http://dx.doi.org/10.1108/IJWIS-11-2017-0081

A Framework for Intelligent Twitter Data Analysis with Nonnegative Matrix Factorization

G. Casalino *[1,3], C. Castiello[1,3], N. Del Buono[2,3], and C. Mencar[1,3]

[1]Department of Informatics, University of Bari Aldo Moro, Italy
[2]Department of Mathematics, University of Bari Aldo Moro, Italy
[3]Member of INDAM Research Group GNCS

Abstract

Purpose In this paper we propose a framework for intelligent analysis of Twitter data. The purpose of the framework is to allow users to explore a collection of tweets by extracting topics with semantic relevance. In this way, it is possible to detect groups of tweets related to new technologies, events and other topics that are automatically discovered.

Methodology The framework is based on a three-stage process. The first stage is devoted to dataset creation by transforming a collection of tweets in a dataset according to the Vector Space Model. The second stage, which is the core of the framework, is centered on the use of Nonnegative Matrix Factorizations (NMF) for extracting human-interpretable topics from tweets that are eventually clustered. The number of topics can be user-defined or can be discovered automatically by applying Subtractive Clustering as a preliminary step before factorization. Cluster analysis and word-cloud visualization are used in the last stage to enable intelligent data analysis.

Findings We applied the framework to a case study of three collections of Italian tweets both with manual and automatic selection of the number of topics. Given the high sparsity of Twitter data, we also investigated the influence of different initializations mechanisms for NMF on the factorization results. Numerical comparisons confirm that NMF could be used for clustering as it is comparable to classical clustering techniques such as spherical k-means. Visual inspection of the word-clouds allowed a qualitative assessment of the results that confirmed the expected outcomes.

*corresponding author, gabriella.casalino@uniba.it

Originality/value The proposed framework enables a collaborative approach between users and computers for an intelligent analysis of Twitter data. Users are faced with interpretable descriptions of tweet clusters, which can be interactively refined with few adjustable parameters. The resulting clusters can be used for intelligent selection of tweets, as well as for further analytics concerning the impact of products, events, etc. in the social network.

1 Introduction

The amount of data available on-line has grown tremendously over the past decades. According to a recent Cisco's survey, the annual global IP traffic will reach about 3.3 ZB (zettabyte, i.e. 10^{21} bytes) per year by 2021. This number appears even more amazing if we consider that in 2016 the annual run rate for global IP traffic was 1.2 ZB per year[1]. Analyzing and extracting information from such data is one of today's biggest challenges. Without proper analysis tools, in fact, it is as though the data does not exist at all (Liu and Motoda, 2007).

On one hand, automatic tools for data analysis are a necessity when facing big volumes of data; on the other hand, when huge amounts of data are involved, it is easy to find correlations that may not be related in a causal way. The right balance is a *collaborative* approach, where automatic mechanisms assist humans in extracting and interpreting useful information. This is the ultimate scope of Intelligent Data Analysis (IDA) as an iterative and interactive process that applies computational methods to understand data, refine questions, and cycling the steps until a satisfactory answer is eventually obtained (Berthold and Hand, 1999; Berthold, Borgelt, Höppner and Klawonn, 2010).

Among the different categories of IDA methods (Holmes and Peek, 2007), we focus on data exploration, concerning the generation of hypotheses from data. Analysts look at data to discover relations among features, trends, anomalies, or outliers in values, as well as relations among features and classes. Most of these techniques use visual tools to represent information. Also, quite often IDA methods incorporate a-priori expert knowledge to allow user interaction for effective data exploration (Casalino, Del Buono and Mencar, 2016).

In this study, we turn our attention to Twitter data. Twitter[2] is a widely used social network which allows millions of users to share short, 140-character messages called tweets[3]. It has been estimated that 500 millions of tweets are produced per day[4]. Tweets roughly correspond to thoughts, ideas, commentaries, short discussions on various topics, personal opinions and comments

[1] *The Zettabyte Era: Trends and Analysis.* Cisco White Paper, `https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html`

[2] `twitter.com`

[3] Twitter is rolling out 280-character tweets to all users except those who tweet in Japanese, Korean and Chinese.

[4] `https://www.omnicoreagency.com/twitter-statistics/`

on several matters and life events. Tweets are an indisputable source of unstructured textual data that are worth to be investigated for either social or commercial purposes (Pak and Paroubek, 2010).

Text processing mechanisms are usually adopted to transform a collection of tweets in a structured source of information which subsequently undergoes some kind of investigations. Both keywords and topic extraction mechanisms can be used as tweet mining tools, but topic extraction enables intelligent document analysis since it allows to classify documents according to their semantic categories. Some topic extraction mechanisms for Twitter have been built to identify and characterize communities (Gupta, Joshi and Kumaraguru, 2012), detect opinion tendency into specific topics (Guo, Zhang, Tan and Guo, 2012), discover user behaviors (Jin, Chen, Wang, Hui and Vasilakos, 2013), understand political inclinations (Shamma, Kennedy and Churchill, 2009; Wong, Tan, Sen and Chiang, 2016), real-time traffic events detection (D'Andrea, Ducange, Lazzerini and Marcelloni, 2015; Ducange, Mannar, Marcelloni, Pecori and Vecchio, 2017). Both in text mining and topic extraction contexts, dimensionality reduction mechanisms – designed to represent data in a reduced space through feature selection and extraction – assume a key role in managing, understanding, and visualizing data. Particularly, Nonnegative Matrix Factorizations (NMF) distinguish from other traditional dimensionality reduction algorithms since they uncover latent low-dimensional structures intrinsic in high-dimensional data and provide a nonnegative, part-based, representation of data enhancing meaningful interpretations of mined information (Alonso, Castiello and Mencar, 2015). The understandability of the results coming from NMF motivates their success in several areas such as bioinformatics, pattern recognition,image analysis, educational data mining and document clustering (Casalino, Del Buono and Mencar, 2014a; Cichocki, Zdunek, Phan and Amari, 2009; Del Buono, Esposito, Fumarola, Boccarelli and Coluccia, 2016; Casalino and Gillis, 2017; Casalino, Castiello, Buono, Esposito and Mencar, 2017), as well as the importance this computational model assumes in IDA (Casalino et al., 2016).

In this paper, we present a framework based on NMF designed to provide an intelligent analysis of Twitter data. The proposed experimental framework aims to standardize the technical steps needed for realizing pattern discovery through NMF methods when Twitter datasets are investigated. As an afterthought, we want to point out the advantages coming from the application of NMF in the context of IDA. By exploiting the nonnegativity property of NMF, in fact, it is possible to derive a kind of factorization which finds an immediate and intuitive interpretation in terms of topics underlying the Twitter data.

The present paper is an extended version of the one presented at the 17th International Conference on Computational Science and Its Applications (ICCSA 2017) (Casalino, Castiello, Del Buono and Mencar, 2017). The main differences of this extended paper consist in:

i. The enrichment of the framework with an additional algorithm for generating initialization for NMF, based on a modified version of Subtracting Clustering. This enables the suggestion of the most appropriate number

of topics to be mined from the collection of Twitter data.

ii. The enhancement of the experimental session which has been extended by considering the ensemble of NMF algorithms incorporated in our framework.

The illustrated case studies witness the effectiveness and the efficiency of the proposed techniques: the results obtained by different combinations of initialization and NMF algorithms have been compared with other traditional clustering algorithms, such as spherical k-means.

The rest of the paper is organized as follows: Section 2 introduces some concepts related to the model employed to translate tweets from their unstructured form into the tweet-term matrix and NMF. The section also includes a brief review of different ways to apply NMF in the field of Twitter data analysis. Section 3 describes the main steps assembling the proposed framework. Section 4 is devoted to a detailed presentation of a case study. Particularly, the framework is used on some newly collected Twitter datasets and its effectiveness in extracting interpretable topics from Twitter data is discussed. The paper ends with some final remarks concerning future research work.

2 Related works

Before being analyzed with any automatic learning mechanisms, social data like tweets need to be collected, pre-processed and then transformed into a more structured format. Tweets are generally short textual messages limited to 140 characters which can be treated as simple textual document.

The Vector Space Model (VSM) (Salton, Wong and Yang, 1975) is among the most employed models to manage text data. In VSM, a *term-document matrix* is built up where documents and terms are represented in columns and rows, respectively. Each term corresponds to a basis in a highly dimensional vector space (being the overall dimension related to the total number of terms), and each element in the matrix can be intended as a weight of a term inside the corresponding document. The VSM provides a useful way to transform unstructured Twitter data into structured data: given a collection of m tweets, it can be encoded into a term-tweet sparse matrix $X \in \mathbb{R}_{+}^{n \times m}$, whose rows are n terms in a selected vocabulary V and whose columns relate to m tweets. Once this matrix is compiled, it can be processed by automatic learning mechanisms for extracting topics from data (where a topic can be intended as a concept associated to a set of terms that are semantically related).

