Article

Knowledge Discovery From Large Amounts Of Social Media Data

Loris Belcastro [1], Riccardo Cantini [1] and Fabrizio Marozzo [1,*]

[1] DIMES, University of Calabria, Rende, Italy
* Correspondence: fmarozzo@dimes.unical.it

Abstract: In recent years, social media analysis is arousing great interest in various scientific fields, such as sociology, political science, linguistics, and computer science. Large amounts of data gathered from social media are widely analyzed for extracting useful information concerning people's behaviors and interactions. In particular, they can be exploited to analyze the collective sentiment of people, understand the behavior of user groups during global events, monitor public opinion close to important events, identify the main topics in a public discussion, or detect the most frequent routes followed by social media users. As an example of the countless works in the state-of-the-art on social media analysis, this paper presents three significant applications in the field of opinion and pattern mining from social media data: *i*) an automatic application for discovering user mobility patterns, *ii*) a novel application for estimating the political polarization of public opinion, and *iii*) an application for discovering interesting social media discussion topics through a hashtag recommendation system. Such applications clearly highlight the abundance and wealth of useful information in many application contexts of human life that can be extracted from social media posts.

Keywords: Big Data; Social media analysis; Big Data analysis; Social media applications; Knowledge Discovery

Citation: Belcastro, L.; Cantini, R.; Marozzo, F. Knowledge Discovery From Large Amounts Of Social Media Data. *Appl. Sci.* **2022**, *1*, 0. https://doi.org/

1. Introduction

Massive amounts of digital data, generated every day on social media platforms, can be used to extract useful information about human interests, opinions and dynamics. [1]. The analysis of Social Big Data [2] belongs to a research area that deals precisely with studying the activities, interests, behaviors and opinions of users by analysing the posts published on social media platforms. A large community of researchers is focused on developing applications for analyzing Social Big Data, usually relying on advanced and scalable algorithms for obtaining accurate results in a reasonable time [3]. Novel machine learning applications are thus defined for extracting useful knowledge in different application fields, including trend discovery, social media analytics, pattern mining, sentiment analysis, and opinion mining. Different surveys have been proposed in the literature [4–6]for trying to summarize and describe the wide range of research papers published in recent years in this area.

This paper presents some of the most recent activities we have carried out in the field of Social Big Data analysis. In particular, it discusses how these research activities can be exploited and, if necessary, combined for the analysis of large amounts of social media data, aimed at extracting different kind of knowledge from three different perspectives: *i*) from the posts published by tourists who visit a city, we can discover the main tourist attractions and also the mobility patterns (i.e., trajectories) across them [7]; *ii*) from public discussions on social media close to important electoral events, it is possible to discover the political orientation of citizens and thus estimate the outcome of a political event [8]; *iii*) from hashtags used by social media users, we discover the main topics underlying social media conversation and how users refer to them in publishing online content [9].

Concerning the *trajectory analysis*, we introduce AUDESOME, an algorithm for detecting user mobility patterns from content posted on social media. Starting from a large set of geotagged posts (e.g., Flickr posts or tweets), it performs a series of operations for detecting the keywords identifying the Places-of-Interest (PoIs) in a given area, the Region-of-Interest (RoI) associated to each PoI, and frequent mobility patterns in user movements across RoIs. Experimental results show that our technique is able to correctly extract the most frequent trajectories and mobility patterns of user groups compared to the techniques present in the state of the art.

About *opinion mining*, we present IOM-NN, a data-driven technique that exploits feed-forward neural networks for estimating the political polarization of social media users during elections. The technique repeatedly creates new categorization rules using a limited number of classification rules, which are produced from an initial set of hashtags that are notoriously in favor of specific political parties or factions. In particular, a classification rule uses the words/hashtags in a post to evaluate if it is in favor of a faction or not. Then, the classified posts of each user are used to determine his/her political alignment and, consequently, from the analysis of a large number of users, to estimate the outcome of a political event. Such a technique was tested on a real case study, achieving results that are more accurate than opinion polls and other approaches proposed in the literature.

Lastly, about *topic discovery*, we present HASHET, a model for recommending a set of hashtags that are suitable for a given post. In particular, we discuss how the recommendation abilities of our model can be leveraged for linking the content published by users to the main discussion topics underlying social media conversation. HASHET uses two latent spaces, independent from each other, to embed the textual content of a post and the hashtags it contains. The first latent space is built through a pre-trained BERT (*Bidirectional Encoder Representations form Transformers*) language representation model, which leverages a self-attention mechanism for detecting semantic and syntactic features of texts. The second space, based on a CBOW (*Continuous Bag of Words*) trained model, is used to extract the contextual relationships between words and hashtags. After having identified in the embedding space of hashtags a clustering structure based on topics, HASHET can be exploited to identify the main topics of discussion as well as the topic to which a given tweet belongs. Several experiments proved that HASHET is the best overall model in identifying the main topics of discussion, overcoming the other state-of-the-art approaches.

All the applications discussed in this paper have been defined and executed in parallel on a Cloud platform, by exploiting ParSoDA [10], a library that enables developers to create Cloud-based parallel applications for analyzing large volumes of social media data. ParSoDA is composed of different packages, which include several functions that are commonly used to process and analyze social media data, so as to discover different types of information (e.g., user opinion, topic trends, user mobility patterns). In addition, the library provides a set of interfaces and abstract classes to be implemented/extended for creating new functions. ParSoDA is based on two of the most popular parallel processing frameworks for Big Data (Apache Hadoop and Apache Spark), which are fundamental to ensure scalability as the amount of data to be processed increases.

