Machine Learning and Artificial Intelligence
J.-L. Kim (Ed.)

doi:10.3233/FAIA220422

Applying Social Network Analysis and Data Mining Techniques to Support Decision-Making: A Case Study

Manuela FREIRE[a, c1], Francisco ANTUNES[b, c] and João Paulo COSTA[a, c]
[a]*CeBER, Faculty of Economics, Av Dias da Silva 165, 3004-512 Coimbra, Portugal*
[b]*Department of Management and Economics, Beira Interior University, Portugal*
[c]*INESCC – Computer and Systems Engineering Institute of Coimbra, Portugal*

Abstract. The key goal of this work is to explore interactions and discursive exchanges between social users, to extract information towards decision support. We analyzed customer-generated data on Facebook, during a period of a ten-day strike, of a well-known airline company. The main goal was to check service and responsiveness of the airline, and also to develop indicators that might enable reviewing and reinforce strategies to be used in customer service response to strike events. The authors aim to investigate the possibility of structuring data, collected from OSN's, incorporating human interaction and network structure, using SNA to study the network from a duo fold manner: the web discourse, which depends on the transmission of information; and the interaction among social users, as information disseminators. Our work intends to determine whether social users and their interactions are consistent with the creation of indicators for decision support.

Keywords. Social network analysis, decision support, web discourse, natural language processing, data mining

1. Introduction

Businesses can use tools to monitor online social networks (OSN), to filter social conversations (web discourse) based on certain keywords [1], such as company name, services, etc. The collected information can be integrated into internal systems using data visualization to improve the view of customer interactions. Automated systems can capture these interactions based on customer needs and historical activity. Additionally, Social Network Analysis (SNA), Natural Language Processing (NLP) and Data Mining (DM) [2] algorithms for OSN content analysis can be used to interpret web discourse.

SNA in particular includes various important centrality metrics associated with social networks' studies (for more details please see Freire, Antunes [1], Wasserman and Faust [3], Moreira, Carvalho [4], Savic, Ivanovic [5]). NLP is a computer science field that uses machines to get knowledge from text written in natural language [6, 7] and DM refers to the automated detection of interesting data structures, including techniques that aim to infer, from data, models that attain specific purposes [2, 8].

[1] Corresponding Author, Manuela Freire, CeBER, Faculty of Economics, Av Dias da Silva 165, 3004-512 Coimbra, Portugal, INESCC – Computer and Systems Engineering Institute of Coimbra, Portugal; E-mail: uc45973@uc.pt.

The travelling restart after COVID-19 pandemic has led to many strikes and lockdowns, determining a worldwide crisis in aviation. According to Bartos and Badanik [9] and SimpliFlying [10], in spite of the fact that OSN has become an important part of airline operations, for building relationships with passengers and to stay in touch with them, OSN data analysis isn't receiving the same attention in supporting decision-making or for aiding in crisis management. In this paper we stand that the application of SNA techniques can allow a deeper understanding of airlines' OSN, thus helping to get useful information for previous mentioned purposes [1].

Our work intends to determine whether indicators for decision support can be developed from social users' interactions. To do so, we analyzed customer-generated data on Facebook, during a period of a ten-day strike, of a well-known airline company. The main goal was to check service and responsiveness of the airline, and also to develop indicators that might enable reviewing and reinforce strategies to be used in customer service response to strike events, by means of SNA.

This paper continues as follows. Next section presents some key concepts on SNA supporting decision-making within an OSN's context and an application case study. In Sect. 3, we discuss the achieved results and present our conclusions in Sect. 4.

2. Supporting decision-making using social network analysis

In this study, we used a framework, earlier proposed by Freire, Antunes [11], to extract, process, structure and analyze data from OSN's. That framework, using recurrent and iterative steps, incorporates two important aspects: human interaction and network structure, combining SNA and automatic DM. It can easily be applied to different research domains (e.g., customer service), to a context-specific situation as an airline company. Briefly, Facebook discourse has at least three entities (user, post and concept) that can be analyzed individually or aggregated. Each entity can be turned into a square matrix. To represent the web discourse, adjacency and affiliation, matrices can be used. Simultaneous analysis of the three entities in a network, involves a transformation of a two-mode network into a one-mode network [12-14]. These three matrices were used to construct a multilevel matrix from which the main network (*user*|*post*|*concept*) was created.

