Exploring Temporal Analysis of Tweet Content from Cultural Events

Mathias Quillot, Cassandre Ollivier, Richard Dufour, Vincent Labatut

Exploring Temporal Analysis of Tweet Content from Cultural Events

Mathias Quillot, Cassandre Ollivier,
Richard Dufour, and Vincent Labatut

LIA, University of Avignon, France
{firstname.lastname}@univ-avignon.fr

Abstract. Online social networking platforms are an important communication medium for cultural events, as they allow exchanging opinions almost in real-time, by publishing messages during the event itself, but also outside of this period. Word embedding has become a popular way to represent and extract information from such messages. In this paper, we propose a preliminary work aiming at assessing the benefits of taking temporal information into account when modeling messages in the context of a cultural event. We perform statistical and visual analyses on two word different representations: one including temporal information (Temporal Embedding), the second ignoring it (Word2Vec approach). Our preliminary results show that the obtained models exhibit some similarities, but also differ significantly in the way they represent certain specific words. More interestingly, the temporal information conveyed by the Temporal Embedding model allows to identify more relevant word associations related to the domain at hand (cultural festivals).

Keywords: Word embedding, Temporal representation, Statistical analysis, Cultural events

1 Introduction

Social networks have become a new way of communicating and sharing information and views that can be accessed by billions of people. These social interactions may take the form of short text messages, such as the Twitter platform, where users have the possibility to instantly send a message (here a tweet) containing around 140 characters.

Thanks to its ease of use, Twitter is currently an essential platform for the exchange of messages. For particular events (news, concerts, festivals, presidential elections, etc.), users are increasingly inclined to express themselves through these short messages. Although this platform has become a formidable object of study for a variety of domains ranging from sociology [8, 13, 10] to automatic information extraction [3, 15, 11], the short format of the messages and the large size of the corpora often both make them difficult to analyze. In this article, we analyze messages exchanged through the Twitter platform in the context of cultural events, with a particular focus on festivals. More precisely, we seek to account for shared content, through the words contained in tweets. The major difficulty of this type of analysis lies in the duration of the considered events: although a festival takes place over a defined period (ranging from a few days to several weeks), the activity of users on social networks can intervene at any time (before, during, or after the festival).

We assume that it is difficult to reveal all the information conveyed through the discussions (the tweets) in a global way without taking into account the temporal aspect of the messages (*i.e.* their emission date). A global model would have a tendency to reveal only the frequently shared information, ignoring the uncommon ones that could nonetheless be important over a particular period of time. This is problematic when these models are used as input for information retrieval tasks, such as automatic event extraction, automatic summarization, etc. Based on this observation we propose a preliminary work close to those initiated in [1, 7] that seeks to compare two word embedding-based models: one ignoring the temporal aspect of the messages, using the state-of-the-art *Word2Vec* model [9], and one taking advantage of the emission date of tweets, using the *Temporal Embedding* approach.

The rest of this article is organized as follows. Section 2 presents the different methods used to analyze the impact of time in the events analysis. We presentthe experimental protocol as well as the results obtained in Section 3. Finally, we conclude and give some perspectives in Section 4.

2 Methods

We seek to highlight the interest of taking into account the temporal information conveyed by messages emitted through social networks, in a context of cultural events analysis. In Subsection 2.1, the two compared word embedding representations are described: one considers the complete set of messages regardless of their emission date (the *Word2Vec* neural network method), while the second one takes into account the chronology of the documents (*Temporal embedding* approach). In Subsection 2.2, we describe the methods used to compare the word embedding models, which include both subjective and objective tools. The goal here is to identify the points on which the models differ or converge. Figure 1 summarizes our overall framework, and is detailed in the rest of this section. Our different models are specified there in bold.

We note N the total number of unique words in the corpus. When performing the analysis, one can focus on a list of n words of interest corresponding to a subset of the corpus lexicon: this allows the end user to adopt either a *verification*-directed approach. If the user has some *a priori* knowledge and would like to check certain assumptions regarding the corpus, or a *exploratory* approach, consisting in using the whole lexicon as the word list (in this case, $n = N$).