This work tries to summarize in a single paper, following a uniform descriptive scheme, three examples of social media applications designed for analyzing data in three different contexts. The aim is to highlight how, starting from the digital contents that people share on social media (e.g., posts, videos or photos), extremely valuable information for many disciplinary fields can be obtained. Starting from a common step-by-step scheme, the presented applications are implemented using a single development framework, also reporting the main results achieved. In this way, the Reader can find, in a single place, three examples of trending applications in the area of social Big Data analysis, useful to give an idea of the vastness of applications that can be implemented in this area.

The paper is organized as follows. Section 2 discusses related work. Section 3 introduces the metadata model proposed for integrating data gathered from different social media. Section 4 introduces the ParSoDA library. Section 5 describes the three applications of trajectory mining, opinion mining, and topic discovery. Finally, Section 6 concludes the paper.

2. Background

Over the past few years, several programmers and researchers have been working on developing new tools, algorithms and programming models to extract relevant information from Big Data. Generally, the volume of data to be processed exceeds the computing capabilities of traditional IT systems, so clusters and Clouds are exploited to effectively run parallel and distributed applications, capable of ensuring scalability and acceptable execution times [3].

Even novel data mining techniques are created for extracting useful knowledge in different application fields, including trend discovery, trajectory mining, predictive data analytics, sentiment and opinion mining. Several surveys on this argument have been proposed in literature [4–6] trying to summarize the innumerable research works that have been published in the last years on this argument. In the following, some examples of recent social media analysis applications are presented.

Social media analysis frequently aims at understanding human dynamics and behaviors, such as understanding the most followed touristic routes and the period of year when touristic attractions are visited [11][12], detecting the crowded areas of a city where transport facilities need to be improved [13], finding the most suitable areas for opening new businesses [14], analyzing the purchasing behavior of users [15], uncovering the behaviors of fans following important sporting events [16].

Many research projects focus not only on the data analysis process, but also on other data processing tasks (e.g., data cleaning, preparation, and transformation), which are fundamentals while developing a data analysis application. These efforts, in particular, aim at assisting researchers and data analysts in implementing all of the phases that compose data analysis applications without having to start from scratch.

SOCLE [17] is a data preparation framework specifically designed for social media applications, which provides a large set of operators for data pruning, data enrichment and data normalization. Cuesta et al. [18] presented a MapReduce framework that provides developers with easy-to-use modules for data collection, data storage and data analytics (e.g., sentiment analysis and reporting). Still in the context of Twitter data, Zhou et al. [19] proposed an unsupervised framework able to discover events from large volumes of tweets through a pipeline process consisting of filtering, extraction and categorization steps. You et al. [13] presented a Cloud-based framework for developing social media analysis applications to support mobility in smart cities. It manages data collection from social media platforms APIs (e.g., Flickr, Foursquare, Twitter) and from other web sources (e.g., websites, blogs). SODATO (SOcial Data Analytics Tool) [20] is a web-based tool for programming data analytics on social media data, which includes some predefined analysis methods, such as sentiment analysis, text analysis, content performance analysis, influencer analytics, and so on.

Taking into account the advances produced by research in this area, this paper presents the definition of three significant applications of Social Big Data analysis in the fields of trajectory mining, sentiment analysis, and topic detection. Such applications analyze large amounts of data and usually require long computation times. Consequently, to get results in a reasonable time, we used ParSoDA for enabling these applications to run on Cloud. The ParSoDA's runtime has been created specifically for dealing with large amounts of data. As a result, it is built on the MapReduce architecture and can run in parallel on distributed computing systems like HPC and Clouds.

3. Metadata model for social media data

One of the main problems to solve when working with social media data is to adopt a standardized data model for representing and integrating data coming from different sources. To this end, Belcastro et al. [10] proposed a unified metadata model for representing different data extracted from social media platforms. According to this model, each social media item (e.g., post, photo, or video) is described by a JSON document [21] organized into two sections. The first section, called *basic*, includes the main descriptive fields available in all major social media platforms (source, item ID, date and time, location coordinates, user information). The second section, called *extra*, contains specific fields that depend on each source. For example, Listing 1 and 2 show the metadata related to a tweet and a photo posted on Flickr. The basic section is the same for the two documents, instead the extra section contains specific information for tweets (e.g., retweet and retweet count) and for Flickr photos (e.g., a list of tags and the photo quality).

```
{
 "BASIC":{
  "SOURCE":"Twitter", "ID":"1234567890123",
  "DATETIME":"2021-02-20T22:19:35.021",
  "LOCATION":{"LNG":-0.1259,"LAT":51.5623},
  "USER":{ "USERID":"0123456789", "USERNAME":"mrpotato"}},
 "EXTRA":{
  "inReplyToScreenName":"djdna", "inReplyToUserId":987654321,
  "inReplyToStatusId":9565757346292993,
  "text":"@djdna perfect sound!",
  "hashtags":[ "#festival", "#music"], "retweets":10, "isRetweet":true}
}
```

Listing 1: Metadata of a tweet.

```
{
 "BASIC":{
  "SOURCE":"Flickr", "ID":"43146791176",
  "DATETIME":"2020-11-09T15:21:12.000",
  "LOCATION":{"LNG":12.567772,"LAT":41.89256},
  "USER":{ "USERID":"987654321@N01", "USERNAME":"samjack"}},
 "EXTRA":{
  "title":"The Colosseum is amazing",
  "description":"The Colosseum in Rome is really amazing"
  "tags":[{"count":1,"value":"trip"},{"count":2,"value":"rome"}],
  "dateTaken":"Oct 10, 2020 15:21:12 AM",
  "accuracy": 15}
}
```

Listing 2: Metadata of a Flickr photo.

