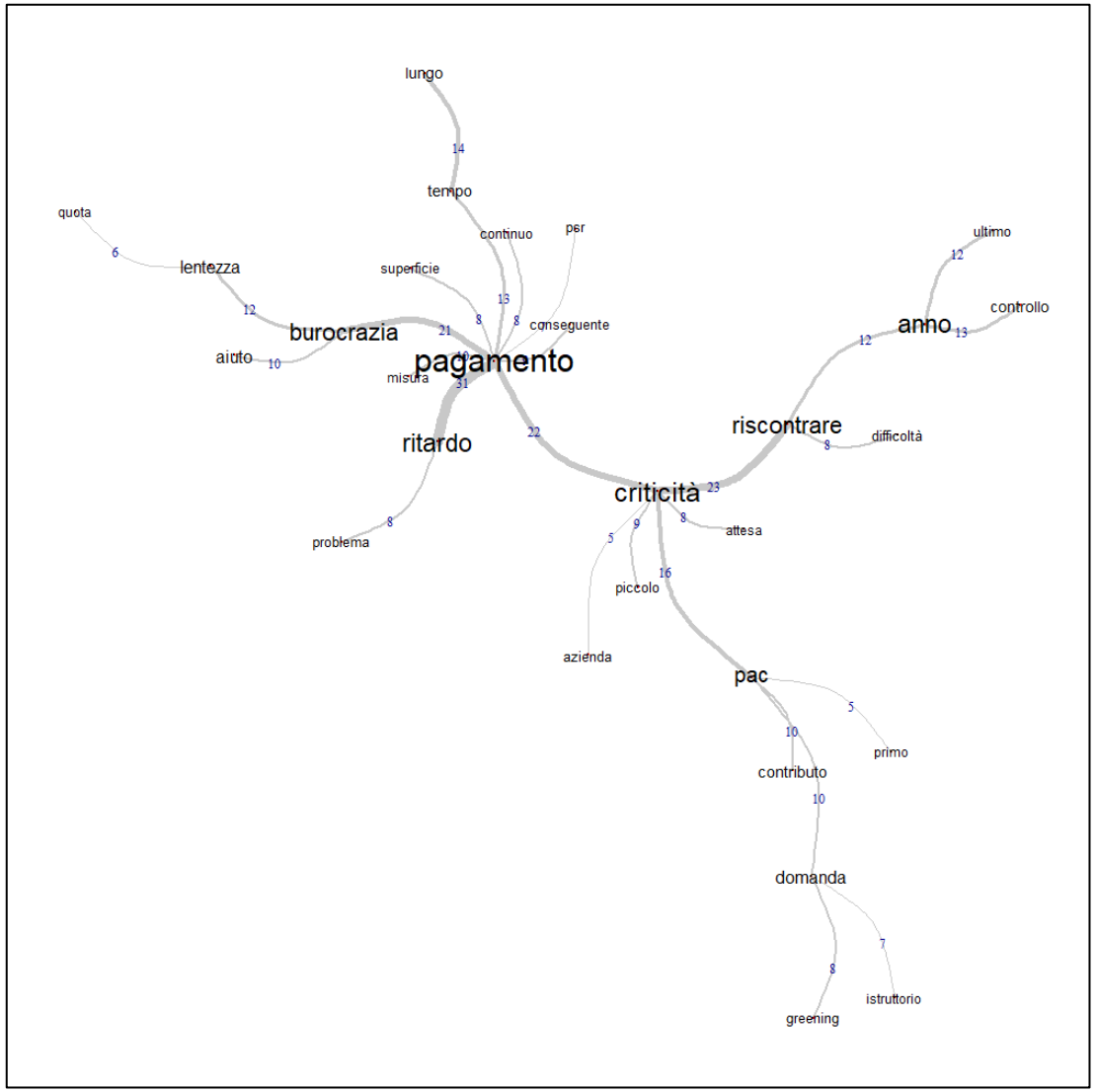


Centro di ricerca
Politiche e Bioeconomia

Results: the cluster «Green»