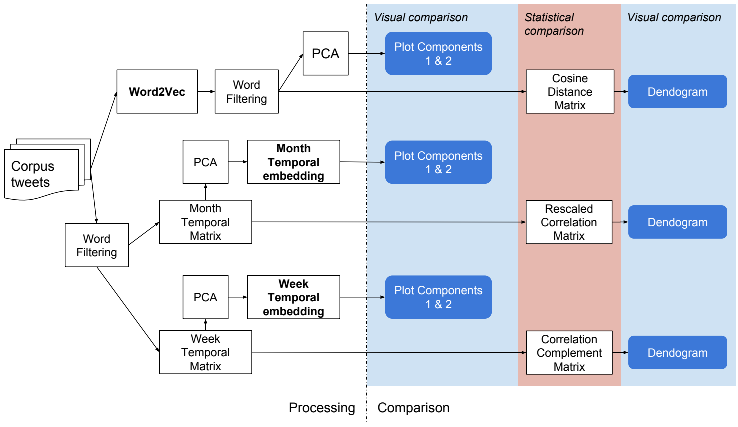

Fig. 1. Overview of the method proposed to process the models (left side) and compare them (right side). The blue boxes represent plots, which are compared visually as explained in Subsection 2.2. The models are represented in bold.

2.1 Word Embedding Representations

We briefly describe the classic *Word2Vec* model, before explaining how the Temporal Embedding model is extracted. Two distinct temporal resolutions are considered in this work (weeks vs. months).

Word2Vec Neural Network. *Word2Vec* models [9] are based on the hypothesis that semantically similar words tend to have similar contextual distributions. Concretely, this context is a window whose size is expressed in a number of words. This is centered on the word of interest. For our experiments we use the CBOW method, it seeks to predict the word reference given a context. For instance, a context of 2 words, the CBOW neural network model takes an input taking the form of a sequence of 4 words w_{i-2}, w_{i-1}, w_{i+1} and w_{i+2}, and outputs a word w_i. We only use the hidden layer of the neural networks, which means each word is represented by a vector. The length of this vector is specified by the user as a parameter d which is 200 by default in literature. We use the complete word vocabulary to train the CBOW model, the method outputs an $N \times d$ matrix. More information about *Word2Vec* models can be found in [9].

Temporal Embedding. Instead of globally taking all the corpus words into account for this method, as for the *Word2Vec* model, we focus on a predefined list of n words of interest (represented by the *Word filtering* box in Figure 1). For each word in this list we count its number of occurrences by time unit where the time unit is either one month or one week, depending on the considered model. In order to avoid zero values, which can be frequent when considering weekly occurrences, we smooth the numbers of occurrences through a moving average. This results in an $n \times m$ occurrence matrix called *Temporal matrix*, where m is the period covered by the corpus expressed in time units. We then perform a *Principal Component Analysis* (PCA) on this matrix. This provides us with the temporal embedding: another $n \times m$ matrix, whose columns are the obtained components ordered by increasing informativeness. The interest of the PCA is to get a more compact representation of the temporal embedding by focusing on the first few components.

2.2 Model Comparison

Comparing models is usually a difficult problem, the most frequent solution being to compare their performance on a targeted task (for example, in speech recognition, the best models are those that allow the lowest word error rate). In this paper, we investigate the interest of including temporal information in word representations from cultural events. To correctly evaluate the impact of each model, an objective ground truth would be necessary, i.e. knowing which words clearly represent a particular event and should be associated with it (and therefore which words have no interest). This would reveal the model that best represents a targeted cultural event. Since no ground truth is available, we first perform an objective comparison relying on statistical tests. These results are generally considered more reliable (and more easily comparable because experiments are reproducible). We seek to know with this first evaluation how close or different the models are. Then, we perform a more detailed analysis through a subjective comparison based on a visual human interpretation in order to investigate the contribution of the temporal information to cultural event representation.

Objective comparison. Two statistical tests are used to compare the models globally: Wilcoxon-Mann-Whitney and Kendall's τ. These non-parametric tests check the hypothesis whether two samples originate from the same distribution. They are complementary, in the sense the former can be considered as a median-based version of the t-test, whereas the latter focuses on rank correlation. We apply them to the Rescaled Correlation and Cosine Distance matrices which were previously computed for the models when extracting the plots and dendrograms. The tests allow us to compare the way the pairs of words are ordered in the different models based on their distances. In other words, they compare the models based on the *relative* positions of the words in the model spaces.

Besides this global comparison, we also perform a local one by focusing on each word separately. For a given word, we perform the same Kendall's and Wilcoxon-Mann-Whitney tests as before, on the distances between this word of interest and the other considered words. When comparing two models, we ultimately identify two groups of words: those whose relative positions are significantly different in these models (using a significance level of .05), and the others, whose relative positions are supposedly similar in both models.