4. Scalable application using ParSoDA

ParSoDA (Parallel Social Data Analytics) [10] is a library for processing and analyzing in parallel large amounts of data collected from social media platforms. The main goal of ParSoDA is to allow programmers to easily extract knowledge from social media data by hiding the difficulty of defining a parallel/distributed application made up of many steps. In fact, programmers have a number of hurdles when designing and implementing these applications, including parallelizing complex algorithms, lowering communication costs, and optimizing memory utilization. For this purpose, the library allows to define a data analysis application starting from a general structure consisting of seven steps:

1. *Data acquisition* for collecting social media items and storing them in a persistent repository (e.g., HDFS [22]).
2. *Data filtering* for filtering social media items according to a set of functions.
3. *Data mapping* for transforming the information contained in each social media item through some functions.

4. *Data partitioning* for partitioning items into shards by a primary key and then sorting them by a secondary key.
5. *Data reduction* for aggregating items contained in a shard according to a function.
6. *Data analysis* for analyzing data using a data analysis function in order to extract the knowledge of interest.
7. *Data visualization* for visualizing data analysis results in a suitable visual format.

ParSoDA provides a predefined set of functions for each step. For example, ParSoDA includes functions for crawling data from Twitter and Flickr, for filtering posts based on location or time of publication, for transforming data from one format to another, functions for classifying and clustering data and so on. However, users can extend this set of functions with their own.

Listing 3 shows the code of the application for extracting the user polarization between rival political factions. First, at line 3, a *SocialDataApp* is instantiated. Then the output path, the file system, and a file containing the keywords used to support the two factions are defined (*lines 4-7*). The rest of the code declares the functions to be used for data filtering, data mapping, data sorting, data reduction and data analysis.

```
1 public class UserPolarizationMain {
2     public static void main(String[] args) {
3         SocialDataApp app = new SocialDataApp("2 Faction User Polarization");
4         app.setOutputBasePath("outputApp");
5         app.setLocatFileSystem();
6         String[] cFiles = {"resources/twoFactionKeywords.json"};
7         app.setDistributedCacheFiles(cFiles);
8         Class[] cFunctions = {FileReaderCrawler.class};
9         String[] cParams = {"-i resources/tweetsFinal.json"};
10        app.setCrawlers(cFunctions, cParams);
11        Class[] mFunctions = {ClassifyTwoFactionsEvent.class};
12        String[] mParams = {"-f twoFactionKeywords.json"};
13        app.setMapFunctions(mFunctions, mParams);
14        String groupKey = "userId";
15        String sortKey = "DATETIME";
16        app.setPartitioningKeys(groupKey, sortKey);
17        Class rFunction = ReduceByTwoFactionsPolarization.class;
18        String rParams = "-t 5";
19        app.setReduceFunction(rFunction, rParams);
20        Class aFunction = TwoFactionsPolarization.class;
21        String aParams = null;
22        app.setAnalysisFunction(aFunction, aParams);
23        app.execute();
24    }
25 }
```

Listing 3: A user polarization application written using ParSoDA.

Using ParSoDA reduces the number of lines of code required to develop complex data analysis applications, as proved in [23]. In particular, ParSoDA allows programmers to save hundreds of lines of code in the main (as they only need to configure the functions to be used at each step) and in the other steps, where built-in functionalities are used and where the programmer needs only to define the function logic. Starting from the application main, ParSoDA creates and executes a chain of tasks for the different steps. Without ParSoDA, programmers must manually control the execution and parallelization of each step.

5. Social Big Data analysis applications

In this section, we show three examples of social media applications designed for analyzing big amounts of data gathered from social media. In particular, we focused on the extraction of knowledge from three different viewpoints. The first application deals with *frequent trajectory mining from social media data*, for extracting the most frequent user trajectories and mobility patterns through a set of Points-of-Interest located in a geographical area (e.g., a city). The second application deals with *opinion mining from social media data*, for discovering the polarization of social media users during a political event (e.g., election, referendum), usually characterized by the competition of two or more political factions. The third application is on *topic discovery from social media*, for the suggestion of a proper set of hashtags for a given post that leverages the state-of-the-art techniques for natural language processing.

For each example, we describe the steps of the analysis by using an uniform descriptive schema and providing technical details with operating information, and the case study addressed with the results achieved and comparison with the main related work present in the state of the art.

5.1. Frequent trajectory mining from social media data

AUDESOME [7] is a method for extracting the most frequent user trajectories and mobility patterns through a set of Points-of-Interest (PoIs) located in a geographical area (e.g., a city). Generally, PoIs refer to tourist attractions, such as monuments, squares or bridges, or to business places, such as airports, shopping malls or train stations. A *trajectory* is a sequence of locations visited by a user. For analyzing users' behavior, it is useful to understand whether a user visited or not a PoI. Since information on a PoI is generally limited to an address or to GPS coordinates, it is hard to match trajectories with PoIs. For this reason, it is useful to define the so-called *Regions-of-Interest* (*RoIs*) that represent the boundaries of the PoIs' area [24].

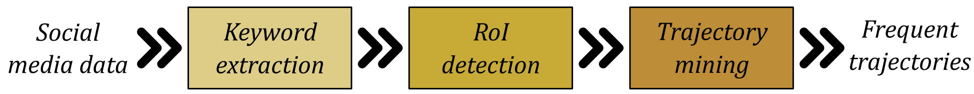

Figure 1. Execution flow of Frequent Trajectory Mining using AUDESOME.