- The "Green" Cluster is made up of 26.3% of terms.
- Most relevant words: "find", "criticality", "payment", "year" and "delay“
- Farms with a "fruit-bearing" production orientation, "mixed_cultivations_livestock" located in "Molise" and "Basilicata"
- Critical situation indicated by the texts and main difficulties:

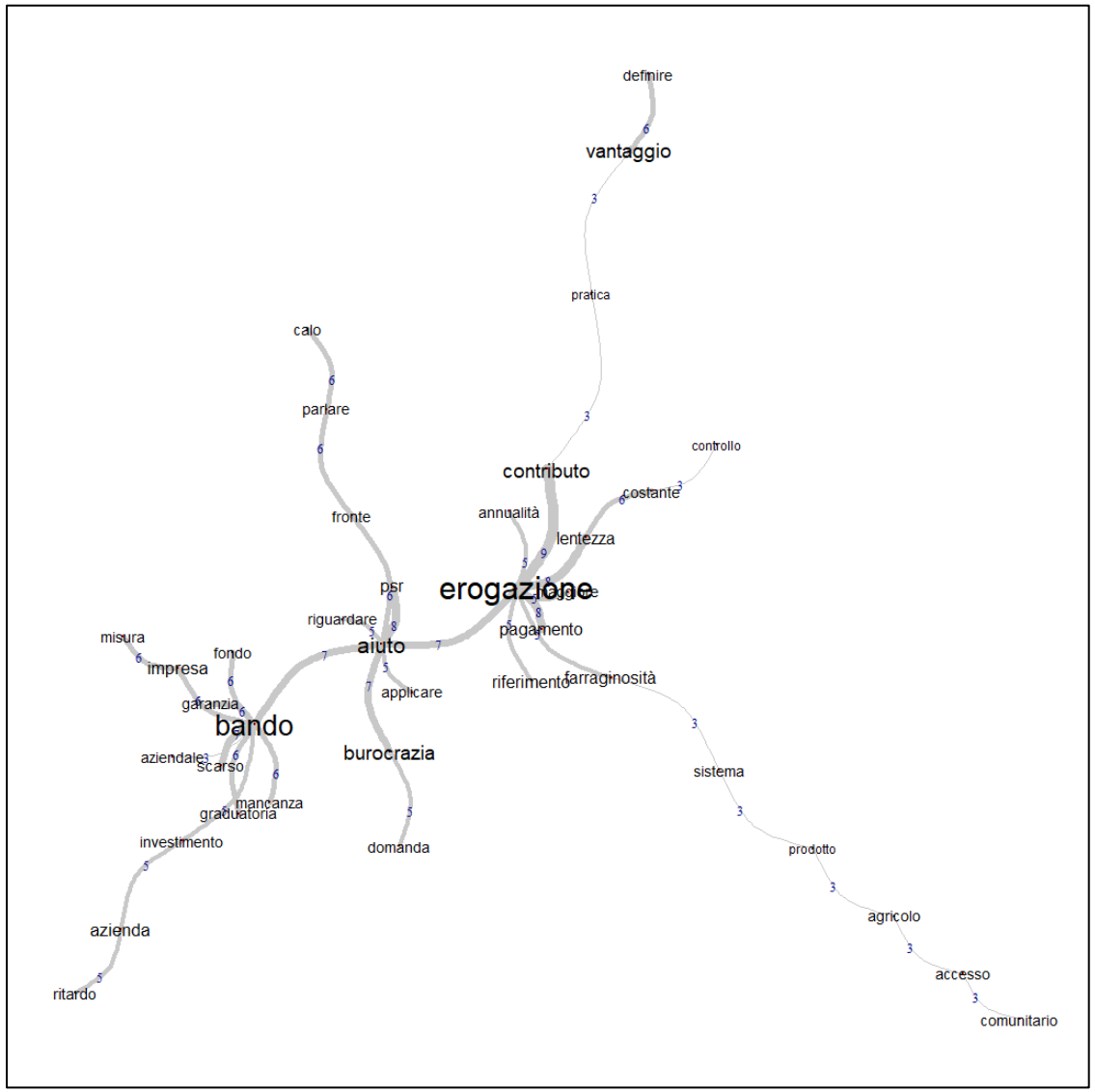
1. In particular for small farms difficulties have been encountered due to the controls which occurred over the last year, they have been considered exaggerated by the farmers.
2. A delay and fragmentation in payments in the use of regional RDP funds and problems related to the calculation of areas (linked to the area premium), in particular for mountain pastures and cereals, where experience strong annual variations.
3. Slowness linked to the bureaucracy that regulates these procedures.
4. Long time required to set up the investigations of the various measures.



