A word embedding approach to explore a collection of discussions of people in psychological distress

1st Rémy Kessler
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
remy.kessler@univ-ubs.fr

2nd Nicolas Béchet
Université Bretagne Sud
CNRS 6074A
56017 Vannes,France
nicolas.bechet@irisa.fr

3rd Gudrun Ledegen
Université Rennes II
PREFics, EA 4246
5043 Rennes, France
gudrun.ledegen@univ-rennes2.fr

4rd Frederic Pugnière-Saavedra
Université Bretagne Sud
PREFics, EA 4246
56017 Vannes, France
frederic.pugniere-saavedra@univ-ubs.fr

***Abstract*—In order to better adapt to society, an association has developed a web chat application that allows anyone to express and share their concerns and anguishes. Several thousand anonymous conversations have been gathered and form a new corpus of stories about human distress and social violence. We present a method of corpus analysis combining unsupervised learning and word embedding in order to bring out the themes of this particular collection. We compare this approach with a standard algorithm of the literature on a labeled corpus and obtain very good results. An interpretation of the obtained clusters collection confirms the interest of the method.**

***Keywords*—word2vec, unsupervised learning, word embedding.**

I. Introduction

Since the nineties, social suffering has been a theme that has received much attention from public and associative action. Among the consequences, there is an explosion of listening places or socio-technical devices of communication whose objectives consist in moderating the various forms of suffering by the liberation of the speech for a therapeutic purpose [1] [2]. As part of the METICS project, a suicide prevention association developed an application of *web chat* to meet this need. The *web chat* is an area that allows anyone to express and share with a volunteer listener their concerns and anguishes. The main specificity of this device is its anonymous nature. Protected by a pseudonym, the writers are invited to discuss with a volunteer the problematic aspects of their existence. Several thousand anonymous conversations have been gathered and form a corpus of unpublished stories about human distress. The purpose of the METICS project is to make visible the ordinary forms of suffering usually removed from common spaces and to grasp both its modes of enunciation and digital support. In this study, we want to automatically identify the reason for coming on the web chat for each participant. Indeed, even if the association provided us with the theme of all the conversations (work, loneliness, violence, racism, addictions, family, etc.), the original reason has not been preserved. In what follows, we first review some of the related work in Section II. Section III presents the resources used and gives some statistics about the collection. An overview of the system and the strategy for identify the reason for coming on the web chat is given in Section IV. Section V presents the experimental protocol, an evaluation of our system and an interpretation of the final results on the collection of human distress.

II. Related works

The main characteristic of the approach presented in this paper is to only have to provide the labels of the classes to be predicted. This method does not need to have a tagged data set to predict the different classes, so it is closer to an unsupervised (clustering) or semi-supervised learning method than a supervised. The main idea of clustering is to group untagged data into a number of clusters, such that similar examples are grouped together and different ones are separated. In clustering, the number of classes and the distribution of instances between classes are unknown and the goal is to find meaningful clusters.

One kind of clustering methods is the partitioning-based one. The k-means algorithm [3] is one of the most popular partitioning-based algorithms because it provides a good compromise between the quality of the solution obtained and its computational complexity [4]. K-means aims to find k centroids, one for each cluster, minimizing the sum of the distances of each instance of data from its respective centroid. We can cite other partitioning-based algorithms such as k-medoids or PAM (Partition Around Medoids), which is an evolution of k-means [5]. Hierarchical approaches produce clusters by recursively partitioning data backwards or upwards. For example, in a hierarchical ascending classification or CAH [6], each example from the initial dataset represents a cluster. Then, the clusters are merged, according to a similarity measure, until the desired tree structure is obtained. The result of this clustering method is called a dendrogram. Density-based methods like the EM algorithm [7] assume that the data belonging to each cluster is derived from a specific probability distribution [8]. The idea is to grow a cluster as the density in the neighborhood of the cluster exceeds a predefined threshold.

Model-based classification methods like self-organizing map - SOM [9] are focus on finding features to represent each cluster. The most used methods are decision trees and neural networks. Approaches based on semi-supervised learning such as label propagation algorithm [10] are similar to the method proposed in this paper because they consist in using a learning dataset consisting of a few labelled data points to build a model for labelling a larger number of unlabelled data. Closer to the theme of our collection, [11] and [12] use supervised approaches to automatically detect suicidal people in social networks. They extract specific features like word distribution statistics or sentiments to train different machine-learning classifiers and compare performance of machine-learning models against the judgments of psychiatric trainees and mental health professionals. More recently, CLEF challenge in 2018 consists of performing a task on early risk detection of depression on texts written in Social Media[1]. However, these papers and this task involve tagged data sets, which is the main difference with our proposed approach (we do not have tagged data set).

