

# Association and Temporality between News and Tweets

Vânia Moutinho<sup>1</sup>, Pavel Brazdil<sup>1</sup> <sup>a</sup> and João Cordeiro<sup>1,2</sup> <sup>b</sup>

<sup>1</sup>LIAAD, INESC TEC – Institute for Systems and Computer Engineering, Technology and Science,  
Rua Dr. Roberto Frias, 4200, Porto, Portugal

<sup>2</sup>HULTIG – Centre of Human Language Technology and Bioinformatics, Universidade da Beira Interior,  
Rua Marquês d'Ávila e Bolama, 6200, Covilhã, Portugal

**Keywords:** Text Mining, Temporal Analysis, Clustering of News, Evolution of Occurrence, Time-wise Differences.

**Abstract:** With the advent of social media, the boundaries of mainstream journalism and social networks are becoming blurred. User-generated content is increasing, and hence, journalists dedicate considerable time searching platforms such as Facebook and Twitter to announce, spread, and monitor news and crowd check information. Many studies have looked at social networks as news sources, but the relationship and interconnections between this type of platform and news media have not been thoroughly investigated. In this work, we have studied a series of news articles and examined a set of related comments on a social network during a period of six months. Specifically, a sample of articles from generalist Portuguese news sources published on the first semester of 2016 was clustered, and the resulting clusters were then associated with tweets of Portuguese users with the recourse to a similarity measure. Focusing on a subset of clusters, we have performed a temporal analysis by examining the evolution of the two types of documents (articles and tweets) and the timing of when they appeared. It appears that for some stories, namely Brexit and the European Football Cup, the publishing of news articles intensifies on key dates (event-oriented), while the discussion on social media is more balanced throughout the months leading up to those events.

## 1 INTRODUCTION


The advent of social media is gradually changing the way information is disseminated and, moreover, possibly shifting the roles of news makers and news recipients. According to (Kaplan and Haenlein, 2010), social media is the set of applications based on the Internet where user generated content (UGC) is created, modified and exchanged in a collaborative and participatory way.


Journalism and social media have become more intricately interconnected. Traditionally, people resort to mainstream media to know what is happening in the world. However, this dynamic has been changing in recent years, at least for some news topics. This is due to the fact that a great proportion of the world population has access to platforms which broadcast real time events. According to data from *statista*<sup>1</sup> (2019), 57.54% are active internet users, and 45.38% are active social media users. Consequently,

the beginning of the process of news generation and dissemination has in some cases changed the way journalism is done. It has been recognized in various studies (Newman, 2009; DVJ Insights and ING The Netherlands, 2015) that journalists spend a considerable amount of their time scouting social media for interesting topics to write about, relying on these platforms as more or less reliable sources.

It is therefore relevant to study how events or news come about on these two types of platforms — news articles and social posts. Of particular interest is the issue of how and when they arise, disseminate, gain strength and die. The main objective of our work is to focus on the timing of their generation and the intensity with which they occur in both media and exploit text mining techniques for this aim.

In this preliminary study, we are especially interested in comparing press publication moments with the community discussion moments, in social media, for different kinds of topics: longstanding, entity-oriented, and event-oriented. What are the new temporal trends? Are events first discussed in social media, becoming later main-stream media, or still in the other way around? Understanding these temporal dy-

<sup>a</sup>  <https://orcid.org/0000-0002-4720-0486>

<sup>b</sup>  <https://orcid.org/0000-0003-0466-1618>

<sup>1</sup><http://www.statista.com>

namics are a crucial issue for many other fields. In our findings, we have observed that the trend in the origin of news stories is shifting from traditional media to social media, at least for certain event-oriented topics.