In order to obtain the most frequent trajectories from a set of geotagged items from social media platforms, the following steps are performed (see Figure 1):

- *Keyword extraction* for discovering the main keywords that are used by social media users to identify the PoIs that are located in the area. Such keywords are used to group social media posts according to the place they refer to. The intermediate output of this step is the sets of keywords identifying the PoIs.
- *RoI detection* for extracting the Regions-of-Interest (RoIs) from social media posts that have been grouped by keywords. Specifically, the geotagged social media posts referring to a PoI are transformed into a series of geographical points and clustered to define RoIs. The intermediate output of this step is the RoIs calculated for the different PoIs, which can also be easily displayed on a map.
- *Trajectory mining* for finding the trajectories across RoIs for each user, and thus obtaining the mobility patterns (i.e., the most frequent trajectories).

5.1.1. Keyword extraction

The keyword extraction algorithm extracts the most relevant keywords used by social users to tag places-of-interest in a given area. The algorithm is composed of three steps:

1. *Keyword discovery*. The area of interest is divided into cells of equal size in order to assign the posts to each cell on the basis of geolocalization. Then, in each cell, the main keywords are found (sorted by frequency) by analyzing the description of the posts associated with a cell. The noisy keywords are then removed in the next step.
2. *Keyword selection*. To distinguish between high and low frequency keywords, a method based on a discrete L-curve [25] is used. Finding the elbow point of this curve allows to distinguish between high and low frequency keywords. The algorithm takes into account both global high frequency keywords (i.e., calculated

over the whole area) and local high frequency keywords (i.e., calculated on each cell) for generating a list of the most representative keywords for each cell.

3. *Keyword grouping*. The most representative keywords are grouped by their textual similarity using the Levenshtein's metric [26]. The algorithm produces a number of sets containing similar keywords, where each set contains the keywords that identify a specific PoI. During the RoI detection phase, each set of keywords is used to find the associated RoI.

5.1.2. RoI detection

The RoI detection algorithm aims at defining the Regions-of-Interest by clustering the geotagged items assigned to the different PoIs. In fact, the geotagged items can be transformed into geographical points (i.e., pairs of ⟨latitude, longitude ⟩), which can be aggregated through clustering. Specifically, an adapted version of DBSCAN [27] is used for RoI mining with an automatic estimation of the parameters required by the algorithm.

The DBSCAN algorithm needs two key parameters: *eps*, the radius of a neighborhood with respect to some point; and $minPts$, the minimum number of points required to form a cluster. These two parameters can be calculated using the following procedure as defined in [27]:

1. Calculate the plot of the k-nearest-neighbor distances (k-dist), computed for each point, and sorted by descending order [28]. As suggested in [27], for bi-dimensional data, k can be set to 4.
2. Choose a *threshold point* on k-dist plot for separating noise points (i.e., all points with a higher k-dist value than threshold) from points that are assigned to some clusters (i.e., all points with a lower k-dist value than threshold). The threshold point is calculated by estimating the noise percentage in the data (*noisePerc*).
3. The k-dist value of the threshold point is used as *eps* value. Concerning $minPts$, it can be set as $k + 1$ [28].

Since DBSCAN calculates one or more clusters on a set of points, we select the points that belong to the largest cluster, and starting from them we return the convex polygon that encloses these points (i.e., a RoI).

5.1.3. Trajectory mining

At this step the input dataset is analyzed for finding behaviors and mobility patterns of users. In order to prepare the data for analysis, the user's trajectories are transformed from sequences of coordinates into sequences of RoIs. For this reason, the quality of the RoIs influences the quality of the trajectories obtained.

Sequential pattern analysis is used to discover the sequences of RoIs that occur most frequently in the data. In sequential analysis, the time dimension and chronological order in which the values appear in the data are crucial. The sequences obtained from the movements of the different users are analyzed together in order to discover the most frequent ones. To this end, support and confidence are calculated for each frequent trajectory found. By setting threshold values for these two factors, we are able to filter and discover the most frequent trajectories followed by users.

5.1.4. Use cases and results

We experimentally evaluated the performance and scalability of AUDESOME using more than 3 million of geotagged items, published in Flickr from January 2006 to May 2020 in the cities of Rome and Paris. Specifically, we evaluated the accuracy in extraction of user trajectories of AUDESOME with respect to four existing techniques (see Table 1): *DBSCAN* [29], *DSets-DBSCAN* [30], Slope [31], and *G-RoI* [24]. The experiments demonstrate that AUDESOME achieves better results than existing techniques. AUDESOME outperformed the other techniques by reaching a mean F1 score of 0.85 in Rome and 0.87 in Paris. For this step, our method turns out to be the most accurate one in finding trajectories, achieving an overall improvement of the F1 score up to 0.39 (i.e., in comparison to DSets-DBSCAN). Also from the point of view of scalability, increasing the number of cores dedicated to the execution of the application leads to a significant reduction in overall execution times, which demonstrates the scalability of our method [32].

Algorithm	Rome dataset (F1)	Paris dataset (F1)
DBSCAN [29]	0.82	0.82
DSets-DBSCAN [30]	0.46	0.69
Slope [31]	0.68	0.77
G-RoI [24]	0.83	0.84
AUDESOME [7]	0.85	0.87

Table 1: Results comparison for the topic discovery task.