III. Resources and Statistics

The association provided a collection of conversations between volunteers and callers between 2005 and 2015, which is called "METICS collection" henceforth.

To reduce noise in the collection, we removed all the discussions containing fewer than 15 exchanges between a caller and a person from the association, these exchanges are generally unrepresentative (connection problem, request for information, etc.). We observe particular linguistic phenomena like emoticons[2], acronyms, mistakes (spelling, typography, glued words) and an explosive lexical creativity [13]. These phenomena have their origin in the mode of communication (direct or semi-direct), the speed of the composition of the message or in the technological constraints of input imposed by the material (mobile terminal, tablet, etc.). In addition, we used a subset of the collection of the French newspaper, Le Monde to validate our method on a tagged corpus. We only keep articles on television, politics, art, science or economics. Figure 1 presents some descriptive statistics of these two collections.

IV. Methodology

A. System Overview

Figure 2 presents an overview of the system, each step will be detailed in the rest of the section. In the first step (module ①), we apply different linguistic pre-processing to each discussion. The next module (②) creates a word embedding model with these discussions while the third module (③) uses this model to create specific vectors. The last module (④) performs a prediction for each discussion before separating the collection into clusters based on the predicted class.

Collection	METICS	Le-Monde
Total number of documents	17 594	205 661
Without pre-processing		
Total number of words	12 276 973	87 122 002
Total number of different words	158 361	419 579
Average words/document	698	424
With pre-processing		
Total number of words	4 529 793	41 425 938
Total number of different words	120 684	419 006
Average words/document	257	201

Fig. 1. Statistics of both collections.

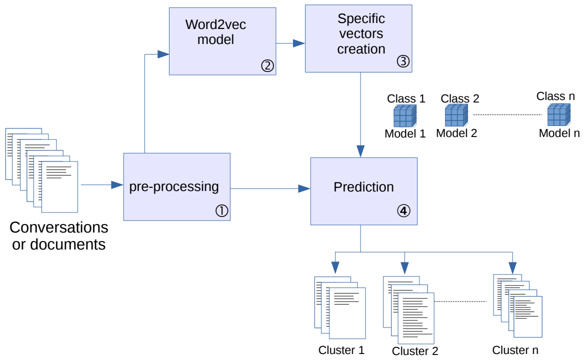


Fig. 2. System overview

B. Normalization and pre-processing

We first extract the textual content of each discussion. In step ①, a text normalization is performed to improve the quality of the process. We remove accents, special characters such as "-","/" or "()". Different linguistic processes are used to reduce noise in the model: we remove numbers (numeric and/or textual), special symbols and terms contained in a stop-list adapted to our problem. A lemmatization process was incorporated during the first experiments but it was inefficient considering the typographical variations described in Section III.

C. word2vec model

In the next step we build a word embedding model using word2vec [14]. We project each word of our corpus in a vector space in order to obtain a semantic representation of these. In this way, words appearing in similar contexts will have a relatively close vector representation. In addition to semantic information, one advantage of such modeling is the production of vector representations of words, depending on the context in which they are encountered. Some words close to a term *t* in a model learned from a corpus *c1* may be very different from those from a model learned from a corpus *c2*. For example, we observe in figure 3 that the first eight words close to the term "teen" vary according to the corpus used. This example also shows that the use of a generic model like Le Monde in French or Wikipedia is irrelevant in our case, since the corpus

[1] http://early.irlab.org/

[2] Symbols used in written messages to express emotions, e.g. smile or sadness

corpus	words
METICS	teenager, young, 15years, kid, school, problem , spoiled, teen,
Le-Monde	sitcom, radio, compote, hearing boy, styx, scamp, rebel

Fig. 3. Eight words closest to the term “teenager” according to the type of corpus in learning.

of the association is noisy and contains a number of apocopes, abbreviations or acronyms. Different parameters were tested and the configuration with the best results was kept[3].