## 2 RELATED WORK

The literature concerning Twitter and news is often directed at using tweets as a single news source. Indeed, there are some studies where Twitter is regarded as a substitute of (rather than complementary platform to) traditional news sources (e.g. (Sankaranarayanan et al., 2009; Zhao et al., 2011; Phuvipadawat and Murata, 2010)). The main reason for this is perhaps the realization that some news break first on Twitter. For example, (Hu et al., 2012) have shown that the capture and death of Osama Bin Laden was made public on Twitter at least 20 minutes sooner than on major U.S. television channels. The authors argue that this may happen due to the role of a particular set of influential users, namely journalists and politicians, whose credibility instantly leads to an immediate reaction on social networks.

Sankaranarayanan et al. (Sankaranarayanan et al., 2009) built a tool called *TwitterStand* with the goal of collecting and diffusing breaking news quicker than conventional news media. This system performs on-line clustering on filtered tweets from a set of manually selected seeders — users that usually post news. In addition, it performs periodic checks to avoid fragmentation and ensure minimal duplication of clusters. Also, it takes advantage of information in the content of the tweet and/or the user’s profile to associate topics to geographic locations. The authors believe that if tweets belonging to a certain cluster mostly come from one location or a set of close locations, then the topic of that cluster is likely to pertain to that geographical area.

Zhao et al. (Zhao et al., 2011) used a corpus of news articles from the journal New York Times (NYT) and tweets from users in Edinburgh, gathered from November 11, 2009 to February 1, 2010, to investigate how similar the topics in Twitter and a traditional news source are. Their results showed some differences regarding the most frequent categories and types of topics: Twitter users tweet mainly about *family and life*, a category not covered by the NYT; *arts* is a topic similarly frequent on both Twitter and the NYT; *world* is much more frequent on the NYT; lastly, while longstanding topics have an equally strong presence, the same does not happen for entity-oriented and event-oriented topics, with Twit-

ter favoring the former and the NYT the latter. Regarding long-lasting topics, there is evidence that their prevalence is not due to an increasing number of users tweeting about them, but to a set of important users who discuss it over time (Kwak et al., 2010).

The above findings bring about relevant aspects of the similarities between Twitter and conventional news sources. Particularly they emphasize the importance of a certain type of users in social networks that foster their role as a news medium. Still, while the reputation and popularity of users is relatively high regarding new information (Hu et al., 2012), the communication structure set upon follower/followee relationships makes Twitter a fast information diffusion network. In fact, this propagation may in some cases not depend entirely on the first user’s network: (Kwak et al., 2010) have found that if a message is retweeted, it quickly reaches an average of 1,000 users, regardless of the number of followers of the first user. This is what the authors call “*the emergence of collective intelligence*”, in the sense that individuals decide what information is good enough to spread and once that decision is made, it almost instantly reaches a massive audience.

None of the above works studied the issue of temporality between news items and Twitter posts. This motivated us to pursue this issue and report our findings in this article.

## 3 METHODOLOGY

The methodology used here is illustrated in Fig. 1. It involves four main stages, from data gathering to association and temporality analysis.

First, tweets and news are collected from news sources and social media. Then, the news items are clustered into groups of similar news. Afterwards, tweets about the same news are associated with the corresponding news clusters. Finally, the association between news and tweets are examined, both regarding the subject and temporality. The most relevant parts of this process are detailed in the following subsections.

### 3.1 Description of the Collected Data

The same story or event can be published in many different articles and shared or commented in many tweets. Therefore, a set of Portuguese news articles and tweets, from the first semester of 2016, was selected from the *POPmine* plat-form, developed in *SAPO Labs*, a large database of news articles and social media content, automatically crawled from the