Classical Latent Semantic Analysis based on the Singular Value Decomposition (SVD) (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990) has been successfully used as a way to realize topic extraction in text applications. However, negative values appearing in such decompositions are difficult to interpret, being sometimes counterintuitive. These drawbacks can be overcome by adopting a NMF approach (Lee and Seung, 1999; Gillis, 2014; Xu, Liu and Gong, 2003; Cichocki et al., 2009). NMF is a dimensionality reduction technique

which decomposes a matrix X into two low-rank factor matrices $W \in \mathbb{R}_+^{n \times k}$ and $H \in \mathbb{R}_+^{k \times m}$ (with rank-factor $k < \min(n, m)$) constrained to have only nonnegative elements and such that $X \approx WH$. The rank k is a user-defined parameter. If X is a term-tweet matrix (in the way it has been introduced before), k defines the number of latent tweet topics to be considered in X, thus providing a semantics for the tweet vector space. Hence, each tweet (namely, a column X_j of the term-tweet matrix X) can be represented as a weighted combination of the columns w_i of the matrix W:

$$X_j \approx h_{1j} w_1 + h_{2j} w_2 + \ldots + h_{kj} w_k, \tag{1}$$

being h_{ij} the elements of the matrix H. It should be observed that NMF factors are not unique. In fact, given a nonnegative pair (W, H) approximating X as in (1), there might exist many equivalent solutions $(WQ, Q^{-1}H)$ for matrices Q with WQ and $Q^{-1}H$ nonnegative matrices. Such transformations lead to different interpretations. To obtain more well-posed NMF pairs different approaches based on the incorporation of additional constraints (such as sparsity and orthogonality) into the NMF factors can be used (Gillis, 2014). Pre-processing and data normalization can also be of some use (Gillis, 2012). Here, we normalize the column vectors of both W and H in L_2 to make the factorization irrespective of data rescaling.

Nonnegativity constraint of the NMF factors allows to interpret equation (1) in terms of topic extraction process (Xu et al., 2003; Kuang, Park and Choo, 2015). In fact, the columns w_i of W stand as the hidden topics embedded into the vector space describing the tweets, whereas each value w_{li} of W expresses the weight of the l-th term to define the semantics of the i-th topic. Obviously, higher weight values correspond to greater degrees of importance associated to the l-th term in defining the hidden topic. To provide a readable interpretation of the topics, for each of them it is possible to consider a subset of terms (in practice, the terms are firstly ranked on the basis of their associated values w_{li}, then the topmost r terms are selected). In this way, the analyst is able to tag each topic with a meaningful label defined through the analysis of the selected terms. The elements h_{ij} of H represent the degree to which each tweet belongs to each topic: if the value h_{ij} is very small, then the corresponding topic is useless in describing that particular tweet. Under some hypotheses (Xu et al., 2003; Chen, Wang and Dong, 2010; Ding, He and Simon, 2005), the topics w_i can be interpreted as prototypes of data clusters, and the elements h_{ij} can be therefore assumed as membership degrees of each tweet to each cluster. Figure 1 illustrates the topic extraction process obtained through a NMF decomposition of a 6×8 term-tweet sample matrix. The topmost three terms have been selected from each column of W.

In the Twitter data analysis scenario, NMF have been used to analyze Twitter networks so as to capture trends (Kim, Seo, Ha, Lim and Yoon, 2013; Pei, Chakraborty and Sycara, 2015), to learn topics from correlation data of terms derived from short texts (Yan, Guo, Liu, Cheng and Wang, n.d.), for emotion detection from text written in Indonesian language (Arifin, Sari, Ratnasari and

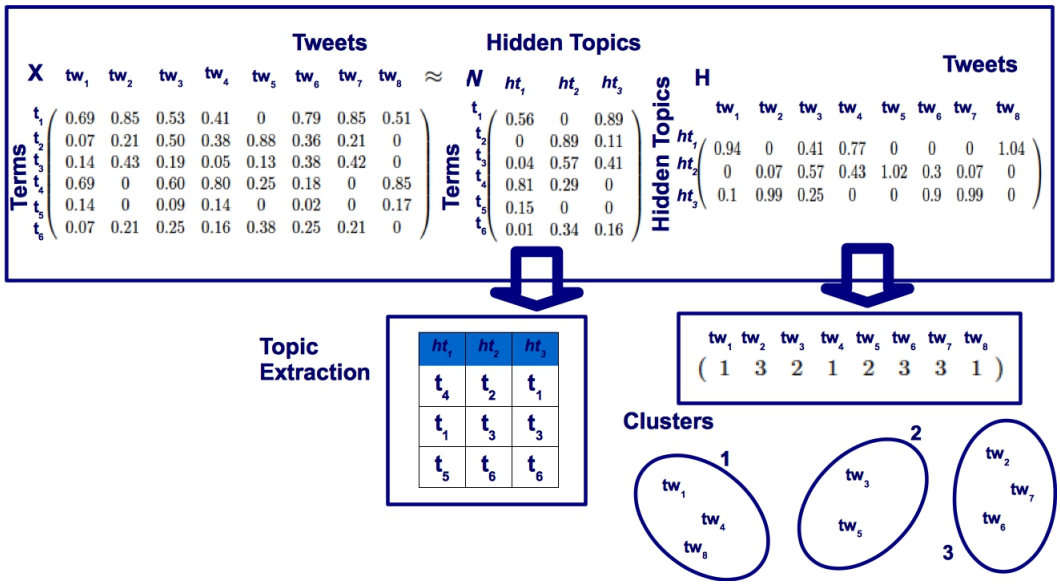


Figure 1: Example of topic extraction with NMF. The term-tweet matrix X is decomposed in the term-topic (W) and topic-tweet (H) factors. Each column of W stands as a topic and can be represented by the r terms with highest height. Tweets can be clustered by assigning the topic with highest membership degree to each tweet.

Mutrofinn, 2014), or to unveil political opinions (Mankad and Michailidis, 2015). Several works have been also proposed to modeling the evolution of topics so as to aid a fast discovery of emerging themes in streaming social media content (Saha and Sindhwani, 2012; Lai, Moyer, Yuan, Fox, Hunter, Bertozzi and Brantingham, 2016; Panisson, Gauvin, Quaggiotto and Cattuto, 2014; Saito, Hirata, Sasahara and Suzuki, 2015; Atsuho, 2017; Shin, Choi, Choi, Langevin, Bethune, Horne, Kronenfeld, Kannan, Drake, Park and Choo, 2017).

NMF proved to be faster than the classical k-means algorithm and yielded more easily interpretable results when mining Twitter data from World Cup Tweets (Godfrey, Johns, Sadek, Meyer and Race, 2014). Also, NMF demonstrated very good performance over other several clustering algorithms when used to analyze Twitter data (Klinczak and Kaestner, 2015; Klinczak and Kaestner, 2016; Ibrahim, Elbagoury, Kamel and Karray, 2017).

Ensemble methods for topic modeling, based on NMF have been proposed to reach stable solutions (Belford, Namee and Greene, 2016; Suh, Choo, Lee and Reddy, 2016; Suh, Choo, Lee and Reddy, 2017). Geo-tagged tweets analysis allows urban monitoring, as urban areas are classified into representative groups (Wakamiya, Lee, Kawai and Sumiya, 2015; Sitorus, Murfi, Nurrohmah and Akbar, 2017). A hashtag recommendation system based on user's usage history and independent from tweets' contents has been proposed by Alvari

(2017). Topic modeling capabilities of Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have been compared by Suri and Roy (2017): the empirical results showed that both the algorithms perform well in detecting topics from text streams. NMF have been also used for Microblog retrieval (Li, Yang and Fan, 2015), topic sense induction and disambiguation on social tags (Iskandar, 2017) and hierarchical clustering (Duong-Trung, Schilling and Schmidt-Thieme, 2017).

3 Twitter data analysis framework

The adopted framework for intelligent analysis of Twitter data is made of three main stages for data creation, NMF decomposition and final data analysis (Casalino, Castiello, Del Buono and Mencar, 2017). The framework is able to collect data from Twitter using specific search criteria, then it appropriately organizes the collected tweets in the term-tweet matrix X using the VSM approach. Once this data matrix is constructed, the core of the process is triggered to factorize X into two nonnegative matrices W and H using some NMF algorithms.

In order to promote intelligent data analysis, the proposed framework allows the selection of different NMF algorithms to inject a-priori knowledge in the factorization process (Casalino et al., 2016). Moreover, different mechanisms for the initialization phase of NMF algorithms are also included into the framework (Casalino, Del Buono and Mencar, 2014b). The obtained factor matrices W and H are then exploited to cluster original tweets into a selected number k of topics (being k the rank of the factorization). The framework integrates also some word clouds visualization tools to allow an easier interpretation of the topic extraction results.