D. *Specific vectors creation and cluster predictions*

In this step, we build vectors containing terms that are selected using the word2vec model described in step IV-C. For each theme in the collection, we build a specific linguistic model by performing a word embedding to reconstruct the linguistic context of each theme. We observe, for example, that the terms closest to the thematic “work” are: “unemployment”, “job”, “stress”. Similarly, for the “addiction” theme, we observe the terms: “cannabis”, “alcoholism”, “drugs” and “heroin”. We used this context subsequently to construct a vector, containing the distance $dist_c(i)$ between each term i and the theme c. Each of these models is independent, so the same term can appear in several models. In this way, we observed that the word “stress” is present in the vector “suicide” and in that of “work”, however, the associated weight is different. We varied the size of these vectors between 20 and 1000 and the best results were obtained with a size of 400. In the last step ④, the system computes an S_c score for each discussion and for each cluster according to each linguistic model such as:

$$S_c(d) = \sum_{i=1}^{n} tf(i) \cdot dist_c(i) \tag{1}$$

with i the considered term, $tf(i)$ frequency of i in the collection, and $dist_c(i)$ is the distance between the term i and the thematic c. In the end, the class with the highest score is chosen.

V. Experiments and results

A. *Experimental protocol*

To evaluate the quality of the obtained clusters, we used a subset of the texts of the Le-Monde newspaper, described in Section III, each article having a label according to the theme. For these experiments, we configured the specific vectors (SV) approach with the optimal parameters, as defined in Sections IV-C and IV-D. We also tested the specific vectors without weighting to test the particular influence of this parameter. To highlight the difficulty of the task, we compare our system with a *baseline* which consists in a random draw, and with the k-means algorithm [3], commonly used in the literature, as mentioned in Section II. To feed the k-means algorithm, we transformed our initial collection into a *bag of words* matrix [15] where each conversation is described by the frequency of its words. Each of the experiments was evaluated using the classic measures of Precision, Recall and F-measure, averaged over all classes (with $beta = 1$ in order not to privilege precision or recall [16]). Since the k-means algorithm does not associate a tag with the final clusters, we have exhaustively calculated the set of solutions to keep only the one yielding the highest F-score.

B. *Results*

	Prec.	Recall	F-score
Without pre-processing			
Baseline	0.18	0.16	0.17
k-means	0.23	0.20	0.22
Without weighting	0.54	0.50	0.52
Specific Vectors	0.53	**0.54**	0.53
With pre-processing			
k-means	0.30	0.21	0.25
Without weighting	**0.55**	0.51	0.53
Specific Vectors	0.54	**0.54**	**0.54**

Fig. 4. Results obtained by each system.

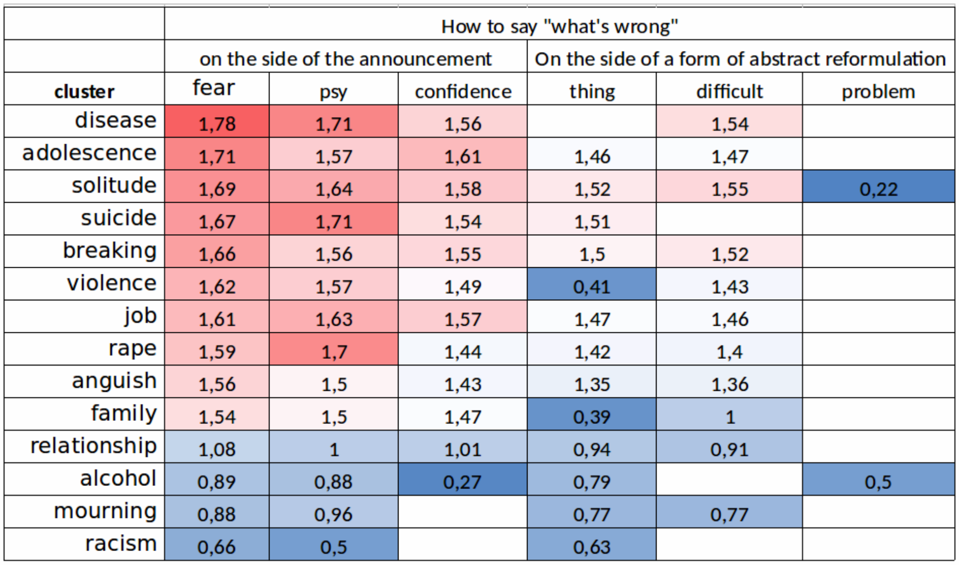
Figure 4 presents a summary of the results obtained with each systems. We first observe that baseline scores are very low, but remain relatively close to the theoretical random (0.2) given by the number of classes. Linguistic pre-treatments are not very efficient individually, but improve overall the results of other experiments. The k-means algorithm obtains slightly better results in terms of F-score, but remains weak. Specific vectors get excellent results that outperform other systems with an F-score of 0.54. The execution without weighting improve slightly the recall.

C. *Cluster Analysis*

Initial objective of this work was the exploration of the METICS collection, we apply the whole process with the specific vectors approach to automatically categorize all the conversations. We use the Latent Dirichlet Allocation [17] in order to obtain the main topic of each cluster. Figure 5 presents average weight of each thematic keywords according to each clusters.