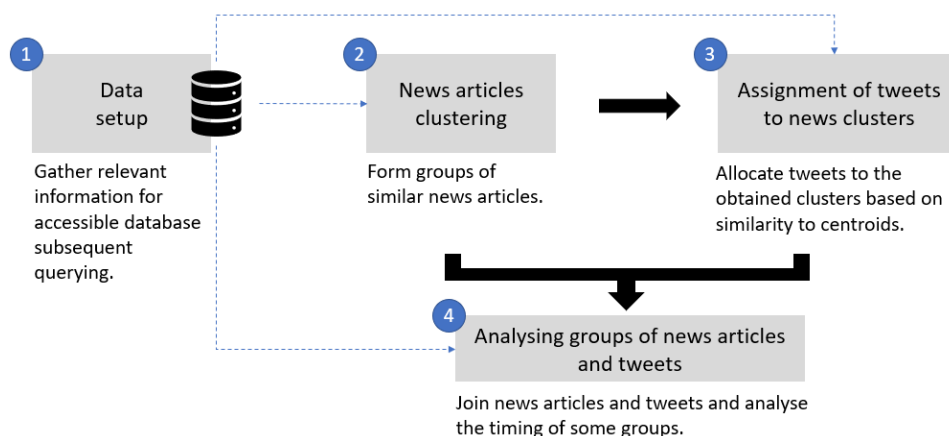


Figure 1: The four main stages of our methodology.

Web (Saleiro et al., 2015).

To fulfill our research objectives, we targeted our data collection on a sample of news stories containing some degree of discussion on Twitter. This sample comprises news stories directly referenced in Twitter, through specific URL pointing to that story, as well as those without any direct link. We are assuming that if a tweet contains a link to a news article, then that tweet is about the same story, naturally meaning that in terms of temporal order the story came out first on mainstream news media. But in this study, we also aim to measure the likelihood of the inverted temporal order, a new phenomenon with the emergence of social media. That is, estimating the number of events and stories that break out first on social media, becoming mainstream media later.

Thus, the final sample of data contains 6,074 news articles and 11,328 tweets, from the same period (the first semester of 2016), eventually associated with those news articles. These were selected from an original set of near 600 thousand elements, observing the following three criteria:

- We have observed a prevalence of sports topics in the news articles from the considered period. In order to obtain a wider range of topics we have decided to focus only on generalist news sources.
- Despite using document vectorization with TF-IDF and cosine similarity, we noticed that longer articles are more likely to be grouped together, as they contain more features, promoting thus document similarity. After several clustering trials, we have decided to only admit documents having a length in the range of 100 to 3,349 characters. The upper bound of 3,349 enables the exclusion of upper outlier documents in terms of document length. The lower bound of 100 characters is used to exclude rather short and uninfor-

mative articles, such as the following examples: “Dados são relativos à zona euro e à União Europeia em geral.”; “Veja na íntegra o debate entre os três candidatos presidenciais, transmitido na SIC Notícias.”. It also prevents some articles that may not have been fully or correctly collected from entering the sample.

- The final criterion for selecting the articles was to include both articles that were shared on Twitter and also articles that were not shared. Hence, 50% of the final sample is comprised of online news articles whose URL was shared in at least one tweet and 50% of online news articles whose URL was not found in any of the tweets. There were 3,037 online news articles with a link to at least one tweet. The other 50% was randomly assembled, resulting in a final sample of 6,074 articles.

Most of the original 19.4 million Portuguese tweets from the first semester of 2016 were discarded, as they report more on personal, family and life subjects, and so irrelevant for our study, as they are not associated with any new stories. The selection of tweets potentially relevant for our study was done as follows. First, 5,664 tweets were selected based on the presence of an URL to one of the previously selected news articles. Secondly, an equal number of tweets without any URL were also selected, following a process that was both random and controlled. First, 250 thousand tweets with more than 20 characters were randomly chosen. Then, only those containing at least one keyword from the clusters (see Section 3.2) of previously selected news, were kept (approximately 50%). Finally, a random subsample of these was selected. The decision of this selection process was therefore a compromise between the processing resources available and the identification of promis-

ing tweets.