In step 1 of the framework, a data extraction process is started. In this step users' activities were collected. We also collected supplementary data from all post's comments. Step 2 calls for data processing and also the representation of social data. Throughout this step, structured data are stored directly in a graph database. Unstructured web discourse and semi-structured interaction data (typically posts), in contrast, are cleaned and enter a normalization process in Step 3, before network graph analysis and the final execution of data visualization techniques. Finally, we analyzed data and obtained output files and results.

2.1. Data extraction

Data extraction was performed using NodeXL [15], collected periodically (every two days) during the ten days strike and the following two days. Only one type of raw network (*user*|*user*) was allowed when collecting data. So, for each dataset, we created two more networks. One (*user*|*post*), to identify the most relevant posts and another (*post*|*concept)*, to identify the concepts that are used the most.

To represent the web discourse, the matrices of the three entities were aggregated and the main network (*user|post|concept*) was semi-automatically built with DM and relational database techniques, using Excel. At the end, we analyze the main network and two other sub-networks for all datasets. **Table 1** summarizes the structural characteristics of each dataset in terms of nodes and edges.

Table 1. Structural characteristics, nodes and edges, of the 6 networks.

Dataset	Node type			Total Nodes	Edges between entities		
	Concept	Post	User		user|post|concept	user|post	post|concept
1st and 2nd of May	2,043	189	394	**2,626**	9,223	614	4,213
3rd and 4th of May	1,973	138	216	**2,327**	4,416	360	3,387
5th and 6th of May	2,426	201	315	**2,942**	7,164	546	4,739
7th and 8th of May	1,580	144	168	**1,892**	3,763	311	3,020
9th and 10th of May	1,315	92	159	**1,566**	3,010	251	2,115
11th and 12th of May	1,294	74	106	**1,474**	2,421	186	1,870

2.2. Data processing and interpretation

Processed data was initially analyzed using Gephi [16] and then SNA techniques were used to study the network. Gephi was used as an SNA tool, avoiding the use of a programming language for data manipulation and visualization, and simplifying the analysis of networks. SNA involves the investigation of social structures, using networks and graph theory. Network structures are characterized in terms of nodes, whether they are users or other entities, and the connections between them [3]. SNA metrics identified the most relevant nodes (in this study user, post or concept), classified them according to their relevance within a community, showed and described results of sub-communities. All data was anonymized taking into account ethical, legal and privacy issues.

Semantic processing transformed unstructured data, from the obtained 769 posts, into a standardized form. We then used the cleaning database and the algorithm proposed in Freire, Antunes [1] to perform the semantic analysis.

The final output posts, with a total of 20,438 concepts, were written in 7 different languages (Dutch - NL, English - UK), French - FR, German - DE, Italian - IT, Portuguese - PT, Spanish - ES). This also implied that the cleaning database was configured for each lexical language.

3. Results and data visualization

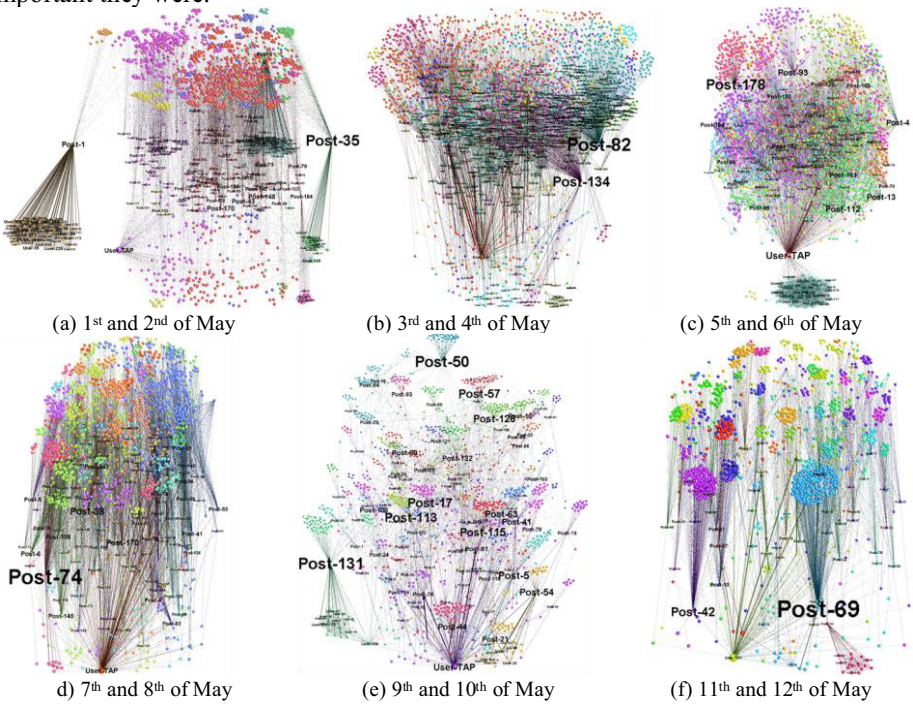
For each dataset, after creating the networks with Gephi, we applied the centrality measure in-degree, out-degree, PageRank, and eigenvector to identify the most influential user and the most relevant post (available at https://influential user & post).