5.2. Opinion Mining from social media data

IOM-NN [8] is a recent methodology aiming at uncovering the polarization of social media users during a political event (e.g., election, referendum), usually characterized by the competition of two or more political factions. It is based on an iterative and incremental procedure that exploits feed-forward neural networks, aimed at determining the political orientation of users by analyzing the political polarization of social media posts. IOM-NN is open source and an implementation is publicly available at: https://github.com/SCAlabUnical/IOM-NN. When using social media data for the estimation of the political polarization of public opinion, several issues several issues need to be addressed:

- *Language barrier*. Our technique is keyword-based, thus resulting language-independent.
- *Data imbalance*. A random sampling approach is used to balance the dataset across the different factions.
- *Data reliability*. We assessed the statistical significance of the collected data in order to understand whether geo-located users under analysis can actually be considered voters.
- *Misclassification*. Only rules that present a high likelihood in the association between a post and its assigned faction are used.

Figure 2. Execution flow of IOM-NN.

As shown in Figure 2, the proposed methodology is comprised of of three main steps:

1. *Keyword definition*: a set of keywords, related to the political event under analysis, is defined in order to gather data from the social media platforms (see Section 5.2.1).
2. *Classification of posts*: an iterative procedure based on feed-forward neural networks is leveraged for assigning the collected posts to a specific faction (see Section 5.2.2).
3. *Polarization of users*: starting from the classified posts, the political orientation of the users is calculated (see Section 5.2.3).

5.2.1. Collection of posts

During this phase a set P of social media posts are collected from different sources (e.g., Twitter, Flickr or other microblog platforms). Specifically, posts are searched and collected by using a set of keywords $\mathcal{K}$ that people commonly use to refer to a political event $\mathcal{E}$ on social media (es. faction-specific hashtags). Specifically, given a set of factions $\mathcal{F} = \{f_1, \ldots, f_n\}$, the methodology exploits two types of keywords in $\mathcal{K}$:

- $\mathcal{K}_{context}$, containing generic keywords or hashtags that can be associated to the political event $\mathcal{E}$, but that do not refer to any specific faction (e.g., *#vote*, *#election*);

- $\mathcal{K}_F = \mathcal{K}_{f_1}, \ldots \mathcal{K}_{f_n}$, which contains the keywords used for supporting each faction $f \in \mathcal{F}$ (e.g., *#votehillary*, *#maga*, *#imwithher*, *#votetrump*).

5.2.2. Classification of posts

During this step, social media posts are incrementally classified as in favor of a specific faction by leveraging an iterative process based on a feed-forward neural network, specifically a multilayer perceptron.

As a first step, IOM-NN builds a classification model M_0 based on a small set of manually-defined faction keywords ($K_{\mathcal{F}}$). Such a model is then used for classifying a part of the posts. Specifically, at this iteration, the algorithm classifies a post in favor of a faction if it contains only positive keywords related to that faction. This means that, at the end of the first iteration, just a small amount of posts are classified, since not all users use the positive keywords in K_F for expressing their support to a faction.

In the subsequent iterations, IOM-NN iteratively generates new classification rules aimed at classifying posts that are not yet assigned to any faction. These rules are extracted by a multilayered perceptron, which is specially trained to find out hidden relationships between hashtags used by social media users and their political alignment. The training phase exploits all the posts that have been classified at the previous iterations. Afterwards, new posts are classified if they can be assigned to a specific faction with a high probability (≥ 0.9, by default), computed by a softmax activation on the set of factions $\mathcal{F}$. The procedure iterates until a maximum number of iterations are made or convergence is reached, i.e. there are no more posts to be classified or the percentage of classified posts in the current iteration does not exceed a predetermined threshold.

5.2.3. Polarization of users

In this final phase, IOM-NN exploits the posts classified in the previous step in order to estimate the political orientation of social media users who wrote those posts. Then, starting from this estimate, the outcome of the political event $\mathcal{E}$ is predicted.

As a first step, the classified posts are grouped by user to get the list of classified posts for each user. Starting from this, for all users the algorithm computes the number of posts published in favor of each faction $f \in \mathcal{F}$. Subsequently, users are filtered based on how active they are on the social platform and how significant their political alignment is with the assigned faction. Specifically, a user is considered only if he/she fulfills the following criteria:

- He/she posted at least a minimum number of posts which show a political alignment.
- There exists a faction for which he/she has published more than 2/3 of his/her posts.

Afterwards, the algorithm determines a faction score for each user, defined as the percentage of posts he/she wrote in favor of his/her preferred faction. At the end, all faction scores are combined and normalized to obtain the overall polarization percentages for each faction $f \in \mathcal{F}$, which represents the prediction of the outcome of the event $\mathcal{E}$.

5.2.4. Use cases and results

The effectiveness of the proposed methodology has been assessed on a real-world case study, aimed at analyzing the polarization of a large number of Twitter users during the 2016 US presidential election. In particular, our analysis focused on data collected for ten US Swing States: *Colorado*, *Florida*, *Iowa*, *Michigan*, *Ohio*, *New Hampshire*, *North Carolina*, *Pennsylvania*, *Virginia*, and *Wisconsin*. The reason behind the choice of these states is linked to their greater political uncertainty, which implies a marked strategic importance, as their votes are more likely to be the deciding factor in a presidential election. We leveraged the Search Twitter API for the extraction of tweets published in a given area or place through geo-referencing, collecting about 820 thousands of tweets posted by 140 thousands users.