The choice of the rank k is crucial for the quality of the results, since it defines the number of clusters and the hidden topics NMF extracts from the data matrix. The original framework proposed by Casalino, Castiello, Del Buono and Mencar (2017) is here expanded to include a peculiar initialization method for NMF based on Subtractive Clustering (Casalino et al., 2014b), which is able to suggest a suitable number of clusters for a given dataset and to provide better initialization for NMF algorithms w.r.t. classical approaches.

Figure 2 sketches the main modules constituting the tweet data analysis framework, i.e. Dataset Creation, NMF Decomposition and Data Analysis, which are described in the following. All the activities in each module are performed sequentially.

1. **Dataset Creation**. This module conducts all the activities related to collect tweets, pre-process them and finally represent the extracted dataset in a structured matrix form. The output of Dataset Creation module is a term-tweet nonnegative real matrix X of proper dimensions, which relates each tweet with a collection of terms belonging to an automatically extracted vocabulary V in accordance with the VSM. More precisely, the tasks performed in this module are described as follows.

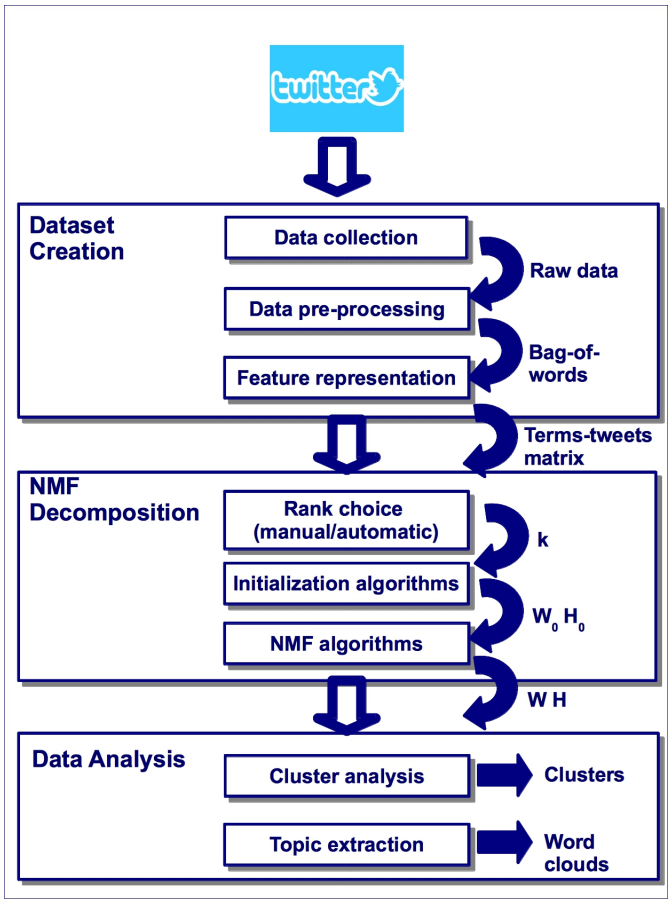


Figure 2: Framework of Twitter data analysis framework based on NMF.

Data collection. A set of m tweets are collected from Twitter through the API[5] (Application Program Interface) on the basis of user-defined keyword search criteria. The result of these operations is a collection of "raw" tweets which have to undertake some pre-processing before being definitely represented as a structured dataset.

Data pre-processing. The collected "raw" tweets contain some useless meta-information and additional text which needs to be pre-processed. The pre-processing phase is carried out by the following steps:

(a) *Meta-information removal.* All the re-tweets[6], URLs, "emojis",

[5] `https://dev.twitter.com/apps`

[6] Tweets that a user received in her stream and shared to her followers.

mentions[7] to other users, as well as any non-alphabetical and numerical characters, are removed;

(b) *Tokenization.* Each tweet is represented by a sequence of tokens (i.e. words in the sense of the "bag-of-words" VSM model).

(c) *Normalization.* The sequence of tokens are normalized to a limited character set, i.e. $[\mathtt{a} - \mathtt{z}]$.

(d) *Stop-word filtering.* Text elements such as articles, conjunctions, prepositions, pronouns are deleted; both English and Italian stop-word lists are considered.

(e) *Stemming.* Each word is reduced to its root form by a standard stemming algorithm.

The output of the pre-processing phase is a set of terms of the vocabulary V which is used the derive the term-tweet matrix.

Matrix representation. According to the VSM, the extracted n terms in the vocabulary V are used to create a structured vector representation of the collected m tweets in order to codify them into the term-tweet matrix $X \in \mathbb{R}_+^{n \times m}$. Each elements x_{ij} represents the "weight" of the i-th term in describing the j-th tweet and it is computed using the tf-idf (term frequency - inverse document frequency) weighting function. The tf-idf value of the i-th term into the j-th tweet is given by:

$$\text{tf} - \text{idf}(i, j) = \text{tf}(i, j) \times \log(|m|/\text{df}(i)),$$

where $\text{tf}(i, j)$ is the frequency of the term i in the tweet j and $\text{df}(i)$ is the number of tweets in which the term i appears.

2. **NMF Decomposition**. This module is the core of the framework and it is responsible of the factorization of the data matrix X obtained as output from the Dataset Creation module.

   From a computational viewpoint, NMF can be carried out through a number of algorithms (Berry, Browne, Langville, Pauca and Plemmons, 2007). We have included into the framework a selection of NMF algorithms which can be selected by the user. They are:

   - Multiplicative *NMF* algorithm, based on the Euclidean distance which is considered the baseline method for NMF (Lee and Seung, 1999; Lee and Seung, 2001);
   - Alternating Nonnegative Least Squares Projected Gradient (*ALS*) (Lin, 2007);
   - Sparse Nonnegative Matrix Factorization (*SNMF*), which is able to control the sparsity of the factors W and H (Kim and Park, 2007)[8];

[7] Text beginning with the symbol '@' followed by any unique user name.

[8] In the experiments we have set the sparsity of the matrices W and H as 0.7 and 0.3 respectively.

- Nonsmooth Nonnegative Matrix Factorization (*NSNMF*), which is able to extract highly localized patterns in data, forcing the global sparseness of the factors W and H (Pascual-Montano, Carazo, Kochi, Lehmann and Pascual-Marqui, 2006). [9]

All NMF algorithms are iterative mechanisms, hence they require some initial matrices as starting point. The initialization phase is critical for the quality of the final results of NMF decomposition and different initialization algorithms lead to different solutions of NMF. As a consequence, a throughout experimental analysis is required to choose the correct initialization scheme for the problem at hand.

Among different initialization mechanisms proposed in literature (Casalino et al., 2014b; Sauwen, Acou, Bharath, Sima, Veraart, Maes, Himmelreich, Achten and Van Huffel, 2017), in our framework we included:

- three different random initialization algorithms (which require low computational costs, but usually generate poor informative initial matrices), namely RAND, RAND_C and RAND_VCOL initialization (Albright, Cox, Duling, Langville and Meyer, 2006),
- NNDSVD initialization (Boutsidis and Gallopoulos, 2008), which is a deterministic initialization mechanism (though it is more computationally expensive than random methods),
- Subtracting Clustering initialization, which was recently proposed by Casalino et al. (2014b) as a new initialization method for NMF when data possess special meaning as in document clustering. This initialization method is able to automatically discover the rank k by fuzzily grouping data according to their Euclidean distance. It can be considered as a new strategy for solving the choice of the most appropriate rank factor k for each given dataset.

3. **Data Analysis**. This module performs topic extraction, interpretation and tweet clustering employing the matrix factors W and H given by the NMF Decomposition module. Some appropriate graphical tools are integrated in this module to effectively visualize clusters and display the semantic of the extracted clusters to users.

   Topic extraction and tweets clustering is exemplified in Figure 1, while cluster visualization is performed using a word-cloud representation mechanism (it shows selected words using different font sizes: the more a word is important in a tweet, the bigger and bolder it appears in the word cloud).

   Each tweet exhibits multiple topics with different relevance. Through NMF it is possible to suggest the importance of each topic in each tweet. The encoding matrix H maps the hidden topics (rows of H) with the tweets (columns of H), and elements h_{ij} indicate the importance (weight)

[9] In the experiments we have set the degree of nonsmoothing as 0.3.

that the i-th topic has in the j-th tweet. In the example in Figure 1, the first tweet tw_1 is about the hidden topics ht_1 and ht_3 with weights 0.94, and 0.1, respectively, while it does not refer to the topic ht_2.

Each tweet is represented as a vector in the sub-space spanned by th vectors w_j and hard document clustering can be obtained by assigning the tweets to the nearest basis in the space (Xu et al., 2003; Shahnaz, Berry, Pauca and Plemmons, 2006). This is equivalent to assigning each tweet to the topic with the highest weight in the column of H. Referring to Figure 1, tweets tw_1, tw_4, tw_8 are assigned to the first cluster (whose semantic is mostly derived by terms t_4, t_1 and t_5); tweets tw_3 and tw_5 are assigned to the second cluster, while tweets tw_2, tw_6 tw_5 are assigned to the third cluster.

