A Network-based Model and a Related Approach to Represent and Handle the Semantics of Comments in a Social Network

(Discussion Paper)

Gianluca Bonifazi[a], Francesco Cauteruccio[a], Enrico Corradini[a], Michele Marchetti[a], Giorgio Terracina[b], Domenico Ursino[a] and Luca Virgili[a]

[a] *DII, Polytechnic University of Marche*
[b] *DEMACS, University of Calabria*

Abstract

In this paper, we propose a network-based model and a related approach to represent and handle the semantics of a set of comments expressed by users of a social network. Our model and approach are multi-dimensional and holistic because they manage the semantics of comments from multiple perspectives. Our approach first selects the text patterns that best characterize the involved comments. Then, it uses these patterns and the proposed model to represent each set of comments by means of a suitable network. Finally, it adopts a suitable technique to measure the semantic similarity of each pair of comment sets.

Keywords

Comment analysis, Social Network Analysis, Text Pattern Mining, Semantic Similarity, Utility Functions

1. Introduction

In the last few years, the investigation of the content of comments expressed by people in social media has increased enormously [1]. In fact, social media comments are one of the places where people tend to express their ideas most spontaneously [2]. Consequently, they play a privileged role in allowing the reconstruction of the real feelings and thoughts of a person, as well as in building a more faithful profile of her [3, 4, 5][1]. Spontaneity is both the main strength and one of the main weaknesses of comments. In fact, they are often written on the spur of the moment, with a language style that is not very structured, apparently confused and in some cases contradictory. In spite of these flaws, the set of comments written by a certain user

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

g.bonifazi@univpm.it (G. Bonifazi); f.cauteruccio@univpm.it (F. Cauteruccio); e.corradini@pm.univpm.it (E. Corradini); m.marchetti@pm.univpm.it (M. Marchetti); terracina@mat.unical.it (G. Terracina); d.ursino@univpm.it (D. Ursino); luca.virgili@univpm.it (L. Virgili)
0000-0002-1947-8667 (G. Bonifazi); 0000-0001-8400-1083 (F. Cauteruccio); 0000-0002-1140-4209 (E. Corradini); 0000-0003-3692-3600 (M. Marchetti); 0000-0002-3090-7223 (G. Terracina); 0000-0003-1360-8499 (D. Ursino); 0000-0003-1509-783X (L. Virgili)

[1] In this paper, we focus on comments expressed by people through their well-defined accounts. We do not consider anonymous comments because they are less reliable and, in any case, not useful for the objectives of our research.

provides an overview of her thoughts and profile. Reconstructing the latter from the apparent “chaos” inherent in the comments is a challenging issue for researchers working in the context of the extraction of content semantics.

In this paper, we want to make a contribution in this context by proposing a model, and a related approach, to detect and handle the content semantics from a set of comments posted on a social network. We argue that our model and its related approach are able to extract from the apparent “chaos” of comments the thoughts of their publisher and, eventually, to reconstruct the corresponding profile. However, the latter is only one of the possible uses of our model and approach. In fact, if we widen our gaze to the comments written by all the users of a certain community, we are able to understand the dominant thoughts in it. If we consider all the comments on a certain topic (e.g., COVID-19), we can reconstruct the various viewpoints on such a topic. Again, if we consider all comments in a certain time period (e.g., the first three months of the year 2022) we can determine what are the dominants thoughts in that period. Furthermore, the reconstruction of thoughts is only one of the possible applications of our model and approach. Other ones may include, for example, constructing recommender systems, building new user communities, identifying outliers or constructing new topic forums. Some of the most interesting applications are described in [6].

This paper is organized as follows: In Section 2, we present an overview of our proposal. In Section 3, we illustrate our model. In Section 4, we describe our approach. Finally, in Section 5, we draw our conclusions and have a look at some possible future developments. Due to space limitations, we cannot describe here the experiments carried out to test our model and approach. However, the interested reader can find them in [6].

2. An overview of our proposal

Our approach consists of two phases, namely *pre-processing* and *knowledge extraction*.