State	Real	Polls	IOM-NN
Colorado	**Clinton**	**Clinton**	**Clinton**
Florida	**Trump**	**Trump**	Clinton
Iowa	**Trump**	**Trump**	**Trump**
Michigan	**Trump**	Clinton	**Trump**
New Hampshire	**Clinton**	**Clinton**	**Clinton**
North Carolina	**Trump**	Tie	**Trump**
Ohio	**Trump**	**Trump**	**Trump**
Pennsylvania	**Trump**	Clinton	Clinton
Virginia	**Clinton**	**Clinton**	**Clinton**
Wisconsin	**Trump**	Clinton	**Trump**
Tweets	-	-	818,403
Users	-	$\approx 10,000$	141,959
Correctly classified	-	6/10	8/10

Table 2: Results comparison in terms of winning candidate. The candidate is written in bold when it is correctly identified.

Table 2 provides a comparison between the results obtained by IOM-NN, the average opinion polls collected before voting and the real results of the political event. The achieved results show the greater accuracy of IOM-NN, which correctly identified the winning candidate in 8 out of 10 cases It is also worth to note that IOM-NN allows the analysis of a much larger number of users at a lower cost, with respect to opinion polls, by exploiting the information-rich contents published by social media users.

5.3. Topic discovery from social media

HASHET [9] is a hashtag recommendation model designed for the suggestion of a proper set of hashtags for a given post, which leverages the state-of-the-art techniques for natural language processing, such as self-attention mechanism in transformer encoders and transfer learning from pre-trained language representation models.

In microblogging platforms, a hashtag is a generic string of characters and numbers that starts with the # symbol. It is generally used to label posts, linking them to trending topics, thus facilitating research and creating communities of like-minded users. However, due to the absence of constraints in choosing hashtags, it is often hard for users to select the appropriate ones, which leads to a huge number of posts characterized by the absence of a representative hashtag. This phenomenon can affect the quality of the results achieved by hashtag-based techniques, such as IOM-NN, and can be mitigated by the use of appropriate recommendation models. In addition, such models can be used to link posts to discussion topics, by identifying a topic-based clustering structure in which semantically-related hashtags are grouped together.

The HASHET models relies on the mapping between two independent semantic spaces, obtained through the embedding of sentences and words/hashtags. Thanks to this mapping, the model can learn the hidden relationships that link a given post to the latent representation of the hashtags it contains. Differently from the other state-of-the-art models based on deep learning architectures, HASHET do not relies on a softmax-loss setting, but exploits a novel concept of locality in the latent space of hashtags. By following this approach the model becomes fully aware of both the semantic relationships among hashtags and the underlying topic-based clustering structure, which leads to a significant improvement in hashtag prediction compared to other techniques. In particular, the recommendation is performed by identifying the projection of the latent vector, associated to the input post, into the embedding space of hashtags. Afterwards, a set of hashtags to be recommended is found in this space using k-nearest neighbor search and semantic expansion.

Figure 3 depicts the main steps of HASHET, a description of which is provided in the following.

Figure 3. Execution flow of HASHET.

5.3.1. Semantic mapping

HASHET relies on the projection (i.e. semantic mapping) of embedded representation of posts into the hashtag embedding space obtained by training a CBOW Word2Vec model [33]. Specifically, the embedding of a post is computed by exploiting the pre-trained BERT encoder, taking the hidden representation of the *CLS* token as the sentence embedding. The semantic mapping is performed by training a multi-layer perceptron to minimize the cosine distance between two latent vectors: $i)$ the projection of the embedded representation of the post into the hashtags latent space, and $ii)$ the embedded representation of its hashtags. In particular, the training process is performed as follows:

1. The BERT encoder is used for computing the embedded representation of the training posts, which are translated by the multi-layer perceptron, trained from scratch with a cosine distance loss until convergence is reached.
2. The entire semantic mapping model, comprising the BERT encoder and the multi-layer perceptron, is fully fine-tuned in an end-to-end fashion, in order to adapt the pre-trained feature of BERT to this particular downstream task. For this reason, in this training step, a small learning rate is used, in order to prevent pre-trained features from being distorted by large weight updates.

5.3.2. Hashtags recommendation

This phase concerns the use of the HASHET model for recommending hashtags to social media users. Specifically, given an input post, the corresponding embedding vector is computed, which is then projected into the hashtag latent space. Given this projection, named *target vector*, its k nearest hashtags are found and ordered by cosine similarity. As a last step, a semantic expansion process is performed, aimed at maximizing the hit rate of the recommendation system. In particular, given an expansion factor n, it includes in the output set the top-n semantically similar hashtags to those computed by the nearest neighbor search. As a result, the output set will consist of $k + n$ hashtags which are representative of both the semantic content of the input post and the spatial relationships of the hashtag embedding space.

5.3.3. Topic discovery

At this step, HASHET is exploited to identify the main topic of discussion of a given post. Firstly, we visualized the main topics present in the hashtag embedding space, identified by the different clusters formed by the spatial co-location of semantically similar hashtags. In order to achieve an effective visualization, we reduced the dimensionality of the latent vectors, projecting them into a 2D space using *Principal Component Analysis* and *t-distributed Stochastic Neighbor Embedding* techniques. Afterwards, we used *OPTICS*, a density-based clustering algorithm, identifying a partitioning of hashtags in a set of clusters $\mathcal{C} = \{c_1, \ldots, c_m\}$, each related to a different topic of discussion. The main reason behind the choice of this clustering algorithm is linked to the density-based approach it leverages, which leads to the identification of clusters of arbitrary shape. Moreover, through the *cut-clustering* mechanism, OPTICS allows the identification of clustering structures at different levels of density (i.e., detail), which is particularly useful for dealing with the presence of micro-topics in social media conversation.