Standard pre-processing techniques were applied to both news articles and tweets, including tokenization, lower case conversion, punctuation and numbers removal, Portuguese stopwords removal and stemming. Additionally, we applied parts-of-speech tagging to identify and keep verbs, nouns or proper nouns. The recognition of named entities, such as personalities, names of events and locations, was also included, as we expected these features to help in the representation and pattern recognition of groups of stories or events. This task was done using the PAMPO method (Rocha et al., 2016), built for the Portuguese language.

### 3.2 News Clustering and Labeling

To obtain groups of similar stories clustering techniques were applied to the selected sample of online news articles. We have identified similar news articles and tweets and characterized each group with a number of keywords.

In terms of clustering, the hybrid buckshot method (Cutting et al., 1992) was used, which combines hierarchical clustering with k-means. The initial centers for the k-means clustering are chosen by first running the hierarchical clustering on a sample of  $\sqrt{k} \cdot N$  news articles, where  $N$  represents the number of articles, and  $k$  the number of centroids. The ideal  $k$  value was chosen based on the representation of the aggregation indices of the hierarchical clustering and the representation of the explained inertia, which can be expressed as a ratio B/T. The term B represents *between-class dispersion*, and is measured as the sum of squared distances of the cluster centers (centroids) to the center of gravity  $g$ . The term T represents the *total dispersion of the data*, and is measured as the sum of squared distances of every observation to the center of gravity  $g$ :

$$B = \frac{1}{n} \sum_{h=1}^k n_h d(g_h, g)^2 \quad T = \frac{1}{n} \sum_{i=1}^n d(I_i, g)^2 \quad (1)$$

The explained inertia (B/T) naturally increases with the number of clusters. The goal is to find the value of  $k$  for which this value starts to marginally decrease (the elbow method) (Bholowalia and Kumar, 2014). In our case this was around  $k=50$ . The clusters identified exhibit non-uniform distribution of documents, some with large numbers and others with only a few.

The groups of news articles were labeled using the most significant terms as keywords. The number of keywords varies according to the cluster size, so that larger clusters were represented by a larger

set of keywords. Keywords were ordered according to their average TF-IDF values within the cluster. As an example, Fig. 2 shows clusters of (a) articles related to the *Brexit* referendum, (b) the first news about the run of António Guterres' for *United Nations Secretary-General*, (c) the *European Football Cup* held in France and (d) the *Brussels terrorist attack*. The names of the clusters were given after an examination of the list of keywords. We highlight the importance of named entities identification in this study, performed using PAMPO (Rocha et al., 2016), since terms such as *Reino Unido* (United Kingdom), *União Europeia* (European Union), *Nações Unidas* (United Nations) and *Secretário-Geral* (Secretary-General) are often very indicative of the cluster's main subject.

In this work we have selected a subset of four clusters for further analysis, based on their size, the topic addressed and the number of associated tweets. They were the *Brexit*, *European Football Cup* and *Air transports incidents* presented above, and also a cluster on *Politics*.

### 3.3 Assigning Tweets to Clusters

In order to analyse when a certain text or group of similar texts appear in social networks in comparison to its press release, there is a need to assign social media posts to a particular group of similar texts. In this case, we used a collection of tweets posted during the same period as the online news articles and assigned them to the appropriate clusters of articles.

Tweets were assigned to the clusters of news articles using the cosine similarity measure between each tweet and the cluster centroids, computed using feature vectors with normalized TF-IDF values. For each tweet, the five closest cluster centroids were identified, and the cluster that was most similar to the tweet in question was chosen. The sample of tweets is larger than the sample of articles, and the mean ratio is 1.8 tweets per article. For the clusters under observation, the ratio is larger than the mean ratio (from 2.1 for cluster *Brexit*, to 3.9 for cluster *Air Transport incidents*), with the exception of cluster *Politics* (0.8). This is some evidence that the chosen clusters have a place in the social network discussion.