Subjective comparison. The second comparison is visually performed by humans, based on graphical representations of the models. We consider two of them: 1) the projection of the words in a 2D space, and 2) dendrograms. The former representation allows us to identify opposition between words, whereas the second one focuses on their associations.

The 2D space representation is obtained by considering the first 2 components (*i.e.* the most informative ones) of a PCA. Since the Temporal Embedding model includes a PCA, no additional process is required. For the *Word2Vec* approach, some additional processing is needed: we extract the n rows corresponding to the word list from the $N \times d$ matrix (this step is represented as the *Word filtering* box in Figure 1), resulting in the $n \times d$ so-called *Filtered Matrix*, on which a PCA is performed. The subjective comparison is conducted by checking how the words are spatially separated by the plot axes and how this differs from one model to the other. Put differently, if an axis separates two pairs of words and these words are away from this axis, we will consider that they are in opposition. If an opposition is present in both models, we consider this as a similarity between the models. On the contrary, if an opposition is found in one model only, we consider it as a difference between the models.

The dendrograms allow us to identify how the models gather words and organize them hierarchically. We obtain them using the standard hierarchical clustering algorithm available in the `R` Language. Note that this implementation requires a dissimilarity matrix as its input. Moreover, we use the complete linkage approach in order to favor compact clusters with small diameters. In the case of the temporal embedding, we first build an $n \times n$ correlation matrix based on the (week/month) temporal matrix (*i.e.* without the PCA). The correlation between two words is obtained by processing Pearson's coefficient for the two rows associated to these words in the matrix. When the correlation is negative, we set it to zero: this is a common practice when dealing with temporal series because in this context they are generally considered as noise. The dissimilarity is then obtained through the following rescaling: $1 - Cor$, which in our case produces values ranging from 0 (similar) to 1 (dissimilar). This gives us an $n \times n$ matrix, which is called *Rescaled Correlation Matrix* in Figure 1. For the *Word2Vec* model, we build an $n \times n$ *Distance Matrix* based on the previously computed $n \times d$ filtered matrix. Each element of this distance matrix corresponds to the Cosine distance between two words of the list. Let us note w_1 and w_2 the respective word embeddings of these two words, then the distance is given by: $d(w_1, w_2) = 1 - Sim(w_1, w_2)$, where $Sim(w_1, w_2)$ is the classic Cosine similarity between the words. Here the Cosine approach seems more appropriate than the Correlation used on the temporal data because we know that the rows of the considered matrix are inter-dependent by construction. After having generated the dendrograms, we make the visual comparison by checking if two words which are contiguous in one dendrogram are also placed together in another dendrogram.

3 Experiments

In this section, we briefly present the data analyzed in this study (Subsection 3.1), before describing and discussing the obtained results (Subsection 3.2).

3.1 Corpus and List of Words

We use the corpus provided for the MC2 CLEF 2017 lab[1] which contains 70 million tweets[6]. These were automatically retrieved from Twitter using a predefined set of keywords related to cultural festivals in the world. They cover a period ranging from May 2015 to November 2016 and are composed by 134 different languages[5, 4].

We focus on a manually curated list of words of interest. It was originally designed by cultural sociologists to focus on festivals. We extended the list in the following way. Firstly, we added certain cities of interest based on various general and specialized sources: Wikipedia's *List of the world's most liveable* (31 cities) [17], BFM Business's *List of the top* 20 *European cities* [2], and festival cities from Wikipedia's *List of theatre festivals* (30 cities) [18], Red Bulletin's *List of the top* 15 *music festivals* (12 cities) [12], Sky Scanner's *List of the top* 10 *music festivals* [14], and Temps de Vivre's *List of top cinema festivals* [16]. Second, we added other words related to the concept of festival in general: "theater", "music", "film". Third, we added some commercial brands also related to festivals, such as *apple* or *deezer* (33 words). In total, the list contains 119 different words.

3.2 Results

In this subsection we compare two word embedding representations under the form of 3 distinct models[2]. One *Word2Vec* model and two Temporal Embedding models based on two different time units (weeks and months respectively). We first discuss the outcome of the statistical tests before presenting the visual comparison of the plots and dendrograms (see Section 2).