Starting from the HASHET model and the set of clusters $\mathcal{C}$ the discussion topic for a given post p is determined as follows. Firstly the recommendation model is exploited in order to get a set of k hashtags suitable for that post. Afterwards, p can be classified in three ways:

- *Assignable*, if recommended hashtags belong to one cluster only.

- *Ambiguous*, if recommended hashtags belong to two or more clusters.
- *Neutral*, if recommended hashtags do not belong to any cluster.

Then, if p was labeled as *assignable*, it is assigned to the topic related to its corresponding cluster. Otherwise, if p is labeled as *neutral*, semantic expansion is iteratively used to recommend n additional hashtags, until a maximum expansion factor n is reached. Finally, *ambiguous* are not assigned to any specific topic, as we only focus on single-topic posts.

5.3.4. Use cases and results

In this section we show the results achieved by exploiting the recommendation abilities of HASHET for identifying the discussion topic of tweets related to COVID-19 pandemic. As a first step, as explained in section 5.3.3, we extracted a set of discussion topics from the input dataset, which are representative of the online discussion about the health emergency. In particular, we identified discussion groups about anti-vaccine protests in the USA (*#protests*, *#losangeles*), pro-vaccination campaigns (*#covidvaccine*,*#healthcare*), smartworking (*#workfromhome*, *#remotejobs*), and COVID-19 prevention rules (*#wearamask*, *#washyourhands*).

Afterwards, we maintained only single-topic tweets, i.e. those having hashtags belonging to exactly one cluster. Then we masked the real hashtags and leveraged the HASHET model in order to identify the topic for each tweet as explained in Section 5.3.3. Finally, discovered topics were compared to the real ones. Specifically, a topic assignment is correct if the recommended hashtags and the real ones belong to the same cluster, i.e., they are related to the same topic; otherwise the assignment may be neutral, ambiguous, or incorrect. This last case occurs when a unique but wrong cluster label can be determined from the recommended hashtags, i.e., the tweet can be assigned to a topic other than the one it actually belongs to. The results achieved by HASHET was compared with the main techniques present in the literature, which include unsupervised techniques and neural models based on different implementations of the attention mechanism.

Model	Correct	Incorrect	Neutral	Ambiguous
HF-IHU [34]	0.28	0.35	0.35	0.02
DBSCAN [35]	0.48	0.32	0.04	0.16
LDA-GIBBS [36]	0.51	0.22	0.05	0.22
TCAN [37]	0.66	0.08	0.11	0.15
BERT [38]	0.68	0.08	0.09	0.15
HASHET	**0.77**	**0.07**	**0.03**	**0.13**

Table 3: Results comparison with state-of-the-art techniques.

From the carried out comparison, shown in Table 3, we found what follows. Firstly, the HF-IHU technique achieved the worst results, while LDA-Gibbs and DBSCAN-based models performed better, being able to model the corpus of tweets in an unsupervised manner, identifying an underlying structure through topic modeling and clustering analysis. Deep learning models based on the attention mechanism achieved even better results, as they can fully exploit the semantic information contained in the analyzed tweets, by learning a representative latent representation of them. Finally, HASHET proved to be the best overall model in uncovering the main hashtag-based discussion topics, achieving both the highest percentage of correct and the lowest amount of incorrect and neutral classifications, which confirms the effectiveness of the proposed approach.

6. Conclusions

This paper discussed how different methodologies of social media analysis can be exploited and executed on Cloud for extracting a rich set of knowledge about users. In particular, we focused on the extraction of knowledge from three different viewpoints: *i*) from the posts published by tourists who visit a city, we discovered the main tourist attractions and also the mobility patterns across them; *ii*) from public discussions on social media close to important electoral events, we estimated the political orientation of citizens and the outcome of a political event; *iii*) from hashtags used in posts, we discover the main topics underlying social media conversation and how users refer to them in publishing online content. Starting from a common step-by-step scheme, the presented applications have been implemented and executed on Clouds using a single development framework (ParSoDA), also reporting the main results achieved. The obtained results highlight the abundance of valuable information that can be extracted, in different application contexts, from social media data by using parallel computing approaches.

As future works, there are still many open challenges that can be faced in the upcoming years. For example, given that some data collected from social media platforms may be unreliable, new automatic techniques can be defined to select the most reliable and representative data before being used in decision-making processes. Furthermore, new algorithms for the geotagging of posts starting from the text could be introduced, as well as algorithms that suggest the best hashtags to be associated with a post to increase its diffusion within the social network. The algorithms and techniques that make it possible to distinguish the content published by bots and humans could also be improved, in order to automatically and accurately separate these two types of content. Possible critical issues that could limit the definition of new algorithms and applications for social media analysis are linked to the need to use large amounts textual and geo-referenced data, which could be difficult to obtain due to the scarce use of some social media platforms in some areas of the world, or the restrictions imposed on data acquisition.