The pre-processing phase is aimed at cleaning and annotating the available comments and, then, selecting the most meaningful ones. During the cleaning activity, bot-generated content, errors, inconsistencies, etc., are removed, and comment tokenization and lemmatization tasks are performed. The annotation activity aims to automatically enrich each lemmatized comment with some important information, such as the value of the associated sentiment, the post to which it refers, the author who wrote it, etc.

The selection of the most significant comments is based on a text pattern mining technique. While most of the approaches to perform such a task proposed in the past consider only the frequency of patterns [7], our technique considers also, and primarily, its utility [8, 9], measured with the support of a utility function. Regarding this function, we point out that our technique is orthogonal to the utility function used. As a consequence, it is possible to choose different utility functions to prioritize certain comment properties over others. A first utility function could be the sentiment of comments; it could allow, for instance, the identification of only positive comments or only negative ones. A second utility function might be the rate of comments; it might allow, for instance, the selection of patterns involving only high rate comments or only low rate ones. A third utility function could be the Pearson’s correlation [10] between sentiment and rate; it could allow, for instance, the selection of patterns involving only comments with discordant (resp., concordant) sentiment and rate. More details on utility functions can be found in Section 3.

Once the comments and patterns of interest have been selected, it is necessary to have a model for their representation and management. As mentioned in the Introduction, in this paper we propose a new network-based model called CS-Net (Content Semantic Network). The nodes of a CS-Net represent the comment lemmas. Its arcs can be of two types, which reflect two different perspectives on the investigation of comment semantics. The first one, based on the concept of co-occurrence, reflects the past results obtained by many Information Retrieval researchers [11]. It assumes that two semantic related lemmas tend to appear together very often in sentences. The second one, based on the concepts of semantic relationships and semantically related terms, reflects the past results obtained by many researchers working in Natural Language Processing [12]. Actually, the CS-Net model is extensible so that, if in the future we wanted to add additional perspectives for investigating comment content, it would be sufficient to add a new type of arc for each new perspective. The CS-Net model is described in detail in Section 3.

After selecting the comments and patterns of interest, and after representing them by means of CS-Nets, a technique to evaluate the semantic similarity of two CS-Nets is necessary. This technique operates by separately evaluating, and then appropriately combining, the semantic similarity of each pair of subnets obtained by projecting the original CS-Nets in such a way as to consider only one type of arcs at a time. The combination of the single components is done by weighting them differently, based on the extension of the CS-Net projections from which they are derived. This extension is determined by the number of the corresponding arcs. In particular, our technique favors the most extensive component, because it represents a larger portion of the content semantics than the other. Analogously to the CS-Net model, our technique for computing the similarity of two CS-Nets is extensible if one wants to add new perspectives of semantic similarity evaluation. In fact, to obtain an overall semantic similarity value, it is sufficient to compute the components related to each perspective separately, and then combine them according to the procedure mentioned above. In evaluating the semantics of two homogeneous subnets (i.e., subnets of only co-occurrences or subnets of only semantic relationships), our technique considers two further aspects, namely the topological similarity of the subnets and the similarity of the concepts expressed by the corresponding nodes. To compute the former, we adopt an approach already proposed in the literature, i.e., NetSimile [13]. To compute the latter, we use an enhanced version of the Jaccard coefficient, capable of considering synonymies and homonymies as well. Adding these two further contributions to co-occurrences and semantic relationships makes our approach even more holistic. A detailed description of our technique for evaluating the semantic similarity of two CS-Nets can be found in Section 4.

3. Proposed model

Let $\mathcal{C} = \{c_1, c_2, \cdots c_n\}$ be a set of lemmatized comments and let $\mathcal{L} = \{l_1, l_2, \cdots, l_q\}$ be the set of all lemmas that can be found in a comment of $\mathcal{C}$. Each comment $c_k \in \mathcal{C}$ can be represented as a set of lemmas $c_k = \{l_1, l_2, \ldots, l_m\}$; therefore, $c_k \subseteq \mathcal{L}$. A text pattern p_h is a set of lemmas;

therefore, $p_h \subseteq \mathcal{L}$.