#### 3.3.1 Assignment

As all the data used in this study are unlabeled, results cannot be directly evaluated. We addressed this issue by including tweets with links to clustered news articles. It can be assumed that if a tweet shares a specific news article, it should belong to the same cluster. So, we consider the cluster as the real class of



Figure 2: Keyword clouds for four news clusters.

the tweet and compare it with the results of the assignment based on similarity to cluster centroids. Accuracy, precision, recall and F1 measures were used to evaluate these results. In addition, we borrowed the concept of precision at  $n$  ( $P@n$ ) from the information retrieval field (Schütze et al., 2008). In this study, we made the following adaptation: each observation leads to  $n$  predictions, based on the distance to the closest news centroids. If the true class of a tweet is in the  $n$ -topmost predictions, it is considered as a true positive. The  $P@n$  is therefore the percentage of observations whose true class is present in the top  $n$  predictions.

We present the values for  $P@n$ , for  $n$  up to five (see Table 1). In this case,  $P@5$  is 43.3%, which means that, for 43.3% of the assigned tweets with link to a news article, the correct cluster was in the top five predictions.

Regarding the detailed analysis of our subset of clusters, *Brexit* has the highest precision value (43.2%) (see Table 2). However, the value of *F1* is higher for *Air transport incidents* (39.7%), *Football - Euro 2016* (37.0%), and *Brexit* is in third place (34.7%). These performance values point to the clusters that are probably the most reliable for the temporality analysis discussed in the next section.

Table 1: Global evaluation of tweets assignment to clusters.

Measure	Value	$P@n$	Value
Accuracy	12.7%	$P@1$	14.2%
MAV Precision	13.8%	$P@2$	21.5%
MAV Recall	33.4%	$P@3$	29.2%
MAV F1	14.7%	$P@4$	36.1%
		$P@5$	43.3%

Table 2: Performance of the selected subset of clusters.

Cluster	Precision	Recall	F1
<i>Air trans. inc.</i>	27.4%	72.2%	39.7%
<i>Foot. Euro 2016</i>	27.4%	57.0%	37.0%
<i>Brexit</i>	43.2%	28.9%	34.7%
<i>Politics</i>	29.7%	1.2%	2.3%

## 4 TEMPORAL ANALYSIS

The main goal of this work was to identify similar stories or events and analyze when they come about on the news and social media. Given the rise of user generated content and the current trend of journalists scouting social media for crowd checking and news monitoring (DVJ Insights and ING The Netherlands, 2015), the hypothesis is that these two environments are interconnected. In this section we present the timeline for all documents in each cluster, in order to gain insights into the evolution of news generation and sharing/commenting on Twitter in Portugal.

For this analysis we focused both on tweets that do not have any link to a news article, as well as on news articles that were not shared on Twitter. This prevents a possible bias towards the hypothesis that, for a given cluster, the discussion on Twitter occurred after its publication by the press.

Fig. 3 presents the temporal evolution of tweets and articles for the four clusters under observation. We recall that *Football - Euro 2016*, *Brexit* and *Air transport incidents* (includes Brussels attack) were the clusters with the best *F1* scores on the evaluation of tweets assignment, and that *Politics* is the largest cluster, albeit with a lower performance evaluation (see Table 2).

It is possible to observe that clusters *Football - Euro 2016* and *Brexit* have peaks in the number of articles