Statistical methods. We apply both the Wilcoxon-Mann-Whitney and Kendall's tests on all 3 models: *Month vs. Week Temporal Embeddings*, *Month Temporal Embedding vs. Word2Vec*, and *Week Temporal Embedding vs. Word2Vec.* All tests return a p-value smaller than 10^{-15}: for these implementations of the tests, this means that they always reject model independence. Kendall's τ, which is an association measure ranging from -1 to $+1$, is 0.68 when comparing the *Month vs. Week Temporal Embeddings*: this corresponds to a strong correlation between these models. For the *Word2Vec vs. Month and Week Temporal models*, we get $\tau = -0.06$ for both comparisons, which means that Word2Vec is almost independent from our temporal models. According to these tests, the information encoded in the temporal models is not the same as the one conveyed by Word2Vec. Thus, they can be considered as complementary and are likely to lead to different results depending on the task at hand.

After the global analysis we switch to individual words to compare the models in terms of which words have significantly different relative positions in the two considered models or a similar position in both models. Based on Kendall's

[1] http://mc2.talne.eu/

[2] http://tac.talne.eu

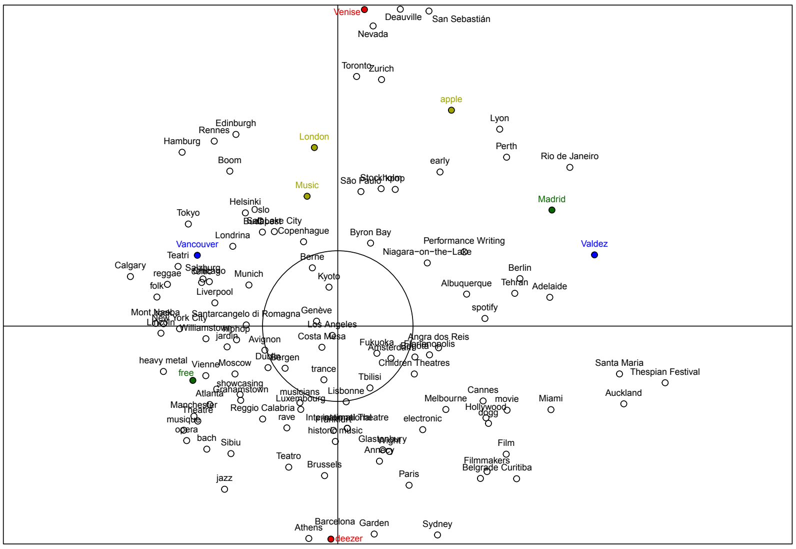

Fig. 2. Two first principal components of the Month Temporal Embedding model.

test, we consider a similar representation when $p < 0.05$ and $\tau > 0.7$. For the *Word2Vec vs. Month Temporal Embedding* models, out of the 119 words constituting our list, only 27 words are represented differently in the models. For the *Month Temporal Embedding vs. Week Temporal Embedding* and *Week Temporal Embedding vs. Word2Vec*, 74 and 25 words are represented differently. With Wilcoxon-Mann-Whitney, we test at $p < 0.05$ and get (in the same order): 64, 25 and 33 differing words. These results show that most words have the similar representation in the different tested model which in accordance to our global test results, means that the difference for the remaining words are very important.

2D representation. We now switch to the visual comparison starting with the 2D plot based on the 2 main components obtained from the PCA. The semantics of these components are not available for the *Word2Vec* model so we do not discuss them. We only consider the positions of word pairs in the graphical representation, and stress the presence of oppositions.

We first compare the Month and Week Temporal Embedding models whose PCA are shown in Figures 2 and 3, respectively. Globally, they seem to present the same oppositions. For instance, "Deezer" vs. "Venise" (in red in all the figures), and "Vancouver" vs. "Valdez" (in blue). This comparison is in line with our statistical results and seems to indicate that it is not necessary to use the week as a time unit because the month-based model requires less data and captures roughly the same information. We then compare both temporal models with the *Word2Vec* one, whose PCA is shown in Figure 4. Visually, the oppositions seem to differ more than previously. For instance, "deezer" and