3.1. The web discourse Network: user|post|concept

Figure 1 evidences the results obtained from all 6 analyzed networks. We use a modularity class algorithm to identify communities in the network and used centrality metrics to perceive who were the most important users and their position within the network. In each graph, the nodes at the base of the networks represent the users. The size of the nodes was dimensioned with the betweenness metric, and the label dimensioned with the in-degree metric. The first metric revealed the most influential

users in the dissemination and control of information. The second identified the most important posts, the most commented or viewed.

The graphs show that all the 6 networks had a different structure. The first network (Figure 1a) evidenced two clusters that were isolated from other members of the network. Network 2, 3 and 4 (Figure 1b, c, d, respectively), were the densest. In networks 5 and 6 (Figure 1e and 2f) multiple clusters of smaller dimensions can be viewed. This indicated that some users were more active in the network than others and identified who and how important they were.



(a) 1st and 2nd of May (b) 3rd and 4th of May (c) 5th and 6th of May

(d) 7th and 8th of May (e) 9th and 10th of May (f) 11th and 12th of May

Figure 1. Network: *user|post|concept* (available at *https://user|post|concept*). In the figure, the networks are composed by the three entities of the web discourse, where nodes are user, post and concept, and edges are the connections between the entities.

3.2. The interactions Network: user|post

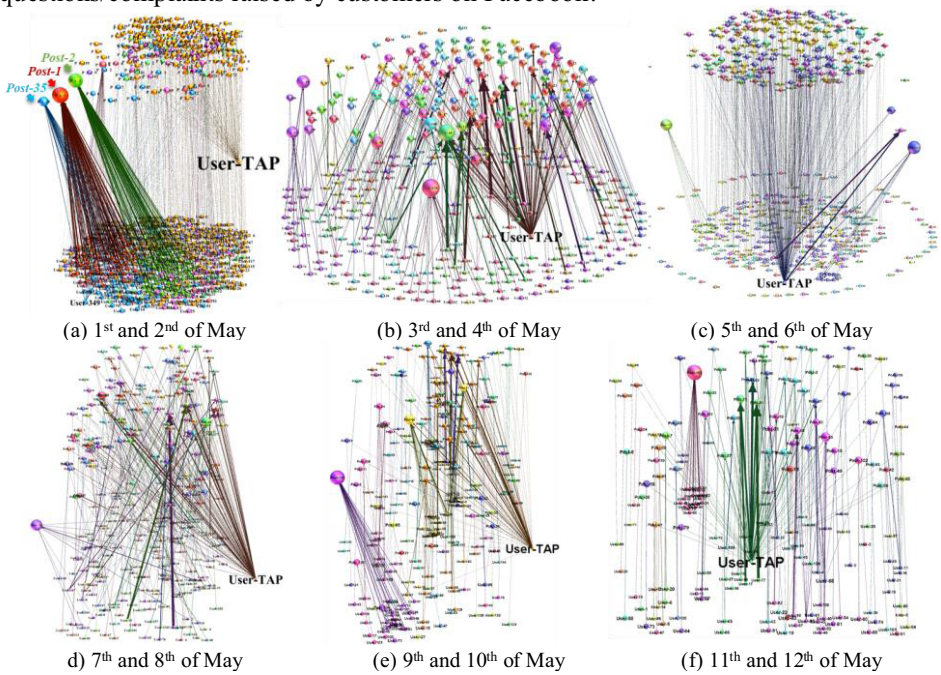
With SNA we can *zoom* on the most relevant social users, in terms of their structural position in the network, and in terms of their posts' content. To check the number of responses provided by the airline company (responsiveness) to customer questions, we used the following metrics:

- In-Degree - to determine the number of contacts that the airline company received. This concept means how other users send information or maintain ties with a particular user, clearly showing the importance of such user's posts. When these posts are broadly commented or displayed, their significance is ascertained.
- Out-Degree - to assess the number of responses given by the airline company (responsiveness). This concept is normally used as a measure of how influential

the user can be within the network, which corroborates the activity that the user has to react to others.