We are interested in patterns with frequency values and utility functions belonging to appropriate intervals. In particular, as far as frequency is concerned, we are interested in patterns whose frequency value is greater than a certain threshold. Instead, for what concerns the utility function, the scenario is more complex, because it depends on the utility function adopted and the context in which our model is used. For example:

- We could employ as utility function the average sentiment value of the comments to which the pattern of interest refers. We call $f_s(\cdot)$ this utility function. It allows us to select patterns characterized by a compound score (and, therefore, a sentiment value) very high (e.g., positive patterns), very low (e.g., negative patterns) or belonging to a given range (e.g., neutral patterns).
- We could adopt as utility function the Pearson's correlation [10] between the sentiment and the score of the comments in which the pattern of interest is present. We call $f_p(\cdot)$ this utility function. It allows us to select: *(i)* patterns having a high sentiment value and stimulating positive comments; *(ii)* patterns having a low sentiment value and stimulating negative comments; *(iii)* patterns having a high sentiment value and stimulating negative comments; *(iv)* patterns having a low sentiment value and stimulating positive comments. Clearly, in the vast majority of investigations, the patterns of interest are those related to cases *(i)* and *(ii)*. However, there may be rare cases where the patterns of interest are those related to cases *(iii)* and *(iv)*.

In the following, we denote by $\mathcal{P}$ the set of the patterns of interest, whose values of frequency and utility function belong to the intervals of interest for the application that is being considered.

We are now able to formalize our model. In particular, a Content Semantics Network (hereafter, CS-Net) $\mathcal{N}$ is defined as $\mathcal{N} = \langle N, A^c \cup A^r \rangle$.

N is the set of nodes of $\mathcal{N}$. There is a node $n_i \in N$ for each lemma $l_i \in \mathcal{L}$. Since there exists a biunivocal correspondence between n_i and l_i, in the following we will use these two symbols interchangeably.

A^c is the set of co-occurrence arcs. There is an arc $(n_i, n_j, w_{ij}) \in A^c$ if the lemmas l_i and l_j appear at least once together in a pattern of $\mathcal{P}$. w_{ij} is a real number belonging to the interval $[0, 1]$ and denoting the strength of the co-occurrence. The higher w_{ij}, the higher this strength. For example, w_{ij} could be obtained as a function of the number of patterns in which l_i and l_j co-occur.

A^r is the set of semantic relationship arcs. There is an arc $(n_i, n_j, w_{ij}) \in A^r$ if there is a semantic relationship between l_i and l_j. w_{ij} is a real number in the interval $[0, 1]$ denoting the strength of the relationship. The higher w_{ij}, the higher this strength. w_{ij} could be computed using ConceptNet [14] and considering both the number of times l_j is present in the set of "related terms" of l_i and the values of the corresponding weights.

An observation on the structure of the CS-Net model is necessary. As specified above, our goal is to model and manage the semantics of the content of a set of comments. CS-Net is a model tailored exactly to that goal. For this reason, it considers two perspectives derived from the past literature. The former is related to the concept of co-occurrence. It indicates that two semantically related lemmas tend to appear very often together in sentences. This perspective is probably the most immediate in the context of text mining. In fact, here, it is well known that the frequency with which two or more lemmas appear together in a text represents an index of their correlation. The potential weakness of this perspective lies in the need to compute the frequency of each pair of lemmas. Moreover, this computation must be continually updated whenever a new comment is taken into consideration. The latter is related to the concepts of semantic relationships and semantically related terms. These refer to several researches conducted in the past in the contexts of Information Retrieval [11] and Natural Language Processing [12]. In this perspective, the meanings of the terms, and thus their semantics, are taken into consideration. Indeed, semantic relationships between terms (e.g., synonymies and homonymies) are a very common feature in natural languages. The main weakness of this perspective lies in the need to have the availability of a thesaurus, which stores the semantic relationships between terms. If such a tool exists, the computation of the strength of the semantic relationship is straightforward. Clearly, additional perspectives could be considered in the future. This is facilitated by the extensibility of our model. Indeed, if one wanted to consider a new perspective, it would be sufficient to add to A^c and A^r a third set of arcs representing the new perspective.