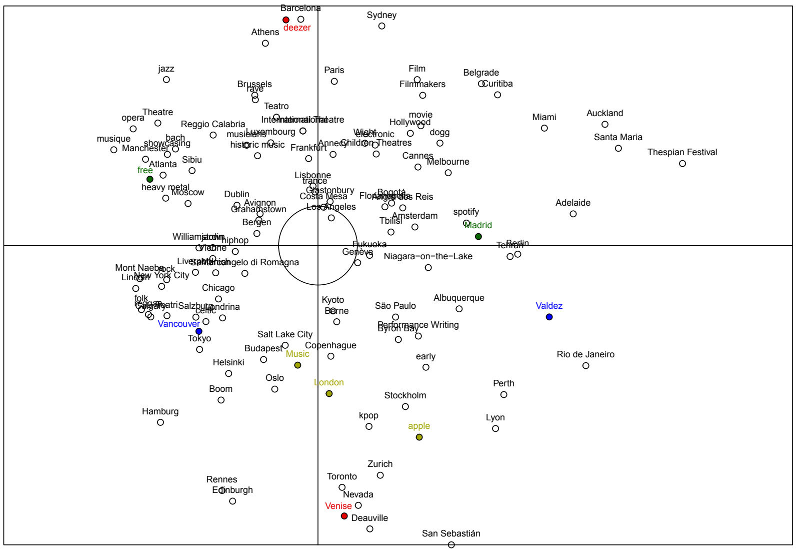

Fig. 3. Two first principal components of the Week Temporal Embedding model.

“Venise“ are not opposed anymore, while “Vancouver“ and “Valdez“ are less opposed. However we can also see some oppositions such as “Madrid“ vs. “free“ (in green), which are common to all 3 models.

Let us focus on an illustrative example: the *Apple Music Festival*, which takes place in London. Both Temporal Models tend to group the 3 concerned words “London“, “Apple“ and “music“ (represented in yellow in the figures), whereas *Word2Vec* does not. This means that temporal models tend to gather words from a same cultural event. If we examine finely the *Word2Vec* representation, we can observe it groups words by semantic category: there are several clusters of cities, whereas “Film“, “Filmmakers“, “Hollywood“ and “movie“ are together (cinematic items), and so are “musique“, “opera“, “theatre“, “jazz“ and “Bach“ (musical items). This is of course consistent of how *Word2Vec* is supposed to work. In conclusion of this visual comparison, we can state that not only are the temporal and *Word2Vec* models different as shown by the statistical tests, but the chronological information encoded by the former also allows to identify relevant groups of words relatively to what we know of the studied corpus.

Dendograms. The dendrograms of the Month and Week Temporal models, which are represented in Figures 5 and 6, respectively, look strongly similar. In particular, we get the same direct connection between “Kyoto“ and “Albuquerque“ (in blue in all the figures), or “Frankfurt“ and “Teatro“ (in purple). Note that they nevertheless differ on some pairs of words such as “Berne“ and “Teatro“ (in green), which are directly connected in the month model whereas “Berne“ is connected to “hiphop“ in the week model. Like before, there are visible differences between the representations of the temporal models and that of the *Word2Vec*,

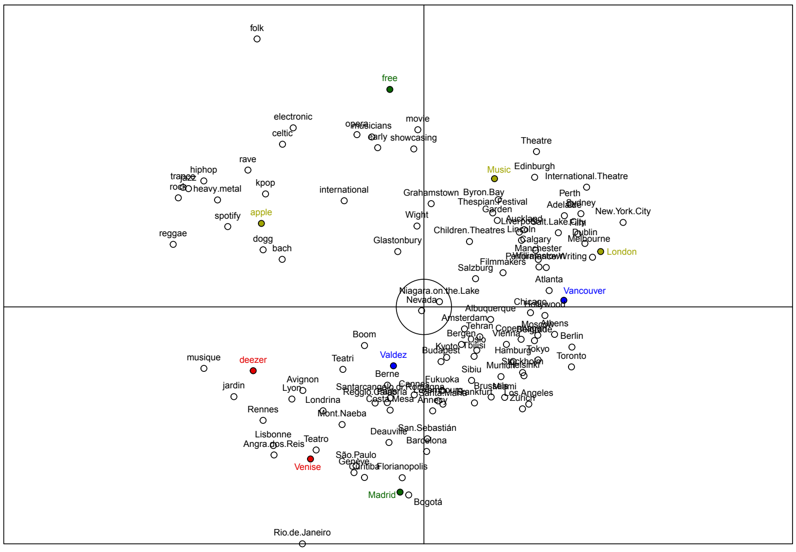

Fig. 4. Two first principal components of the Word2Vec model.