When we look at the interactions between users and posts it was possible to perceive isolated nodes did that not receive any response from the airline company. A deeper analysis on Figure 2a allowed to verify that the red post is not a complaint or a request for information. Looking from the side of the airline company, identifying these posts may be important (even if the only purpose could be to eliminate them). Nodes that have an established connection (link) with the airline company are the ones that were answered. All networks in Figure 2 evidence that the airline company, during all of the strike days, answered many questions on Facebook.

For these networks, subgroups were identified with the modularity class, the size of nodes and labels were defined with the in-degree and out-degree metrics respectively. The nodes that had an established connection with the airline were the posts that were "assisted" by an operator, and, for that reason, the users got a response. By viewing the graphs, it was possible to verify that the company, replied to a considerable number of questions/complaints raised by customers on Facebook.



(a) 1st and 2nd of May (b) 3rd and 4th of May (c) 5th and 6th of May

d) 7th and 8th of May (e) 9th and 10th of May (f) 11th and 12th of May

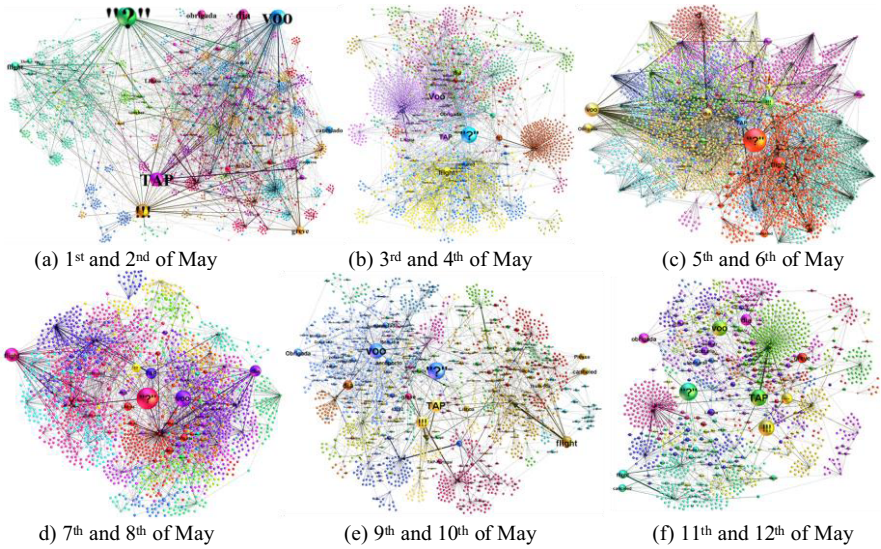
Figure 2. Network: *user|post* (available at *https://user|post*). Each network showing, in the base, the entity user and, at the top, the entity post. Connections represent user's feedback, manifested by different actions such as comment replies, likes, shares, etc. These reactions captured user's interests, as well as status, popularity and relevance of the posts that were established between the entity's *user* and *post*.

3.3. The semantic Network: post|concept (out-degree)

Considering that keywords can be seen as a summary of text within posts, we created a semantic network with all identified concepts and respective post. Figure 3 depicts this sub-network implementation to identify the most relevant concepts and to look for specific information in posts. Using the collected data, we selected posts containing text and built up a semantic network for each two days.

With a semi-automated textual analysis, we were able to extract and rank a list of commonly used keywords (list can be found at https://top 10 concepts). Using the modularity class algorithm, all concepts belonging to the same post are seen as a community. The out-degree metrics were used due to our interest in the nodes that had the most direct votes, since they carry information about the number of times that a concept has been used in a post.

Analyzing the posts' content, we understood that they could be categorized, which would have given the airline company hints on how to better focus/create task groups to address complaints, requests for help or support in changing reservations and other generic information requests (about hotel, refund procedures, etc.).



(a) 1st and 2nd of May (b) 3rd and 4th of May (c) 5th and 6th of May

d) 7th and 8th of May (e) 9th and 10th of May (f) 11th and 12th of May

Figure 3. Network: *post|concept* (available at *https://post|concept*). The networks were built with the entities posts and concepts. The dimension of nodes and labels was defined with the out-degree metric.

When analyzing the most used concepts in each dataset it was found that the question mark, the exclamation mark, and the ellipsis they were widely used. This is justified because, OSN's users use abbreviations, symbols, images etc. to communicate quicker and express some emotions. Particularly within OSN's, users tend to combine linguistic and non-linguistic resources which can speed up communication [17].