In figure 5, fear, shrink and trust are present designations for each cluster with a largely significant rank; yet, does the writer still express fear when he writes, ”I’m afraid of being sick”? Do these designations not participate in opening and constructing spheres of meanings around these pivotal words? Conversely, with a lower rank, but also significant, the designations of thing, difficult, and problem are more vague, but more reformulating to take up the elements involved in writing what is wrong.

[3] The best results were obtained with the following parameter values: vector size: 700, sliding window size: 5, minimum frequency: 10, vectorization method: skip grams, and a soft hierarchical max function for the model learning.



	How to say "what's wrong"					
	on the side of the announcement			On the side of a form of abstract reformulation		
cluster	fear	psy	confidence	thing	difficult	problem
disease	1,78	1,71	1,56		1,54	
adolescence	1,71	1,57	1,61	1,46	1,47	
solitude	1,69	1,64	1,58	1,52	1,55	0,22
suicide	1,67	1,71	1,54	1,51		
breaking	1,66	1,56	1,55	1,5	1,52	
violence	1,62	1,57	1,49	0,41	1,43	
job	1,61	1,63	1,57	1,47	1,46	
rape	1,59	1,7	1,44	1,42	1,4	
anguish	1,56	1,5	1,43	1,35	1,36	
family	1,54	1,5	1,47	0,39	1	
relationship	1,08	1	1,01	0,94	0,91	
alcohol	0,89	0,88	0,27	0,79		0,5
mourning	0,88	0,96		0,77	0,77	
racism	0,66	0,5		0,63		

Fig. 5. Distribution of discursive routines by cluster.

VI. Conclusion and Future Work

In this article, we presented an unsupervised approach to exploring a collection of stories about human distress. This approach uses a word embedding model to build vectors containing only vocabulary from the linguistic context of the model. We evaluated the quality of the approach on a collection labeled with classical measures. The detailed analysis showed very good results (average Fscore of 0.54) compared to the other systems tested. This method of analysis has also made it possible to highlight semantic universes and thematic groupings. We first intend to study in more detail the influence of each of the parameters on the results obtained. We are also planning to be able to assign several tags to each discussion, which would allow thematic overlaps to be taken into account. The analysis reinforces the cluster approach to highlight the defining features of this type of speech production and to reveal its inner workings. This entry by the discursive routines is only one example which will then make it possible to approach other explorations with a particular focus on the argumentative forms and on the forms of intensity.

References

[1] D. Fassin, “Et la souffrance devint sociale,” in *Critique*. 680(1), 2004, pp. 16–29.

[2] ——, “Souffrir par le social, gouverner par l’écoute,” in *Politix*. 73(1), 2006, pp. 137–157.

[3] MacQueen, J., “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Statistics*. USA: University of California Press, 1967, pp. 281–297.

[4] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[5] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids*. Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics, 1987. [Online]. Available: https://books.google.fr/books?id=HK-4GwAACAAJ

[6] G. N. Lance and W. T. Williams, “A general theory of classificatory sorting strategies1. hierarchical systems,” *The Computer Journal 4*, pp. 373–380, 1967.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” in *Journal of the royal society, series B*, 1977, pp. 1–38.

[8] J. D. Banfield and A. E. Raftery, “Model-based gaussian and non-gaussian clustering,” in *Biometrics*, vol. 49, 1993, pp. 803–821.

[9] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biological Cybernetics*, pp. 59–69, Jan 1982.

[10] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks.” *Physical review. E, Statistical, nonlinear, and soft matter physics*, p. 036106, 2007.

[11] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew, “Sentiment analysis of suicide notes: A shared task,” *Biomedical Informatics Insights*, pp. 3–16, 2012.

[12] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, and P. Poncelet, “Mining twitter for suicide prevention,” in *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings*. Springer, 2014, pp. 250–253.

[13] R. Kessler, J.-M. Torres, and M. El-Bèze, “Classification thématique de courriel par des méthodes hybrides,” *Journée ATALA sur les nouvelles formes de communication écrite*, 2004.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of NIPS’13*. USA: Curran Associates Inc., 2013, pp. 3111–3119. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999959

[15] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[16] C. Goutte and E. Gaussier, “ A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation,” *ECIR 2005*, pp. 345–359, 2005.

[17] M. Hoffman, F. R. Bach, and D. M. Blei, “Online learning for latent dirichlet allocation,” in *Advances in Neural Information Processing Systems*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. 23, 2010, pp. 856–864.