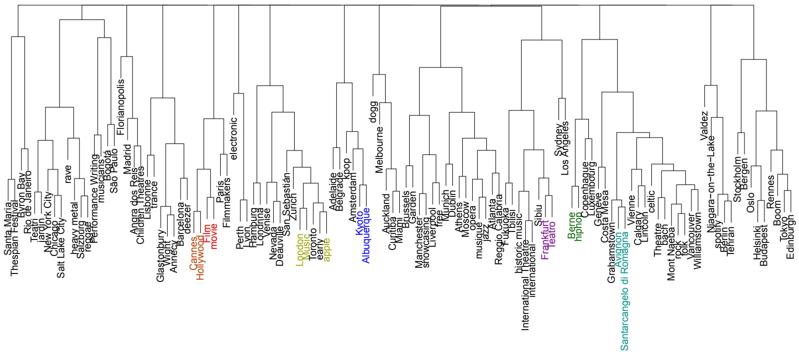

Fig. 5. Dendrogram of the Month Temporal Embedding model.

represented in Figure 7. Focusing on the same pairs of words, we see that "Kyoto" is not connected with "Albuquerque" anymore, and neither are "Frankfurt" and "Teatro". Moreover, "Berne" is neither connected with "Teatro" or "hiphop", but rather with "MontNaeba" (a Japanese mountain).

In the case of *Word2Vec* the words are grouped by semantic similarity. The hierarchical nature of the groupings seem to connect them according to hyper/hyponymy relationships. For instance, "rock" is connected to "heavy.metal" (in orange in the figure), the latter is a subgenre of the former. Both are close to "Jazz" in the dendrogram, itself connected to "reggae" (in red) which can be considered as musical style strongly influenced by certain forms of jazz. Elsewhere, "Tokyo" is connected to "Kyoto" (both Japanese cities, Kyoto is in blue),

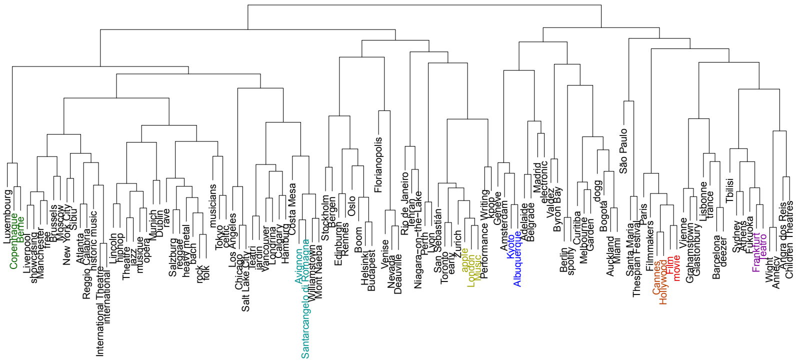

Fig. 6. Dendrogram of the Week Temporal Embedding model.

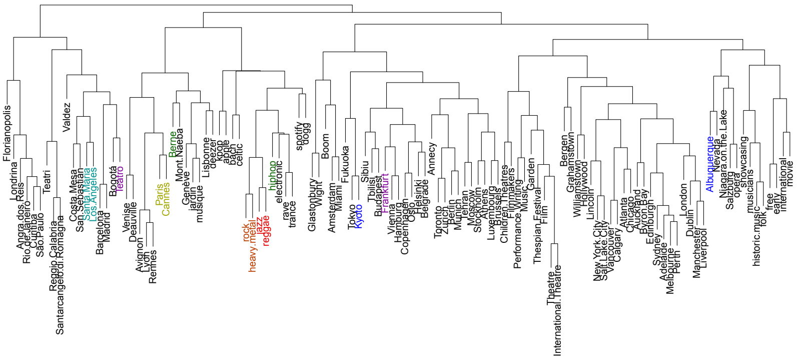

Fig. 7. Dendrogram of the Word2Vec model.

“Santa Maria“ to “Los Angeles“ (both Californian cities, in Cyan) or “Paris“ and “Cannes“ (both French cities, in yellow).

Unlike *Word2Vec*, the temporal models lead to clusters of words which are consistent with the fact that the corpus is related to festivals. For instance in the dendrograms of both temporal models, we observe that “Cannes“ (French city hosting a Cinema festival) and “Hollywood“ (Californian center of the movie industry) are directly connected (both in orange in the figures), they are close to “Film“ and “movie“ (themselves directly connected aas well and represented in red). We also have “Avignon“ (French city) connected to “Santarcangelo di Romagna“ (Italian city): both cities (in cyan in the figures) host a renown theater festival. An other example is “London“, directly connected to “Music“ and close to “Apple“ (all in yellow): this would represent the previously mentioned *Apple Music Festival* in London. Generally speaking in the dendrograms, festival cities are associated to words which are semantically related to the nature of their festival.