Results: the cluster «Red»

- The "Red" Cluster, made up of 22.0% of terms.
- More relevant words: "tender", "advantage" and "disbursement".
- Farm with olives located in "Tuscany" and "Sicily".
- Main critical point indicated:

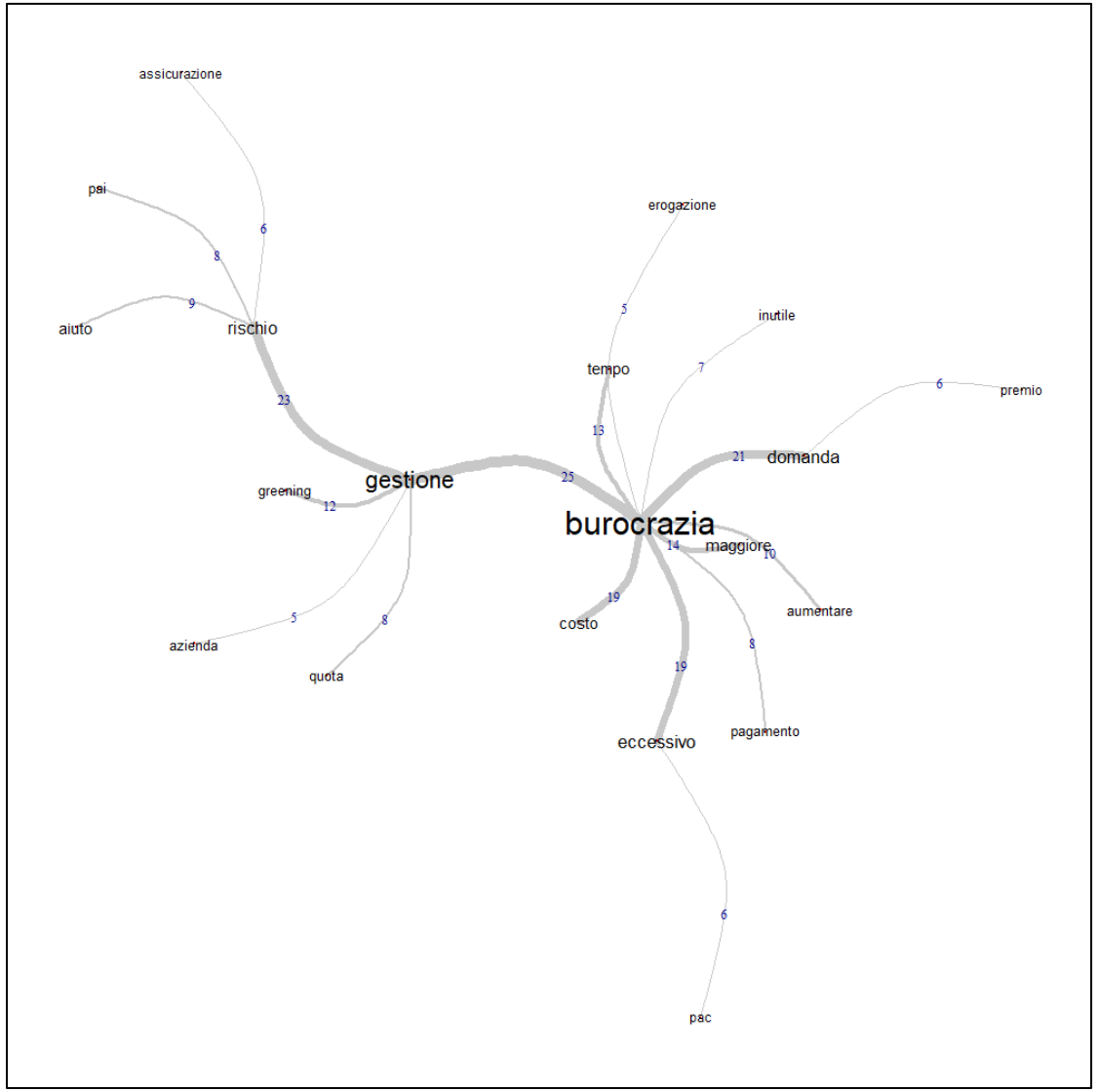
1. Both a lack of funds for all farms and difficulties related to the criteria for determining the rankings.
2. Delays in payments do not allow investments to be planned, with difficulty in accessing credit due to lack of guarantees.
3. A decline in resources, slow payments and cumbersome bureaucracy, especially for small companies.