It should be pointed out that topics are automatically discovered by analyzing the original tweets since they usually are not known in advance but are learned from data.

3.1 User and automatic rank selection

As previously observed, k is crucial parameter into the proposed framework. In fact, it specifies the low-rank dimension of the factor matrices W and H which approximate the term-tweet matrix X and defines the number of clusters and the hidden topics to be extracted from it. The proposed framework has been enlarged with the possibility of selecting k either as user defined parameter or in an automatic way. The automatic selection tool for k is obtained by adding the Subtracting Clustering initialization into the NMF decomposition module. This adjoint component allows to inject a-priori knowledge in the factorization process and could be of aid for users that are not able to manually provide any particular value of k; this is especially useful in real-world applications where no information about a *ground truth* is available.

The initialization algorithm based on Subtractive Clustering has been proven to suggest a suitable number of clusters for a given dataset and to provide a more informative initial pair of matrices W_0 and H_0 (Casalino et al., 2014b; Casalino, Del Buono and Mencar, 2011). It works on the basis of two hyper-parameters, namely r_a and r_b, whereas r_a stands as the minimum distance that is acceptable for two samples to belong to different clusters, while the parameter r_b is the minimum distance that is acceptable for two cluster prototypes.

The two parameters r_a and r_b are therefore the hyper-spherical cluster and penalty radius in the data space, respectively, and they can be estimated on the basis of the distances among the tweets in the term-tweet matrix. This choice reflects a stable behavior of the SC scheme and suggests a number of clusters more suitable from an interpretability point of view.

3.2 Implementation details

The modules of the proposed framework have been implemented partly in Matlab (R2014b) and Python 3.5. In particular, we used the following Python libraries:

- TWEEPY[10]: this library allows the direct access to the public stream of tweets, which can be downloaded according to some search criteria;
- NLTK[11]: this library is used to implement all the tweet pre-processing steps (Bird, Klein and Loper, 2009);
- SCIKITLEARN[12]: this library is adopted to compute the tf-idf weights

We used Matlab implementations of NMF initializations and algorithms, and cluster evaluation measures[13].

4 Using the framework: a case study

In this section we illustrate the results obtained by using the proposed framework for intelligent analysis of some Twitter datasets. Two different sets of experiments were conducted to demonstrate the capability of the framework to dealing with both user-defined rank k (corresponding to some a-priori data information) and the rank value automatically provided by Subtractive Clustering for initialization (as described in Section 3.1). Furthermore, all the experiments aimed to numerically compare the influence of different NMF initializations and algorithms on the clustering results and on the semantic meaning of the topics extracted from the collected Twitter data. We repeated each experimental session 10 times in order to smooth out random effects. For each experimental session, we retained the run providing the lowest reconstruction error. All the experiments have been run on a machine equipped with an Intel Core 2 Duo 2.40 GHz, 8 GB of RAM.

Three different datasets of Italian tweets were acquired and transformed into the corresponding term-tweet matrices using the *Dataset Creation* module. Four groups of tweets were acquired using, as search criterion, the presence of four Italian keywords for each group as showed in Table 1; Table 2 reports some elementary statistics on data. To better investigate the capability of NMF in topic extraction, very general meaning keywords were used to select tweets. Figure 4 shows the pre-processing steps applied on a tweet acquired by the key-word RELIGIONE, the Italian word for "religion" as illustrated in Figure 3. It should be observed that term-tweet matrices obtained as output of the first

[10] `http://docs.tweepy.org/en/v3.5.0/`

[11] `http://www.nltk.org/py-modindex.html`

[12] `http://scikitlearn.org`

[13] NMI:`https://it.mathworks.com/matlabcentral/fileexchange/29047-normalized-mutual-information` and Silhouette coefficient `https://it.mathworks.com/help/stats/clustering.evaluation.silhouetteevaluation-class.html`

Figure 3: Example of a tweet acquired by the key-word 'religione'. English translation: "This is a holy tree for the #religion and #culture in #Madagascar".

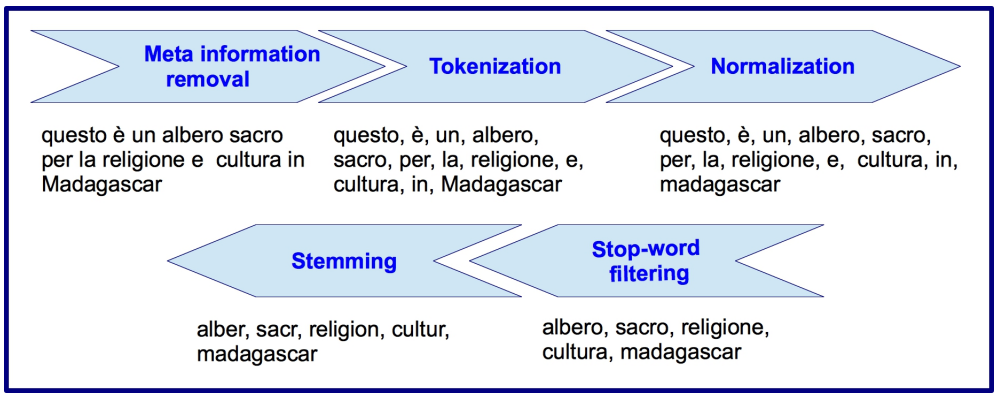


Figure 4: Example of the pre-processing phase of the proposed framework on a tweet.

module present an high degree of sparsity (more than the 99% of the entries are zero).

Both quantitative and qualitative analysis have been performed to evaluate the performances of NMF algorithms and their clustering capabilities. The used evaluation measures are:

- Initial Error. It evaluates the error obtained by approximating the original matrix X with the initial pair W_0,H_0 obtained by an initialization

Dataset	keyword 1	keyword 2	keyword 3	keyword 4
1	RELIGIONE (religion)	TECNOLOGIA (technology)	SCUOLA (school)	AMORE (love)
2	AMORE (love)	SPORT (sport)	VIAGGIO (travel)	MUSICA (music)
3	AMORE (love)	SCUOLA (school)	CLIMA (climate)	CIBO (food)

Table 1: Selected keywords used as research keys for extracting tweets by Data Creation module.

Dataset	#terms	#tweets	sparsity
1	4219	2272	99.81%
2	2840	995	99.74%
3	4350	2312	99.82%

Table 2: Elementary statistics on the three datasets.

algorithm. This error is computed as

$$\frac{\|X - W_0 H_0\|_F}{\|X\|_F}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. This measure has been used to compare initialization methods included into the NMF decomposition module.

- Execution Time. It measures the time (in seconds) needed by the initialization algorithm to construct the initial matrices and NMF algorithms to reach their stopping criterion (number of iterations > 1000 or error reduction $< 10^{-6}$).
- Final Error. It evaluates the approximation error of the final factors W and H in reconstructing the original matrix X and is computes as

$$\frac{\|X - WH\|_F}{\|X\|_F}$$

- Iterations Number. It is the number of iterations required by the algorithm to reach the stopping criterion.
- *Normalized Mutual information (NMI).* It is an external cluster evaluation measure based on entropy, comparing the obtained labeling with the *a-priori* known classes. It is a measure of the mutual dependence between the two groups. It has values in $[0, 1]$, where 0 means no mutual information and 1 perfect correlation.
- *Silhouette coefficient.* It is an internal measure evaluating *cluster cohesion* (i.e. intra-cluster distance) and *separation* (inter-cluster distance) (Rousseeuw, 1987). The coefficient is determined by the average measure of the silhouette value of each point. It has values in $[-1, 1]$ where 1 indicates high separation and cohesion, -1 a wrong number of clusters, and 0 similar inter and intra clusters distances.

Both *NMI* and *Silhouette coefficient* are used to evaluate cluster results. Additionally, to better appreciate the semantics of the hidden topics returned by NMF, each extracted topic (columns w_i of W) has been represented with the topmost 10 terms (ordered accordingly to the weights in the corresponding column of W). These topic can be illustrated using the the *word-clouds* visualization tool included into the *Data Analysis* module.

	Dataset 1		Dataset 2		Dataset 3	
Init. Alg	Init. Err	Time	Init. Err	Time	Init. Err	Time
RAND	192.96	8.67e-4	158.77	5.69e-4	198.17	9.99e+4
RAND_C	2.60	0.07	1.71	0.02	2.71	0.10
RAND_VCOL	2.60	0.16	1.71	0.05	2.71	0.24
NNDSVD	0.98	0.39	0.98	0.17	0.91	0.37

Table 3: Comparisons of the performance of initialization algorithms.

4.1 Results for user-defined rank value

The first experimental session was performed with the factorization rank $k = 4$. This value corresponds to the number of keywords used to acquire the tweets and represents an a-priori knowledge on the semantic categories embedded into the tweets.