4. Evaluation of the semantic similarity of two CS-Nets

In this section, we illustrate our approach for computing the semantic similarity of content related to two sets of comments represented by means of two CS-Nets $\mathcal{N}_1$ and $\mathcal{N}_2$. It receives two CS-Nets $\mathcal{N}_1$ and $\mathcal{N}_2$ and returns a coefficient σ_{12}, whose value belongs to the real interval $[0, 1]$. It measures the strength of the semantic similarity of the content represented by $\mathcal{N}_1$ and $\mathcal{N}_2$; the higher its value, the higher the semantic similarity. Our technique behaves as follows:

- It constructs two pairs of subnets $(\mathcal{N}_1^c, \mathcal{N}_2^c)$ and $(\mathcal{N}_1^r, \mathcal{N}_2^r)$. The former (resp., latter) is obtained by selecting only the co-occurrence (resp., semantic relationship) arcs from the networks $\mathcal{N}_1$ and $\mathcal{N}_2$. Specifically: $\mathcal{N}_1^c = \langle \mathcal{N}_1, A_1^c \rangle$, $\mathcal{N}_2^c = \langle \mathcal{N}_2, A_2^c \rangle$, $\mathcal{N}_1^r = \langle \mathcal{N}_1, A_1^r \rangle$, and $\mathcal{N}_2^r = \langle \mathcal{N}_2, A_2^r \rangle$. If, in the future, we want to add a new perspective, and therefore a new set of arcs beside A^c and A^r, it will be sufficient to build another pair of subnets corresponding to the new perspective.
- It computes the semantic similarity degree σ_{12}^c and σ_{12}^r for the pairs of networks $(\mathcal{N}_1^c, \mathcal{N}_2^c)$ and $(\mathcal{N}_1^r, \mathcal{N}_2^r)$, respectively. The approach for computing σ_{12}^x, $x \in \{c, r\}$ should be as holistic as possible. To this end, it is necessary to define a formula capable of considering as many factors as possible, among those that are believed to influence the semantic similarity degree of two networks $\mathcal{N}_1^x$ and $\mathcal{N}_2^x$, $x \in \{c, r\}$. In particular, it is possible to consider at least two factors with these characteristics.

  The first factor concerns the topological similarity of the networks, i.e., the similarity of their structural characteristics. The structure of a network is ultimately determined by its nodes and arcs. In our networks, nodes are associated with lemmas, while arcs represent features (e.g., co-occurrences or semantic relationships) contributing significantly to define the semantics of the lemmas they connect. This reasoning is further reinforced by the fact that the semantics of a lemma can be contributed by the lemmas to which it is related in the network (in this observation, the application to the CS-Net of the principle of homophily, which characterizes social networks, takes place). The second factor is much more immediate. In fact, it concerns the semantic meaning of the concepts expressed by the nodes of the CS-Net, each representing a lemma of the set of comments associated with it.

Regarding the first factor, many approaches for computing the similarity degree of the structures of two networks have been proposed in the past literature. We decided to adopt one of these approaches, i.e., NetSimile [13]. This choice is motivated by the fact that the latter has a much shorter computation time than the other related approaches. At the same time, it guarantees an accuracy level adequate for our reference context. NetSimile extracts and evaluates the structural characteristics of each node by analyzing the structural characteristics of its ego network. Therefore, in order to return the similarity score of two networks, it computes the similarity degree of the corresponding vectors of features.

Regarding the second factor, we decided to consider the portion of nodes with the same or similar meaning present in the two subnets of the pair. A simple, but very effective, way to do this is the computation of the Jaccard coefficient between the sets of lemmas associated with the nodes of the two CS-Nets. Actually, the Jaccard coefficient only considers equality between two lemmas, while we can also have lexicographic relationships (e.g., synonymies and homonymies) between them [15]. These can modify the semantic relationships between two lemmas and, therefore, must be taken into consideration. To do so, our technique uses an advanced thesaurus, i.e., ConceptNet [14], which includes WordNet within it. Based on this thesaurus, we redefine the Jaccard coefficient and introduce an enhanced version of it, which we call J^*. It behaves as the classic Jaccard coefficient but takes lexicographic relationships into account.