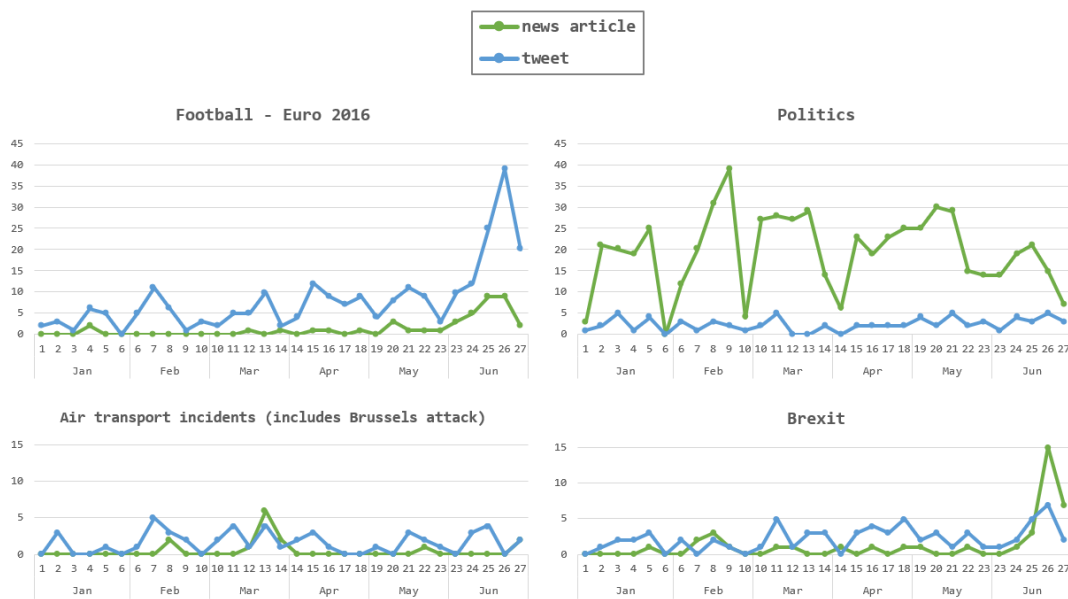


Figure 3: Evolution of the number of elements.

and tweets at the expected moment: the European Football Cup started on the 10th of June (week 24) and the Brexit Referendum took place on the 23rd of June (week 26). Naturally, both of these events were subject to discussion in the previous months, as the National team prepared for the competition and debates concerning Brexit and its consequences intensified. Tweets assigned to *Football - Euro 2016* always surpassed the number of published articles on a weekly basis, with a global ratio of six tweets to one news article. This highlights the importance of this event (Euro 2016) and topic (football) on the Portuguese discussions on Twitter. These time series show a correlation of 0.84, which may indicate that football is referred to with the same intensity in the news and on Twitter. A smoother trend of tweets surpassing the number of articles is noticed for *Brexit*, with the exception of week 26, when the referendum occurred, where the number of assigned tweets is approximately 50% lower than the clustered news articles.

*Air transport incidents* (includes Brussels attack) is the smallest cluster under observation, albeit having scored the highest F1 value. It shows a small rise at week 13, which was when the bombings in Brussels Airport and Maalbeek metro station happened. We remark, nevertheless, that if we included shared articles and tweets with link to them in this analysis, the rise at week 13 would be significantly larger (18 tweets and 10 articles versus an average of 0.2 and 1.6 in the weeks prior to this event).

*Politics*, the largest cluster of articles and news,

shows a rather smooth evolution in the number of elements, especially on the tweets side. It shows that for the kind of study further partitioning of this cluster would be desirable in order to identify patterns in texts possibly different from the ones revealed. This is the cluster with the lowest ratio of assigned tweets to articles, which may also be a sign that for this topic, the keywords generated at the articles level may not be sufficiently discriminative at the tweets level. The evaluation measures did, indeed, reveal a large proportion of false negatives for this class. Another possible line of interpretation is that Portuguese Twitter users do not, in fact, talk as much about politics when compared to its importance to the press.