4. Conclusions

Social networks and technologies based on OSN's can create insight for decision-making. Post analysis can provide specific means and hints that might help the company to focus and/or create task groups to deal with client complaints, provide help and support on the different issues arising from customer requests (on hotels, luggage collection, refund procedures, etc.). Visualization and keyword identification, as seen from this case, can be used to strengthen the companies need for providing constant information to their clients (using, for instance, online channels or service counters) on: canceled or delayed flights; list of available flights; alternatives to reschedule or change the date of a flight.

Numerous customers complained about the call center not answering calls, no replies to e-mails being sent, airline service counters being closed, and no information about the strike in the airline's website. Post visualization could have been used to identify important conversations, providing direct and instantaneous perceptions on customers' feelings, and would have allowed a faster reaction to customers' dissatisfaction sentiment. The analysis and visualization of web discourse could also have helped the airline to identify possible service problems in traditional channels (call centers, e-mail, etc.) to address them faster and, thus, has the potential to foster any future breakdown management due to strike situations.

The analysis of multiple networks and the calculation of different SNA metrics granted a richer, more structured view of those engaged in the discourse, and the most used concepts. Our conclusion is that it is possible to extract information conducive to decision support, such as predict customer behavior.

Acknowledgments

CeBER's research is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., Project UIDB/05037/2020.

References

[1] Freire M, Antunes F, Costa JP. Getting decision support from context-specific online social networks: a case study. Social Network Analysis and Mining. 2022; 12(1):41.
[2] Fu X, Luo J-D, Boos M. Social Network Analysis: Interdisciplinary Approaches and Case Studies. USA: Taylor & Francis Group; 2017.
[3] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. New York, USA: Cambridge University Press; 1994.
[4] Moreira J, Carvalho A, Horvath T. A General Introduction to Data Analytics. USA: Wiley; 2019.
[5] Savic M, Ivanovic M, Jain LC. Complex Networks in Software, Knowledge, and Social Systems. Library ISR, editor. Switzerland: Springer; 2019.
[6] Isson JP. Unstructured Data Analytics: How to Improve Customer Acquisition, Customer Retention, and Fraud Detection and Prevention. NJ, USA: Wiley; 2018.
[7] Antunes F, Freire M, Costa JP. Semantic web and decision support systems. Journal of Decision Systems. 2016; 25(1):79-93.
[8] Yang J, Xiu P, Sun L, Ying L, Muthu B. Social media data analytics for business decision making system to competitive analysis. Information Processing & Management. 2022; 59(1):15.
[9] Bartos M, Badanik B. Flying Social Media course. Transportation Research Procedia. 2019; 43(2019):119-28.
[10] SimpliFlying. (2019) Airline Social Media Outlook Report 2019. 2019.
[11] Freire M, Antunes F, Costa JP. A semantics extraction framework for decision support in context-specific social web networks. In: Linden I, Liu S, Colot C, editors. Decision Support Systems VII Data, Information and Knowledge Visualization in Decision Support Systems. Switzerland: Springer; 2017. p. 133-47.
[12] Opsahl T. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Elsevier: Social Networks. 2013; 35:159-67.
[13] Banerjee S, Jenamani M, Pratihar DK, editors. Properties of a Projected Network of a Bipartite Network. International Conference on Communication and Signal Processing (ICCSP); 2017 April 6-8; Chennai, India: 2017 IEEE International Conference on Communication and Signal Processing (ICCSP).
[14] Roth C. Knowledge Communities and Socio-Cognitive Taxonomies. In: Missaoui R, Kuznetsov SO, Obiedkov S, editors. Formal Concept Analysis of Social Networks. Cham: Springer International Publishing; 2017. p. 1-18.
[15] Hansen DL, Shneiderman B, Smith M, Himelboim I. Analysing Social Media Networks with NodeXL: Insights From a Connected World. USA: Morgan Kaufmann; 2020.
[16] Bastian M, Heymann S, Jacomy M, editors. Gephi: an open source software for exploring and manipulating networks. Third International ICWSM Conference; 2009.

[17] Antunes F, Freire M, Costa JP. From Motivation and Self-Structure to a Decision-Support Framework for Online Social Networks. In: Information Resources Management A, editor. Research Anthology on Decision Support Systems and Decision Management in Healthcare, Business, and Engineering. Hershey, PA, USA: IGI Global; 2021. p. 161-81.