Initialization algorithms have been compared to verify whether inexpensive, but less informed algorithms, lead to acceptable results, so that they could be used in place of more informed but computationally expensive algorithms.

Table 3 reports the performances of the initialization algorithms on each dataset. As expected, the simplest random generation of the initial matrices is the fastest but less accurate method. On the contrary, NNDSVD is the slowest, but it has the minimum initial error. NNDSVD requires to compute the truncated SVD which is in fact time consuming compared to the other approaches as witnessed by the remarkable differences in execution time between datasets of different dimensions (namely, datasets 1 and 2 vs. dataset 3). This could represent a problem when dealing with big amounts of data. On the other hand, the semi-informed initialization algorithms RANDOM_C and RANDOM_VCOL give initial error values that are comparable with those provided by NNDSVD, but with a significant reduction of computation time.[14]

Any combination of initialization and NMF algorithm has been evaluated over the three datasets. Tables 3(a)-3(c) report the obtained numerical results. Each pair of initialization-NMF algorithm returns performance values which are comparable, none of them numerically prevails on the others on the considered datasets.

Very accurate solutions may not be the most significant in terms of grouping tweets according to their topics. Table 5 reports the clustering performances obtained by any pair (initialization, NMF algorithm) on the three datasets. Spherical k-means cluster method was also applied as a term of comparison. It should be noted that spherical k-means performs better on Dataset 1, but its results are comparable with those provided by NMF methods on the other datasets. Among the pairs initialization-NMF algorithm, ALS generally gives the best values of NMI.

As an example, Figure 5 shows the word-cloud representation of the four

[14]For a more detailed analysis of the efficiency of initialization algorithms, the interested reader is referred to Casalino et al. (2014b).

(a) Dataset 1

	NMF			ALS			NSNMF			SNMF		
Init.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.
RAND	0.968	123.03	187	0.968	11.21	26	0.97	500.45	580	0.975	259.39	141
RAND_C	0.97	36.12	107	0.968	13.4	32	0.97	500.45	845	0.97	480.34	345
RAND_VCOL	0.97	39.10	65	0.968	15.37	36	0.972	500.43	732	0.97	500.45	580
NNDSVD	0.97	44.42	65	0.968	15.70	36	0.972	500.56	774	0.97	500.45	580

(b) Dataset 2.

	NMF			ALS			NSNMF			SNMF		
Init.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.
RAND	0.973	13.25	61	0.973	13.65	103	0.975	1000	332.65	0.978	52.42	94
RAND_C	0.974	10.43	100	0.97	10.12	100	0.976	281	223	0.978	44.21	101
RAND_VCOL	0.976	10.28	57	0.974	3.17	24	0.978	218.09	1000	0.978	48.72	86
NNDSVD	0.974	3.72	20	0.974	1.70	12	0.975	321.08	1000	0.978	32.05	60

(c) Dataset 3.

	NMF			ALS			NSNMF			SNMF		
Init.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.	Err.	Time	It.
RAND	0.912	300.38	370	0.909	6.92	14	0.918	500.76	485	0.932	813.33	382
RAND_C	0.954	21.4	59	0.909	6.31	17	0.945	502.32	546	0.93	352.7	180
RAND_VCOL	0.971	25.08	38	0.909	7.22	15	0.973	500	716	0.937	312.48	152
NNDSVD	0.909	17.56	19	0.909	5.62	11	0.973	500	716	0.93	217.06	106

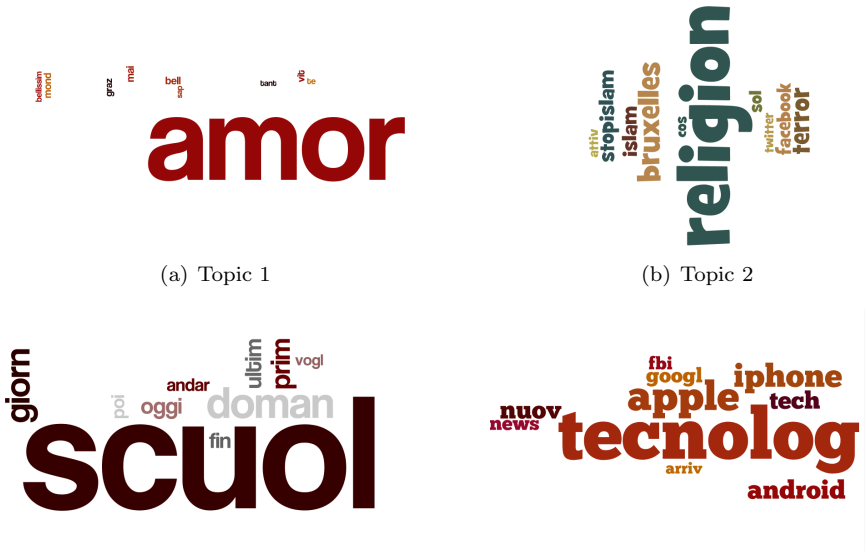
Table 4: Performance of the NMF algorithms initialized with different strategies applied to the three datasets.

Init.-NMF alg.	Dataset 1	Dataset 2	Dataset 3
Rand-NMF	0.701	0.694	0.707
Rand-ALS	0.7	0.696	0.795
Rand-NSNMF	0.653	0.714	0.709
Rand-SNMF	0.598	0.625	0.7
Rand_c-NMF	0.65	0.643	0.689
Rand_c-ALS	0.673	0.651	0.679
Rand_c-NSNMF	0.659	0.701	0.721
Rand_c-SNMF	0.60	0.632	0.703
Rand_vcol-NMF	0.666	0.614	0.639
Rand_vcol-ALS	0.702	0.701	0.795
Rand_vcol-NSNMF	0.653	0.612	0.642
Rand_vcol-SNMF	0.653	0.627	0.507
NNDSVD-NMF	0.666	0.698	0.798
NNDSVD-ALS	0.702	0.701	0.795
NNDSVD-NSNMF	0.653	0.712	0.642
NNDSVD-SNMF	0.653	0.633	0.547
spherical k-means	0.832	0.696	0.718

Table 5: Cluster performance of NMF and initialization algorithms in terms of NMI.

hidden topics extracted from the first dataset by NSNMF algorithm initialized with the NNDSVD. The tweets are grouped in four clusters (in accordance with the rank value $k = 4$) and depicted as four word-clouds of ten terms with the highest weight. As it can be observed, the main terms in each cloud are exactly the Italian (stemmed) keywords used to acquire the tweets (Love, School, Religion, Technology). This confirms that NMF was been able to correctly capture the hidden meaning in the tweets. Furthermore, the terms appearing in each word-cloud are semantically correlated. For instance, taking into account the (stemmed) term RELIGION, it is grouped together with the Italian words TERROR, BRUXELLES, ISLAM and the tag STOPISLAM (figure 5(b)). Even if these terms do not strictly define the concept of *religion*, it should be observed that Twitter data are strictly related to the temporal instants they are acquired, reflecting the current events and the respective people thoughts and feelings. Since the numerical experiments were conducted after the terrorist attacks in Bruxelles (on March 22th, 2016), this explains why those words are grouped together. Similar results can be observed with the (stemmed) keywords TECNOLOG and SCUOL. In particular, the terms connected to TECNOLOG are also related to the the terrorists' facts; in fact in those days the possibility of accessing to confidential information contained in the terrorists' phones was being discussed. That is why the terms IPHON and APPLE have a big weight (i.e. bold font and big size) in the word cloud, but also FBI, though to a lesser account (figure 5(d)). Finally, the terms related to returning to school have been grouped with the keyword SCUOL, because in the days tweets were collected,

the students were coming back to school after Easter holidays (figure 5(c)).



(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 5: Hidden topics obtained with Dataset 1, NNDSVD initialization and NSNMF algorithm.

In conclusion, we also observe that NMF algorithms are able to detect localized patterns in sparse matrices as the term-tweet matrix is. In particular, NSNMF it is able to preserve this sparsity in the factorization process giving more interpretable bases than the other algorithms.

4.2 Automatic selection of the rank factor

Unsupervised learning aims to capture the intrinsic geometry in data. The number of the groups depends on the data structure. We use the Subtractive clustering initialization algorithm to derive suitable factor rank k for each of the three Twitter datasets, and then we compute the cluster results provided by NMF algorithms included into the framework.

The hyper-parameters in the Subtractive Clustering based initialization, that is hyper-sphere cluster r_a and the penalty radius r_b, were estimated on the basis of the distances among the tweets. We varied r_a between the $5th$ and $95th$ percentile of the tweet distance values, while the penalty radius is computed as $r_b = \alpha r_a$, being $\alpha \in [1, 2]$ (Casalino et al., 2014b).

A grid search strategy has been adopted by considering all parameter combinations from the candidate sets and the first value of α stabilizing the number of clusters was selected (as showed in Figure 6). Subsequently, r_a was selected as the value minimizing the initial error with respect to α (as illustrated in Figure 7).