#### 4.1 Time-wise Differences

One way of analyzing the temporality between news and tweets is to consider the time difference of every article in a cluster to its *median tweet*. The median tweet is the tweet with median time in the corresponding cluster. This brings out a question of how the press publication timings compare to the moment when the public discussion is at its highest. The time difference is computed as a lag variable, that, if positive, indicates that the article is older than the median tweet, and the opposite if negative. If the distribution of this variable is skewed to the right, the stories of that cluster have a tendency to be first published by the press; if skewed to the left, the social media discussion happens sooner than the news. Fig. 4 shows the representation of the distribution of this lag vari-

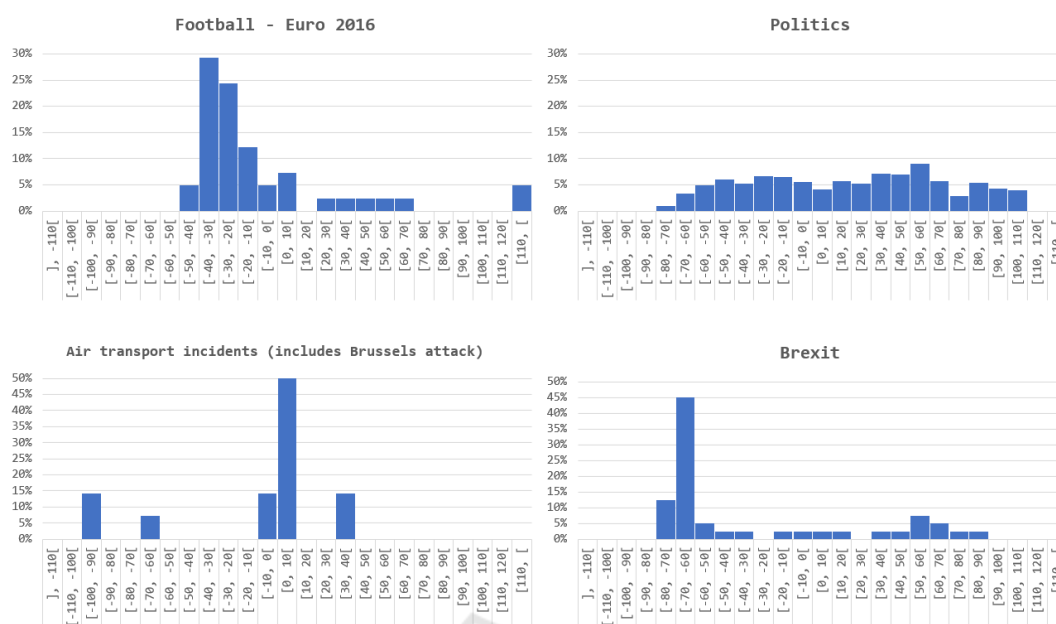


Figure 4: Days difference between articles and the median tweet.

able for the considered subset of clusters.

It can be observed that the majority of the articles belonging to the cluster *Football - Euro 2016* were published after the median date of the tweets assigned to this cluster (a negative days difference). Portuguese Twitter users seem, therefore, to anticipate the discussion of the national team participation in the competition in comparison to what happened in the news. A similar conclusion can be drawn for *Brexit*: the height of discussion of Portuguese users of the staying or leaving of the UK from the European Union happened on Twitter about two months before it happened in the press.

These observations lead to the following conclusion: while the news on *Football - Euro 2016* and *Brexit* were more event-oriented, with peaks of articles at specific points or short periods of time (e.g.: the start of the football competition on the 10th of June; the referendum date announcement in February and the referendum itself on the 23rd of June), exchange of tweets has happened more evenly distributed during the period under study.

The *Politics* cluster does not present any specific pattern. The *Air transport incidents* cluster shows some signs that the press published articles about the incidents first — in total there were nine articles published before the median tweet (positive days difference) and five articles published after the median tweet (negative days difference).

We also did this temporality analysis for the whole dataset, i.e., including shared articles and tweets with

links to them. Conclusions for the selected clusters have not changed.

## 5 CONCLUSIONS

This work aimed at providing some insights into the temporal relationship of texts that appear both as news articles and in social networks in Portugal. The strategy was to investigate what were the main themes published and how they behaved in terms of the number of articles and social media posts during a reasonably long period.

To that end, a sample of online news articles from generalist news sources was subject to text clustering techniques, and 50 groups of similar articles were identified. These were subsequently labelled with keywords. The groups included articles on football related events and teams, terrorist attacks and other international incidents, elections, accidents, investigations, political parties, ministerial actions, economic/financial reports and weather warnings, among others.