The visual comparison of the dendrograms confirm and complete our previous observations. The *Word2Vec* model is definitely different from the temporal ones which are relatively similar to each other. The context-based approach of *Word2Vec* does not capture the temporal information conveyed by the other models. In both cases, words are grouped according to their relative semantics. But in the case of *Word2Vec* these groups are built on semantic proximity. However with the Temporal Embeddings words from the same groups are indirectly related through a festival. In conclusion, the latter models seem more appropriate to a festival-oriented analysis of this corpus.

4 Conclusion

In this preliminary work, our objective was to study how considering temporal information affects the word embedding-based modeling of text corpora. We built two Temporal Embedding models and one *Word2Vec* model based on a corpus of tweets focusing on cultural events. We studied and compared them objectively through statistical tests and visually through PCA plots and dendrograms. It turns out that both temporal models appear to be highly similar according to their PCA plots and dendrograms whereas they seem different from the Word2Vec model. The statistical tests conclude that when taken globally, they are all significantly different from on an other. However, when considering each word separately we show that they differ only on a small proportion of words albeit with a large magnitude. These first results show that temporal information is not completely captured on *Word2Vec* model and *Temporal Embedding* is worth considering, at least when dealing with analyzing temporally messages related to cultural events.

However, a finer analysis is indispensable to correctly characterize the contribution of time information. To this end, we must carry on several steps. Firstly, we could adopt a more exploratory approach by expanding our analysis to the whole lexicon instead of focusing on a predefined list of words. Secondly, we could compare our models built on the considered corpus with out-of-the-box *Word2Vec* models in order to analyze the contribution of the corpus regarding how words are represented.

Acknowledgments

This work was funded by the GaFes project supported by the French National Research Agency (ANR) under contract ANR-14-CE24-0022.

References

1. Basile, P., Caputo, A., Semeraro, G.: Analysing word meaning over time by exploiting temporal random indexing. In: First Italian Conference on Computational Linguistics CLiC-it (2014)

2. BFM Business: List of the top 20 european cities. http://bfmbusiness.bfmtv.com/diaporama/le-top-20-des-villes-europeennes-ou-il-fait-bon-vivre-2832/20-dublin-1/ (2017)
3. Endarnoto, S.K., Pradipta, S., Nugroho, A.S., Purnama, J.: Traffic condition information extraction & visualization from social media twitter for android mobile application. In: ICEEI. pp. 1–4 (2011)
4. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.Y., SanJuan, E.: Cultural micro-blog contextualization 2016 workshop overview: data and pilot tasks. In: CLEF Working Notes. pp. 1197–1200 (2016)
5. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: Overview of the CLEF 2016 cultural micro-blog contextualization workshop. Lecture Notes in Computer Science 9822, 371–378 (2016)
6. Goeuriot, L., Mulhem, P., SanJuan, E.: CLEF 2017 MC2 search and time line tasks overview. In: CLEF Working Notes. pp. 1197–1200 (2017)
7. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096 (2016)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD workshop on Web mining and social network analysis. pp. 56–65 (2007)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)
10. Murthy, D.: Towards a sociological understanding of social media: Theorizing twitter. Sociology 46(6), 1059–1073 (2012)
11. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: Sentiment analysis in twitter. In: SemEval. pp. 1–18 (2016)
12. Red Bulletin: List of the top 15 music festivals. https://www.redbulletin.com/fr/fr/culture/les-15-meilleurs-festivals (2017)
13. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: WWW. pp. 695–704 (2011)
14. Sky Scanner: List of the top 10 music festivals. https://www.skyscanner.fr/actualites/les-10-meilleurs-festivals-de-musique-du-monde (2017)
15. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: ACM SIGIR. pp. 841–842 (2010)
16. Temps de Vivre: List of the top cinema festivals. http://www.tempsdevivre.org/cinebook/festivals.htm (2017)
17. Wikipedia: List of the world's most liveable cities. https://fr.wikipedia.org/wiki/Classements_des_villes_les_plus_agreables_a_vivre (2017)
18. Wikipedia: List of theatre festivals. https://en.wikipedia.org/wiki/List_of_theatre_festivals (2017)