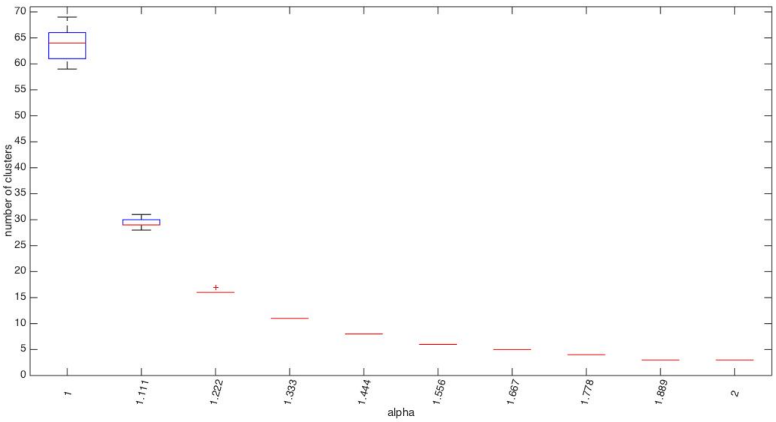


Figure 6: Cluster number variance for different values of the hyper-parameter α, varying r_a in the given ranges, for Dataset 1.

	r_a range	α range	r_a	α	cluster number
Dataset 1	$[1.3651, 1.4142]$	$[1, 2]$	1.3651	1.333	11
Dataset 2	$[1.3740, 1.4142]$	$[1, 2]$	1.374	1.88	5
Dataset 3	$[1.3624, 1.4142]$	$[1, 2]$	1.4053	1.11	3

Table 6: Parameter settings.

Table 6 reports the hyper-parameter settings for the three datasets (that is the candidate sets for the hyper-parameters r_a and α, the chosen values and the suggested number of clusters, respectively). Note that r_a ranges suggest that the tweets are very different each other; indeed the columns of the term-tweet matrices have been normalized in L_2, and the distances among them could vary in $\left[0, \sqrt{2}\right]$. This is a predictable result, due to the intrinsic characteristic of the tweets: few terms from a big vocabulary.

Subtractive Clustering returns a suitable rank for a given dataset together with the initial pair W_0,H_0. In Tables 7 and 8 we compare the performance of NMF when either Subtractive Clustering or NNDSVD are used as initialization algorithm.

Comparing the results reported in Table 7, it should be observed that both initializations algorithms provide comparable results either in terms of reconstruction error and computational effort on the three datasets.

Table 9 reports the quantitative evaluation of the cluster results in terms of *Silhouette Coefficient* Very small values were achieved due to the high distances among the original tweets; however, we observe that the average silhouette values (over the three datasets) of the NMF algorithms initialized with Subtractive Clustering slightly overcome the corresponding values obtained with NNDSVD initialization. Furthermore, it should be observed that the results provided by

	Dataset 1		Dataset 2		Dataset 3	
Init. Alg	Init. Err	Time	Init. Err	Time	Init. Err	Time
SC	1.0	0.8956	1.05	0.1483	1.04	0.3943
NNDSVD	0.96	0.5166	0.97	0.1506	0.91	0.4187

Table 7: Comparisons of the performance of initialization algorithms.

(a) Dataset 1

	NMF			ALS			NSNMF			SNMF		
	Err	Time	It	Err	Time	It	Err	Time	It	Err	Time	It
SC	0.954	224.1	324	0.947	25.83	49	0.958	710.89	1000	0.967	1.87e+3	1000
NNDSVD	0.974	168.57	151	0.970	35.86	55	0.978	501.28	399	0.977	2.28e+3	1000

(b) Dataset 2

	NMF			ALS			NSNMF			SNMF		
	Err	Time	It	Err	Time	It	Err	Time	It	Err	Time	It
SC	0.974	5.09	38	0.970	2.36	13	0.978	328.11	640	0.977	175.02	1000
NNDSVD	0.970	32.38	18	0.970	7.02	34	0.971	397.88	1000	0.976	196.94	273

(c) Dataset 3

	NMF			ALS			NSNMF			SNMF		
	Err	Time	It	Err	Time	It	Err	Time	It	Err	Time	It
SC	0.917	31.16	64	0.915	6.60	11	0.920	774.10	405	0.933	749.63	1000
NNDSVD	0.915	18.09	20	0.915	5.25	7	0.920	500.32	492	0.933	616.84	269

Table 8: Performance of the NMF algorithms initialized with SC and NMF algorithms applied to the three datasets.

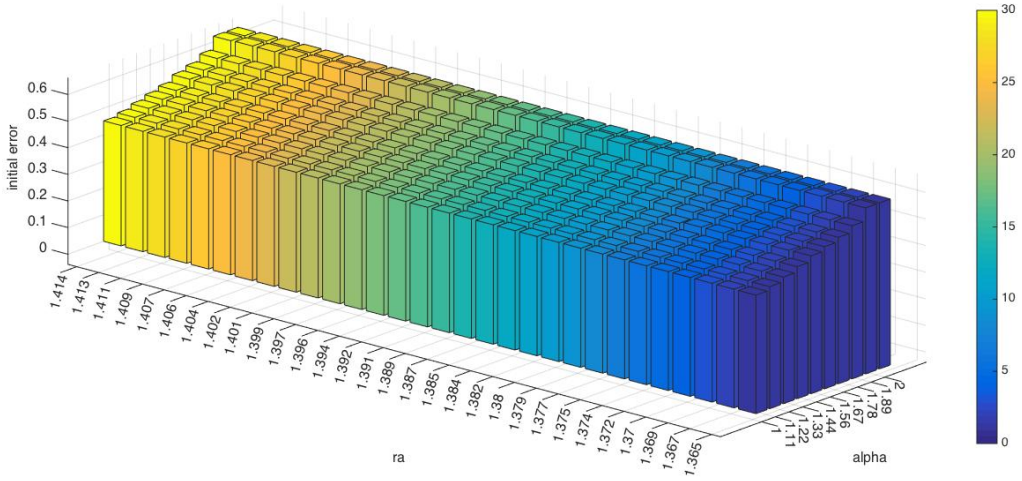


Figure 7: Initial error obtained with the Subtractive clustering initialization method on Dataset 1, varying the hyper-parameters r_a and α in the specified ranges.

NMF algorithms are comparable with those given by the spherical k-means baseline, confirming the applicability of NMF as Twitter data clustering mechanism.

Word-cloud visual tools were used to represent the topic extraction results: in particular, Figure 8 shows the eleven topics extracted from Dataset 1 using the pair (Subtractive Clustering, ALS algorithm).

As it can be observed, Topic 1 (Figure 8(a)) is related to the keyword AMOR (standing for the English "love"). The most important stem *amor* is related with terms as *dolc* (sweet), *bellissim* (beautiful), *vit* (life), *mond* (world) all of these can be associated to the concept of love. Two topics concern the keyword SCUOLA as depicted in figures 8(d)) and 8(j) where the stem SCUOL is more evident. The first one contains words as *student* (student), *piac* (like), *bell* (beautiful), which can be in some way related to the idea that the tweets falling in this topic deal with the "happiness" to go to the school. On the contrary, the second one can be related to the end of Easter holidays when students are not very happy for coming back to school.

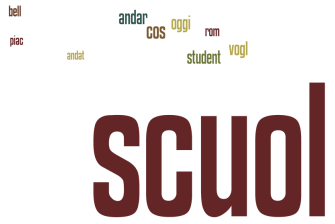
Six separate topics talk about technology (TECNOLOG) from different points of view: safety check available on Facebook after the terror attack in Bruxelles, and in general the activities on the social networks (figure 8(b)); the launch of the new IPhone model and the rumors about the launch of the new smartphone model by Xiaomi with the Android operating system[15] (Figure 8(e)); the death

[15] `https://www.apple.com/apple-events/march-2016/` `http://www.techtimes.com/articles/144899/20160329/xiaomi-mi-5-india-launch-set-for-march-31.htm`

(a) Topic 1

(b) Topic 2

(c) Topic 3



(d) Topic 4

(e) Topic 5

(f) Topic 6

(g) Topic 7

(h) Topic 8

(i) Topic 9

(j) Topic 10

(k) Topic 11

Figure 8: Word-cloud representation of the topics extracted from Dataset 1, with the pair (SC initialization, ALS algorithm).

SC-NMF alg.	Dataset 1	Dataset 2	Dataset 3
SC-NMF	0.0448	0.0288	0.0945
SC-ALS	0.0532	0.0475	0.1197
SC-NSNMF	0.0520	0.0444	0.1197
SC-SNMF	0.0396	0.0274	0.0425
NNDSVD-NMF	0.0407	0.0288	0.0945
NNDSVD-ALS	0.0421	0.0288	0.0945
NNDSVD-NSNMF	0.0410	0.0291	0.0942
NNDSVD-SNMF	0.0369	0.0274	0.0415
spherical k-means	0.0429	0.0206	0.0870

Table 9: Cluster performance of NMF and initialization algorithms in terms of Silhouette.

of the Intel's president Andrew (Andy) Grove[16] (Figure 8(f)); FBI-Apple debate on mobile phone privacy in case of terror attacks[17] (Figure 8(g)); the Apple's night shift function, introduced with the IOS update, to improve sleep quality[18] (Figure 8(i)); the Italian phone provider *Telecom Italia Mobile* that signed an agreement with Google for using the *Google Chromecast* technology on their TV decoders[19](Figure 8(k)).

