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 Several thousand anonymous conversations have been gathered and form a corpus of unpublished stories about human distress.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