Given these premises, we can define the formula for the computation of σ_{12}^x:

$$\sigma_{12}^x = \beta^x \cdot \nu(\mathcal{N}_1^x, \mathcal{N}_2^x) + (1 - \beta^x) \cdot J^*(N_1^x, N_2^x).$$

Here:

- $\nu(\mathcal{N}_1^x, \mathcal{N}_2^x)$ is a function that applies NetSimile for computing the topological similarity of $\mathcal{N}_1^x$ and $\mathcal{N}_2^x$.
- $J^*(N_1^x, N_2^x)$ is the enhanced Jaccard coefficient between $\mathcal{N}_1^x$ and $\mathcal{N}_2^x$.
- β^x represents the weight given to the topological similarity of CS-Nets with respect to the lexical similarity of the lemmas associated with their nodes. A discussion on the possible formulas for β^x based on the objectives one wants to pursue in a specific application can be found in [6].

Note that our approach for computing σ_{12}^x can operate on any projection $\mathcal{N}_1^x$ and $\mathcal{N}_2^x$ of the networks $\mathcal{N}_1$ and $\mathcal{N}_2$. In fact, the only constraint related to it is that the arcs of the two networks involved are of the same type x. This allows it to be extensible. Indeed, if we wish to add a new perspective on modeling content semantics in the future, the similarity degree of the corresponding projections of $\mathcal{N}_1$ and $\mathcal{N}_2$ can be computed using the same formula of σ_{12}^x described above.

- It computes the overall semantic similarity degree σ_{12} of $\mathcal{N}_1$ and $\mathcal{N}_2$ as a weighted mean of the two semantic similarity degrees σ^c_{12} and σ^r_{12}:

  $$\sigma_{12} = \frac{\omega^c_{12} \cdot \sigma^c_{12} + \omega^r_{12} \cdot \sigma^r_{12}}{\omega^c_{12} + \omega^r_{12}} = \alpha \cdot \sigma^c_{12} + (1 - \alpha) \cdot \sigma^r_{12}$$

  In this formula, $\alpha = \frac{\omega^c_{12}}{\omega^c_{12} + \omega^r_{12}}$ weights the semantic similarity obtained through the analysis of co-occurrences against the one derived from the analysis of the semantic relationships between lemmas. The rationale behind it is that the greater the amount of information carried out by one perspective, relative to the other, the greater its weight in defining the overall semantics. Now, since $|N^c_1| = |N^r_1|$ and $|N^c_2| = |N^r_2|$, the amount of information carried out by the two perspectives can be measured by considering the cardinality of the corresponding sets of arcs. On the basis of this reasoning, we have that: $\omega^c_{12} = \frac{\omega^c_1 + \omega^c_2}{2}$, and $\omega^r_{12} = \frac{\omega^r_1 + \omega^r_2}{2}$, $\omega^c_1 = \frac{|A^c_1|}{|A^c_1| + |A^r_1|}$, $\omega^c_2 = \frac{|A^c_2|}{|A^c_2| + |A^r_2|}$, $\omega^r_1 = 1 - \omega^c_1$, $\omega^r_2 = 1 - \omega^c_2$. These formulas essentially tell us that the importance of a perspective in determining the overall content semantics is directly proportional to the number of pairs of lemmas it can involve.
  Finally, note that σ_{12} ranges in the real interval $[0, 1]$. The higher σ_{12}, the greater the similarity of $\mathcal{N}_1$ and $\mathcal{N}_2$.
  Like the other components of our approach, the one for computing σ_{12} is extensible. In fact, in the future, if we wanted to add additional perspectives for modeling content semantics, we would simply add to σ^c_{12} and σ^r_{12} an additional similarity coefficient for each added perspective and modify the weights in the formula of σ_{12} accordingly.

5. Conclusion

In this paper, we have proposed a model and a related approach to represent and handle content semantics in a social platform. Our model is network-based and is capable of representing content semantics from different perspectives. It is also extensible in that new perspectives can be easily added when desired. It first performs the detection of the text patterns of interest, based not only on their frequency but also on their utility. Then, it uses these patterns and the proposed model to represent each set of comments by means of a CS-Net. Finally, it adopts a suitable technique to measure the semantic similarity of each pair of comment sets. The latter information can be useful in a variety of applications, ranging from the construction of recommender systems to the building of new topic forums [6].