Moreover, as it can be observed in Figure 8(h), the topic contains all the terms related to the keyword RELIGIONE (also in this case extracted terms reflect the events currently happened). The last topic (depicted in Figure 8(h)) regards the process to the mafia boss *Bernardo Provenzano* that was postponed due to his health conditions: in this case the algorithm mixed this information with the FBI-Apple fight.

Summarizing, this qualitative analysis shows the effectiveness of the NMF algorithms in topic modeling. The algorithms have been able to detect significant topics in the Twitter collection both with a given rank factor and with the suggested one. The difference is on the granularity of the results. The rank suggested by the Subtractive Clustering initialization allows to capture the real structure of the data without forcing the results in any given classes.

5 Final remarks

In this paper we proposed a framework to intelligently analyze Twitter data. These are a particular kind of textual data that are characterized by a small number of terms belonging to a large vocabulary. An automatic mechanism is necessary to pick the most descriptive words in this vocabulary, to aggregate them in the *bag-of-word* representation to form topics, and to group the tweets

[16] `https://newsroom.intel.com/news-releases/andrew-s-grove-1936-2016/`

[17] `https://www.nytimes.com/2016/03/22/technology/apple-fbi-hearing-unlock-iphone.html`

[18] `http://time.com/4269497/iphone-night-shift/`

[19] `https://www.tim.it/tv/nuovo-decoder-timvision`

according to these topics. In this work we use NMF algorithms as a tool for Intelligent Data Analysis. A case study shows the use of the proposed framework for capturing and analyzing tweets. After retrieving tweets according to some search criteria, they are transformed in a structured matrix form, which is suitable for NMF decomposition. Finally, tweets are clustered in groups related to their hidden topics. We verified the effectiveness of the framework by comparing the results obtained with different initialization and NMF algorithms on three datasets obtained by querying the Twitter repository. Moreover we have investigated the appropriate choice of the factorization rank which is connected to the number of clusters that NMF are able to extract. We used the Subtractive Clustering initialization to determine a suitable rank factor for a given dataset. The proposed experimental framework is mainly devoted to standardize the technical steps one has to perform when NMF are applied for topic extraction from Twitter data. Beside different NMF algorithms forming the core of the proposed framework, also some initialization mechanisms are considered in order to allow the user to chose starting matrices for NMF algorithms. In fact, a correct initialization is critical for the quality of the final results of NMF decomposition in an Intelligent Data Analysis context.

Finally, we have compared the NMF cluster results with the spherical k-means clustering algorithm, showing that NMF give comparable results with a better interpretability that is evidenced by the word cloud representation used to visualize the hidden topics discovered in data.

Future work will be addressed to scale the proposed framework to big data contexts. To this pursuit, we have already shown that the use of Subtractive Clustering provides a convenient initialization for NMF in acceptable time (especially when compared with state-of-art methods, like NNDSVD); however, scaling to big data poses technological challenges that require careful design of all the modules included in the model, in order to keep the computational complexity, both in time and space, under acceptable limits. Scaling to big data also calls for novel solutions for clustering in high-dimensional spaces. To this aim, a careful choice of the metrics used to evaluate the similarity of tweets, as well as the selection of the most suitable parameters, become of paramount importance and require an in-depth investigation.

Acknowledgements

This work has been supported in part by the GNCS (*Gruppo Nazionale per il Calcolo Scientifico*) of Istituto Nazionale di Alta Matematica FrancescoSeveri, P.le Aldo Moro, Roma, Italy.

References

Albright, R., Cox, J., Duling, D., Langville, A. and Meyer, C. (2006). Algorithms, initializations, and convergence for the nonnegative matrix factor-

ization, *Technical report*, NCSU Technical Report Math 81706.

Alonso, J. M., Castiello, C. and Mencar, C. (2015). *Interpretability of Fuzzy Systems: Current Research Trends and Prospects*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 219–237.
URL: *http://dx.doi.org/10.1007/978-3-662-43505-2_14*

Alvari, H. (2017). Twitter hashtag recommendation using matrix factorization, *CoRR* **abs/1705.10453**.
URL: *http://arxiv.org/abs/1705.10453*

Arifin, A. Z., Sari, Y. A., Ratnasari, E. K. and Mutrofinn, S. (2014). Emotion detection of tweets in indonesian language using non-negative matrix factorization, *International Journal of Intelligent Systems and Applications* **6 (9)**: 8.

Atsuho, N. (2017). *The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation*, Springer International Publishing, pp. 161–173.

Belford, M., Namee, B. M. and Greene, D. (2016). Ensemble topic modeling via matrix factorization, *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science, AICS 2016, Dublin, Ireland, September 20-21, 2016.*, pp. 21–32.

Berry, M., Browne, M., Langville, A., Pauca, P. and Plemmons, R. (2007). Algorithms and applications for approximate nonnegative matrix factorization, *Computational Statistics and Data Analysis* **52**(1): 155–173.

Berthold, M. and Hand, D. J. (eds) (1999). *Intelligent Data Analysis: An Introduction*, 1st edn, Springer-Verlag New York, Inc.

Berthold, M. R., Borgelt, C., Höppner, F. and Klawonn, F. (2010). *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, 1st edn, Springer Publishing Company, Incorporated.

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*, 1st edn, O'Reilly Media, Inc.

Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization, *Pattern Recognition* **41**: 1350–1362.

Casalino, G., Castiello, C., Buono, N., Esposito, F. and Mencar, C. (2017). Q-matrix extraction from real response data using nonnegative matrix factorizations, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10404**: 203–216.

Casalino, G., Castiello, C., Del Buono, N. and Mencar, C. (2017). Intelligent twitter data analysis based on nonnegative matrix factorizations, *in* G. O. et al. (ed.), *Computational Science and Its Applications ICCSA 2017*, Vol. 10404 of *Lecture Notes in Computer Science*, Springer.

Casalino, G., Del Buono, N. and Mencar, C. (2011). Subtractive initialization of nonnegative matrix factorizations for document clustering, *in* A. Fanelli, W. Pedrycz and A. Petrosino (eds), *Fuzzy Logic and Applications*, Vol. 6857 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 188–195.

Casalino, G., Del Buono, N. and Mencar, C. (2014a). Part-based data analysis with masked non-negative matrix factorization, *in* B. Murgante, S. Misra, A. M. A. C. Rocha, C. M. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan and O. Gervasi (eds), *Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Guimarães, Portugal, June 30 - July 3, 2014, Proceedings, Part VI*, Vol. 8584 of *Lecture Notes in Computer Science*, Springer, pp. 440–454.

Casalino, G., Del Buono, N. and Mencar, C. (2014b). Subtractive clustering for seeding non-negative matrix factorizations, *Information Sciences* **257**(0): 369 – 387.

Casalino, G., Del Buono, N. and Mencar, C. (2016). *Nonnegative Matrix Factorizations for Intelligent Data Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 49–74.
URL: *https://doi.org/10.1007/978-3-662-48331-2_2*

Casalino, G. and Gillis, N. (2017). Sequential dimensionality reduction for extracting localized features, *Pattern Recognition* **63**: 15 – 29.
URL: *http://www.sciencedirect.com/science/article/pii/S0031320316302667*

Chen, Y., Wang, L. and Dong, M. (2010). Non-negative matrix factorization for semisupervised heterogeneous data coclustering, *IEEE Transaction on knowledge and data engineering* **22**(10): 1459–1474.

Cichocki, A., Zdunek, R., Phan, A. H. and Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley.

D'Andrea, E., Ducange, P., Lazzerini, B. and Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis, *IEEE Transactions on Intelligent Transportation Systems* **16**(4): 2269–2283.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990). Indexing by latent semantic analysis, *JASIS* **41**: 391–407.

Del Buono, N., Esposito, F., Fumarola, F., Boccarelli, A. and Coluccia, M. (2016). *Breast Cancer's Microarray Data: Pattern Discovery Using Non-negative Matrix Factorizations*, Springer International Publishing, Cham, pp. 281–292.
URL: *http://dx.doi.org/10.1007/978-3-319-51469-7_24*

Ding, C., He, X. and Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and k-means - spectral clustering, *Proceedings of the SIAM Data Mining Conference*, SIAM, pp. 606–610.

Ducange, P., Mannar, G., Marcelloni, F., Pecori, R. and Vecchio, M. (2017). A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers, *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6.