Then, to associate tweets with the groups of news articles, we used a sample of tweets and assigned them to the clusters using a similarity measure. This assignment was evaluated using tweets with a news article URL. The fact that tweets contain very few terms makes this task rather difficult. Also, as the number of classes is large, we cannot expect a very high accuracy. The default accuracy for 50 classes

with equal distribution of examples in training would be 1/50, i.e. only 2%. The accuracy achieved was 12.7%, that is substantially higher than the default.

Four clusters for which the evaluation was considerably above average, were chosen for the subsequent study of the temporality between news and tweets. These were *Brexit*, *Football - Euro 2016*, *Air transport incidents* and *Politics*. Regarding *Football - Euro 2016* the football national team was a rather constant subject of discussion on social media in the first semester of 2016, culminating in the final month with its participation in the European Cup. However, the same did not happen on the news side, where the most frequent articles were published towards the end of the period under observation. A similar pattern was observed for *Brexit*. This indicates that, for some texts, the press is more event-oriented, contrasting with the more permanent focus of Twitter users. The analysis *Air transport incidents*, which included Brussels bombings, revealed that the press had a more prominent role in the news diffusion, while comments in social media appeared afterwards. Regarding the cluster *Politics*, no pattern was identified. Perhaps, with a more refined level of partitioning certain timing patterns would be more easily identified.

Regarding future work, we believe that the results could be improved by adopting ontologies that would enable to compute semantic distances between articles and tweets. Besides, tweets could also be expanded with the use of synonyms or through word embeddings (Mikolov et al., 2013). This would enhance the matching process of tweets to the lexicon obtained from the articles.

## ACKNOWLEDGMENT

This work has been supported by the project Centro-01-0145-FEDER-000019 - C4 – Cloud Computing Competences Centre” cofinanced, through the Support System for Scientific and Technological Research - Integrated SR&TD Programs, by the Portugal 2020 Program (PT 2020), in the framework of the Regional Operational Program of the Center (CENTRO 2020) and by the European Union through the European Regional Development Fund (ERDF).

## REFERENCES

- Bholowalia, P. and Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, 105(9):17–24.
- Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *SIGIR 92*, volume 51, pages 318–329, Copenhagen, Denmark. ACM.
- DVJ Insights and ING The Netherlands (2015). Impact of social media on news (#SMING15). Technical report.
- Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., and Ma, K.-L. (2012). Breaking news on twitter. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, pages 275–279, Austin, Texas, USA. ACM.
- Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media? In *WWW '10 Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, Raleigh, North Carolina, USA. ACM.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Newman, N. (2009). The rise of social media and its impact on mainstream journalism. Technical Report September, Reuters Institute for the Study of Journalism, Department of Politics and International Relations, University of Oxford.
- Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in Twitter. In *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2010*, pages 120–123, Toronto, ON, Canada. IEEE.
- Rocha, C., Jorge, A., Sionara, R., Brito, P., Pimenta, C., and Rezende, S. (2016). PAMPO: using pattern matching and pos-tagging for effective Named Entities recognition in Portuguese. *arXiv preprint arXiv:1612.09535*, pages 1–17.
- Saleiro, P., Amir, S., Silva, M., and Soares, C. (2015). POPmine: Tracking Political Opinion on the Web. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM)*, pages 1521–1526, Liverpool, UK. IEEE.
- Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, pages 42–51, Seattle, Washington, USA. ACM.
- Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Evaluation in information retrieval*, volume 39. Cambridge University Press, New York, USA, 1st edition.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-p., Yan, H., and Li, X. (2011). Comparing Twitter and Traditional Media using Topic Models. In *Proceedings of the 33rd European conference on Advances in information retrieval (ECIR'11)*, pages 338–349. Springer, Berlin, Heidelberg.