In the future, we plan to extend this research in various directions. First, we could use our approach as the core of a system for the automatic identification of offensive content of a certain type (cyberbullism, racism, etc.) in a set of comments. In addition, we could study the evolution of CS-Nets over time. This could allow us to identify new trends and topics that characterize a social platform. Finally, we plan to use our approach in a sentiment analysis context. Indeed, in the past literature, there are several studies on how people with anxiety and/or psychological disorders write their comments on social media. We could contribute to this research effort by considering sets of comments written by users with these characteristics, constructing the corresponding CS-Nets and analyzing them in detail. We could also compare these CS-Nets with “template CS-Nets”, typical of a certain emotional state, to support classification activities.

References

[1] X. Chen, Y. Yuan, M. Orgun, Using Bayesian networks with hidden variables for identifying trustworthy users in social networks, Journal of Information Science 46 (2020) 600–615. SAGE Publications Sage UK: London, England.

[2] P. Boczkowski, M. Matassi, E. Mitchelstein, How young users deal with multiple platforms: The role of meaning-making in social media repertoires, Journal of Computer-Mediated Communication 23 (2018) 245–259. Oxford University Press.

[3] F. Cauteruccio, E. Corradini, G. Terracina, D. Ursino, L. Virgili, Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity, Journal of Information Science (2021). doi:https://doi.org/10.1177/01655515211047428, sAGE.

[4] B. Abu-Salih, P. Wongthongtham, K. Chan, K. Yan, D. Zhu, CredSaT: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor, Journal of Information Science 45 (2019) 259–280. SAGE Publications Sage UK: London, England.

[5] S. Ahmadian, M. Afsharchi, M. Meghdadi, An effective social recommendation method based on user reputation model and rating profile enhancement, Journal of Information Science 45 (2019) 607–642. SAGE Publications Sage UK: London, England.

[6] G. Bonifazi, F. Cauteruccio, E. Corradini, M. Marchetti, G. Terracina, D. Ursino, L. Virgili, Representation, detection and usage of the content semantics of comments in a social platform, Journal of Information Science (Forthcoming). SAGE.

[7] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. Koh, R. Thomas, A survey of sequential pattern mining, Data Science and Pattern Recognition 1 (2017) 54–77.

[8] P. Fournier-Viger, J. Lin, B. Vo, T. Chi, J. Zhang, H. Le, A survey of itemset mining, WIREs Data Mining and Knowledge Discovery 7 (2017) e1207. doi:https://doi.org/10.1002/widm.1207, Wiley.

[9] L. Gadár, J. Abonyi, Frequent pattern mining in multidimensional organizational networks, Scientific Reports 9 (2019) 1–12. Nature Publishing Group.

[10] K. Pearson, Note on Regression and Inheritance in the Case of Two Parents, Proceedings of the Royal Society of London 58 (1895) 240–242. The Royal Society.

[11] Y. Djenouri, A. Belhadi, P. Fournier-Viger, J. Lin, Fast and effective cluster-based information retrieval using frequent closed itemsets, Information Sciences 453 (2018) 154–167. Elsevier.

[12] Z. Bouraoui, J. Camacho-Collados, S. Schockaert, Inducing relational knowledge from BERT, in: Proc. of the International Conference on Artificial Intelligence (AAAI 2020), volume 34(05), New York, NY, USA, 2020, pp. 7456–7463. Association for the Advancement of Artificial Intelligence.

[13] M. Berlingerio, D. Koutra, T. Eliassi-Rad, C. Faloutsos, Netsimile: A scalable approach to size-independent network similarity, arXiv preprint arXiv:1209.2684 (2012).

[14] H. Liu, P. Singh, ConceptNet — a practical commonsense reasoning tool-kit, BT technology journal 22 (2004) 211–226. Springer.

[15] P. De Meo, G. Quattrone, G. Terracina, D. Ursino, Integration of XML Schemas at various “severity” levels, Information Systems 31(6) (2006) 397–434.