Results: the cluster «Blue»

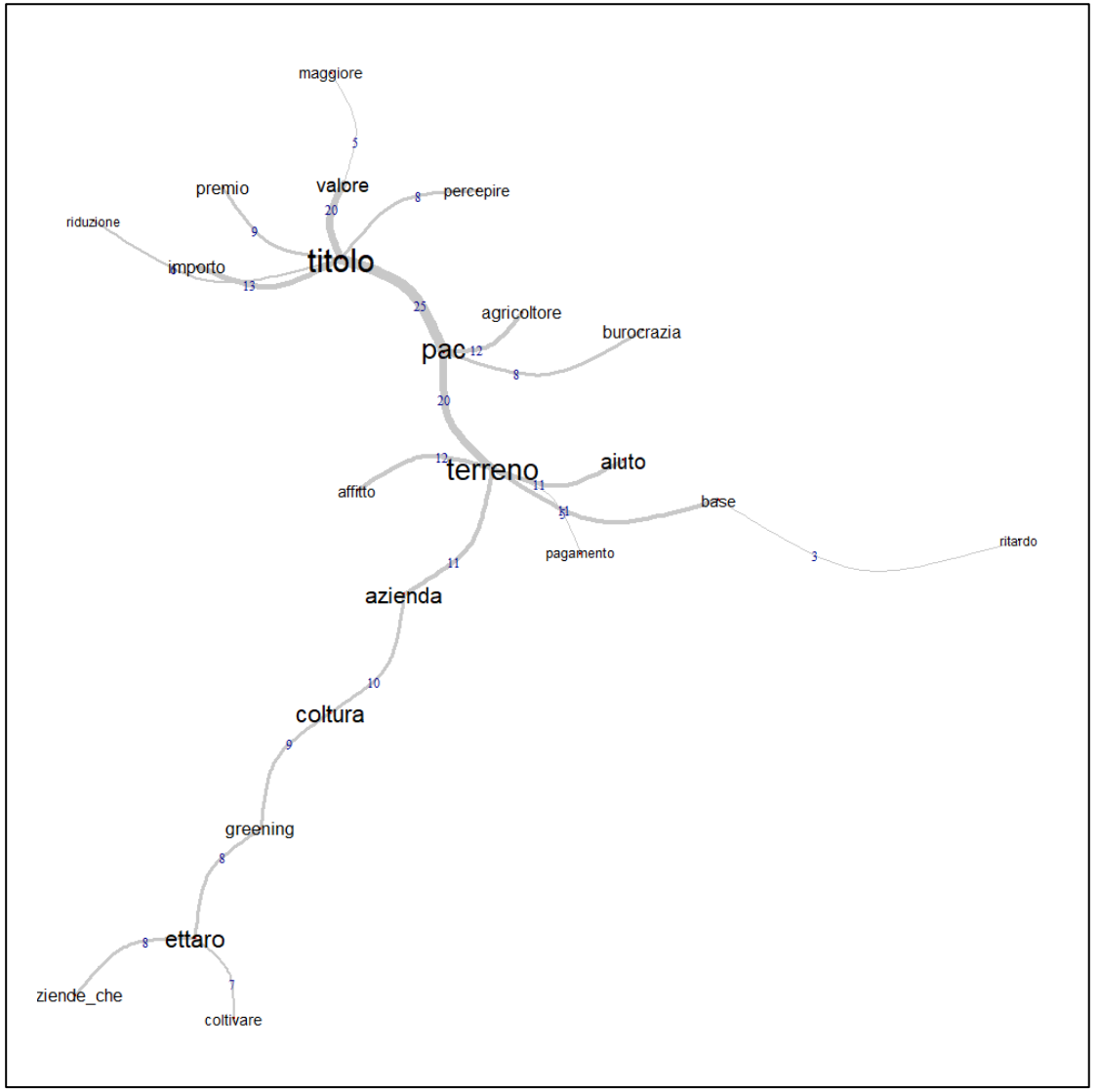
- The “Blue” Cluster is composed of 20.6% of terms.
- Most significant words: “management”, “bureaucracy”, “excessive”, “risk” and “pai” (Individual Plan for Insurance).
- Farms with “dairy cattle" and " arable land" located in "Emilia-Romagna" and "Liguria“.
- Main difficulties highlighted:

1. Difficulties due to requests considered excessive by the officials who follow the measures.
2. All the bureaucratic procedures for risk and PAI management are also considered excessive and cumbersome. Specifically, the pai were deemed not sufficient to make up for the large damages suffered by farms due to the Asian stink bug.
3. The applications are based too much on financial rather than agricultural requirements.
4. The costs incurred by farms to turn to specialized technicians for submitting applications.



Results: the cluster «Purple»

- “Purple” cluster is composed of 31.1% of terms.
- More relevant words: “title”, “hectare”, “crop” and “land”
- Farms with “cereal cultivation" and "herbivores" located in "Veneto", "Puglia", "Trentino" and "Piedmont“.
- Main difficulties emerged:
- The gradual reduction over time of the total amount per hectare of the payment of historical titles that leads to a strong inequality of profitability even between entrepreneurs who practice the same cultivation activity on the same cultivation area.
- The CAP should become more dynamic following the terrain. E.g. if a farmer leases new land he should automatically be entitled to full titles based on the area indicated in the application and not need to search for titles on the market.



Conclusions

- Sentiment Analysis and Text Mining is a new approach to directly understand the farmers' opinions and thus necessities by automatically analyzing the farmers' textual data, that is what the farmers effectively say.
- Reduction of the statistical burden and zero costs.
- This type of analysis allows to collect farmers' opinions on important issues for the agricultural sector and, once the various interested groups and the difficulties encountered have been identified, to better direct policy measures and interventions.
- This would allow to better implement a bottom-up and more participatory approach and achieve better results.

References

- Gupta V., Lehal G.S. (2009). A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence, 1(1).
- Illia L., Sonpar K., Bauer M.W. (2014). Applying Co-occurrence Text Analysis with ALCESTE to Studies of Impression Management. British Journal of Management, Vol. 25, 352–372.
- Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.
- Ratinaud, P. (2014). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Computer software]. Retrieved at http://www.iramuteq.org
- Sharma, R., Nigam, S., & Jain, R. (2014). Polarity detection at sentence level. International Journal of Computer Applications, 86(11).
- Younis E.M.G. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open-Source Tools: An Empirical Study. International Journal of Computer Applications, 112(5).
- Vassallo, M., Gabrieli, G., Basile, V., & Bosco, C. (2019). The tenuousness of lemmatization in lexicon-based sentiment analysis. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), CEUR workshop proceedings: vol. 2481, Proceedings of the sixth Italian conference on computational linguistics, Bari, Italy, November 13-15, 2019. CEUR-WS.org, URL: http://ceur-ws.org/Vol-2481/paper74.pdf

Thank you!

concetta.cardillo@crea.gov.it