Duong-Trung, N., Schilling, N. and Schmidt-Thieme, L. (2017). Finding hierarchy of topics from twitter data, *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings, Rostock, Germany, September 11-13, 2017.*, p. 39.

Gillis, N. (2012). Sparse and unique nonnegative matrix factorization through data preprocessing, *Journal of Machine Learning Research* **13**: 3349–3386.

Gillis, N. (2014). The why and how of nonnegative matrix factorization, *in* M. S. J.A.K. Suykens and A. Argyriou (eds), *Regularization, Optimization, Kernels, and Support Vector Machines*, Machine Learning and Pattern Recognition Series, Chapman and Hall/CRC.

Godfrey, D., Johns, C., Sadek, C., Meyer, C. and Race, S. (2014). A case study in text mining: Interpreting twitter data from world cup tweets.
URL: *https://arxiv.org/pdf/1408.5427.pdfl*

Guo, J., Zhang, P., Tan, J. and Guo, L. (2012). Mining hot topics from twitter streams, *Procedia Computer Science* **9**(Supplement C): 2008 – 2011. Proceedings of the International Conference on Computational Science, ICCS 2012.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050912003456*

Gupta, A., Joshi, A. and Kumaraguru, P. (2012). Identifying and characterizing user communities on twitter during crisis events, *Proceedings of the 2012 Workshop on Data-driven User Behavioral Modelling and Mining from Social Media*, DUBMMSM '12, ACM, New York, NY, USA, pp. 23–26.
URL: *http://doi.acm.org/10.1145/2390131.2390142*

Holmes, J. H. and Peek, N. (2007). Intelligent data analysis in biomedicine., *Journal of Biomedical Informatics* **40**(6): 605–608.

Ibrahim, R., Elbagoury, A., Kamel, M. S. and Karray, F. (2017). Tools and approaches for topic detection from twitter streams: survey, *Knowledge and Information Systems* .
URL: *https://doi.org/10.1007/s10115-017-1081-x*

Iskandar, A. A. (2017). Topic extraction method using red-nmf algorithm for detecting outbreak of some disease on twitter, *AIP Conference Proceedings* **1825**(1): 020010.

Jin, L., Chen, Y., Wang, T., Hui, P. and Vasilakos, A. (2013). Understanding user behavior in online social networks: a survey, *Communications Magazine, IEEE* **51**(9): 144–150.

Kim, H. and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics* **23**(12): 1495–1502.

Kim, Y.-H., Seo, S., Ha, Y.-H., Lim, S. and Yoon, Y. (2013). Two applications of clustering techniques to twitter: Community detection and issue extraction, *Discrete Dynamics in Nature and Society* **2013**: 8.

Klinczak, M. N. M. and Kaestner, C. A. A. (2015). A study on topics identification on twitter using clustering algorithms, *2015 Latin America Congress on Computational Intelligence (LA-CCI)*, pp. 1–6.

Klinczak, M. N. M. and Kaestner, C. A. A. (2016). Comparison of clustering algorithms for the identification of topics on twitter, *Latin American Journal of Computing* .

Kuang, D., Park, H. and Choo, J. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering.

Lai, E. L., Moyer, D., Yuan, B., Fox, E., Hunter, B., Bertozzi, A. L. and Brantingham, P. J. (2016). Topic time series analysis of microblogs, *IMA Journal of Applied Mathematics* **81**(3): 409–431.
URL: *http://dx.doi.org/10.1093/imamat/hxw025*

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization, *Nature* **401**(6755): 788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization, *in* T. K. Leen, T. G. Dietterich and V. Tresp (eds), *Advances in Neural Information Processing Systems 13*, MIT Press, pp. 556–562.

Li, C., Yang, Z. and Fan, K. (2015). BJUT at TREC 2015 microblog track: Real-time filtering using non-negative matrix factorization, *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015.*
URL: *http://trec.nist.gov/pubs/trec24/papers/BJUT-MB2.pdf*

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization, *Neural Comput.* **19**(10): 2756–2779.
URL: *http://dx.doi.org/10.1162/neco.2007.19.10.2756*

Liu, H. and Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC.

Mankad, S. and Michailidis, G. (2015). Analysis of multiview legislative networks with structured matrix factorization: Does twitter influence translate to the real world?, *Ann. Appl. Stat.* **9**(4): 1950–1972.
URL: *https://doi.org/10.1214/15-AOAS858*

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining, *in* N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta.

Panisson, A., Gauvin, L., Quaggiotto, M. and Cattuto, C. (2014). Mining concurrent topical activity in microblog streams, *Proceedings of the the 4th Workshop on Making Sense of Microposts co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 7th, 2014.*, pp. 3–10.

Pascual-Montano, A., Carazo, J. M., Kochi, K., Lehmann, D. and Pascual-Marqui, R. D. (2006). Nonsmooth nonnegative matrix factorization (nsnmf), *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(3): 403–415.

Pei, Y., Chakraborty, N. and Sycara, K. (2015). Nonnegative matrix tri-factorization with graph regularization for community detection in social networks, *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, AAAI Press, pp. 2083–2089.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20**(Supplement C): 53 – 65.
URL: *http://www.sciencedirect.com/science/article/pii/0377042787901257*

Saha, A. and Sindhwani, V. (2012). Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization, *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pp. 693–702.
URL: *http://doi.acm.org/10.1145/2124295.2124376*

Saito, S., Hirata, Y., Sasahara, K. and Suzuki, H. (2015). Tracking time evolution of collective attention clusters in twitter: Time evolving nonnegative matrix factorisation, *PLOS ONE* **10**(9): 1–17.
URL: *https://doi.org/10.1371/journal.pone.0139085*

Salton, G., Wong, A. and Yang, C. S. (1975). A vector space model for automatic indexing, *Commun. ACM* **18**(11): 613–620.
URL: *http://doi.acm.org/10.1145/361219.361220*

Sauwen, N., Acou, M., Bharath, H. N., Sima, D. M., Veraart, J., Maes, F., Himmelreich, U., Achten, E. and Van Huffel, S. (2017). The successive projection algorithm as an initialization method for brain tumor segmentation using non-negative matrix factorization, *PLOS ONE* **12**(8): 1–17.
URL: *https://doi.org/10.1371/journal.pone.0180268*

Shahnaz, F., Berry, M. W., Pauca, V. P. and Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization, *Inf. Process. Manage.* **42**(2): 373–386.

Shamma, D. A., Kennedy, L. and Churchill, E. F. (2009). Tweet the debates: Understanding community annotation of uncollected sources, *Proceedings of the First SIGMM Workshop on Social Media*, WSM '09, ACM, New York, NY, USA, pp. 3–10.
URL: *http://doi.acm.org/10.1145/1631144.1631148*

Shin, D. S., Choi, M., Choi, J., Langevin, S., Bethune, C., Horne, P., Kronenfeld, N., Kannan, R., Drake, B., Park, H. and Choo, J. (2017). Stexnmf: Spatio-temporally exclusive topic discovery for anomalous event detection, *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 435–444.

Sitorus, A. P., Murfi, H., Nurrohmah, S. and Akbar, A. (2017). Sensing trending topics in twitter for greater jakarta area, *International Journal of Electrical and Computer Engineering (IJECE)* **7**(1): 330–336.

Suh, S., Choo, J., Lee, J. and Reddy, C. K. (2016). L-ensnmf: Boosted local topic discovery via ensemble of nonnegative matrix factorization, *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, pp. 479–488.
URL: *https://doi.org/10.1109/ICDM.2016.0059*

Suh, S., Choo, J., Lee, J. and Reddy, C. K. (2017). Local topic discovery via boosted ensemble of nonnegative matrix factorization, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 4944–4948.
URL: *https://doi.org/10.24963/ijcai.2017/699*

Suri, P. and Roy, N. R. (2017). Comparison between lda nmf for event-detection from large text stream data, *2017 3rd International Conference on Computational Intelligence Communication Technology (CICT)*, pp. 1–5.

Wakamiya, S., Lee, R., Kawai, Y. and Sumiya, K. (2015). Twitter-based urban area characterization by non-negative matrix factorization, *Proceedings of the 2015 International Conference on Big Data Applications and Services*, BigDAS'15, ACM, New York, NY, USA, pp. 128–135.
URL: *http://doi.acm.org/10.1145/2837060.2837079*

Wong, F. M. F., Tan, C. W., Sen, S. and Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters, *IEEE Transactions on Knowledge and Data Engineering* **28**(8): 2158–2172.

Xu, W., Liu, X. and Gong, Y. (2003). Document clustering based on non-negative matrix factorization, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, ACM, New York, NY, USA, pp. 267–273.

Yan, X., Guo, J., Liu, S., Cheng, X. and Wang, Y. (n.d.). *Learning Topics in Short Texts by Non-negative Matrix Factorization on Term Correlation Matrix*, pp. 749–757.
URL: *http://epubs.siam.org/doi/abs/10.1137/1.9781611972832.83*