Author Contributions: Conceptualization, L.B., R.C. and F.M.; methodology, L.B., R.C. and F.M.; investigation, L.B., R.C. and F.M.; writing—original draft preparation, L.B., R.C. and F.M.; supervision, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data that support the findings of this study are publicly available. In particular, this data was gathered using Twitter APIs available at https://developer.twitter.com and Flickr APIs available at https://www.flickr.com/services/api/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Talia, D.; Trunfio, P.; Marozzo, F. *Data Analysis in the Cloud: Models, Techniques and Applications*, 1st ed.; Elsevier Science Publishers B. V., 2015.
2. Olshannikova, E.; Olsson, T.; Huhtamäki, J.; Kärkkäinen, H. Conceptualizing Big Social Data. *Journal of Big Data* **2017**, *4*, 3.
3. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. Big Data Analysis on Clouds. In *Handbook of Big Data Technologies*; Zomaya, A.; Sakr, S., Eds.; Springer, 2017; pp. 101–142. ISBN: 978-3-319-49339-8.
4. Adedoyin-Olowe, M.; Gaber, M.M.; Stahl, F. A Survey of Data Mining Techniques for Social Media Analysis. *Journal of Data Mining & Digital Humanities* **2014**, *6*, 25.
5. Hou, Q.; Han, M.; Cai, Z. Survey on Data Analysis in Social Media: A Practical Application Aspect. *Big Data Mining and Analytics* **2020**, *3*, 259–279.
6. Batrinca, B.; Treleaven, P.C. Social Media Analytics: a Survey of Techniques, Tools and Platforms. *Ai & Society* **2015**, *30*, 89–116.
7. Belcastro, L.; Marozzo, F.; Perrella, E. Automatic Detection of User Trajectories From Social Media Posts. *Expert Systems with Applications* **2021**, *186*, 115733.
8. Belcastro, L.; Cantini, R.; Marozzo, F.; Talia, D.; Trunfio, P. Learning Political Polarization on Social Media Using Neural Networks. *IEEE Access* **2020**, *8*, 47177–47187.
9. Cantini, R.; Marozzo, F.; Bruno, G.; Trunfio, P. Learning Sentence-To-Hashtags Semantic Mapping for Hashtag Recommendationon Microblogs. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **2021**, *16*, 1–26.

10. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. ParSoDA: High-Level Parallel Programming for Social Data Mining. *Social Network Analysis and Mining* **2019**, *9*.
11. Bermingham, L.; Lee, I. Spatio-temporal Sequential Pattern Mining for Tourism Sciences. *Procedia Computer Science* **2014**, *29*, 379–389. 2014 International Conference on Computational Science.
12. Kurashima, T.; Iwata, T.; Irie, G.; Fujimura, K. Travel Route Recommendation Using Geotags in Photo Sharing Sites. Proceedings of the 19th ACM International Conference on Information and Knowledge Management; ACM: New York, NY, USA, 2010; CIKM '10, pp. 579–588.
13. You, L.; Motta, G.; Sacco, D.; Ma, T. Social data analysis framework in cloud and Mobility Analyzer for Smarter Cities. Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics, 2014, pp. 96–101.
14. Ancillai, C.; Terho, H.; Cardinali, S.; Pascucci, F. Advancing Social Media Driven Sales Research: Establishing Conceptual Foundations for B-to-B Social Selling. *Industrial Marketing Management* **2019**, *82*, 293–308.
15. wen Shen, C.; Min Chen.; chen Wang, C. Analyzing the Trend of O2O Commerce by Bilingual Text Mining on Social Media. *Computers in Human Behavior* **2019**, *101*, 474–483. doi:https://doi.org/10.1016/j.chb.2018.09.031.
16. Cesario, E.; Marozzo, F.; Talia, D.; Trunfio, P. SMA4TD: A Social Media Analysis Methodology for Trajectory Discovery in Large-Scale Events. *Online Social Networks and Media* **2017**, *3-4*, 49–62.
17. Amer-Yahia, S.; Ibrahim, N.; Kengne, C.K.; Ulliana, F.; Rousset, M.C. SOCLE: Towards a Framework for Data Preparation in Social Spplications. *Ingénierie des Systèmes d'Information* **2014**, *19*, 49–72.
18. Cuesta, Á.; Barrero, D.F.; R-Moreno, M.D. A Framework for Massive Twitter Data Extraction and Analysis. *Malaysian J. of Computer Science* **2014**, *27*, 1.
19. Zhou, D.; Chen, L.; He, Y. An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., 2015, pp. 2468–2475.
20. Hussain, A.; Vatrapu, R. Social Data Analytics Tool: Design, Development, and Demonstrative Case Studies. 2014 IEEE 18th International Enterprise Distributed Object Computing Conference Workshops and Demonstrations, 2014, pp. 414–417.
21. ECMA. ECMA-262: ECMAscript Language Specification. Fifth Edition. *ECMA (European Association for Standardizing Information and Communication Systems)* **2009**.
22. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop Distributed File System. 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). IEEE, 2010, pp. 1–10.
23. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. A High-Level Programming Library for Mining Social Media on HPC Systems. Post-Proc. of the High Performance Computing Workshop 2018, Cetraro, Italy, Advances in Parallel Computing. IOS Press, 2019, Vol. 34, *Advances in Parallel Computing*, pp. 3–21.
24. Belcastro, L.; Marozzo, F.; Talia, D.; Trunfio, P. G-RoI: Automatic Region-of-Interest Detection Driven by Geotagged Social Media Data. *ACM Transactions on Knowledge Discovery from Data* **2018**, *12*, 27:1–27:22.
25. Hansen, P.C. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Review* **1992**, *34*, 561–580.
26. Levenshtein, V.I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. 1966, Vol. 10, *Soviet Physics Doklady*, pp. 707–710.
27. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, KDD'96, pp. 226–231.
28. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems (TODS)* **2017**, *42*, 19.
29. Zheng, Y.T.; Zha, Z.J.; Chua, T.S. Mining Travel Patterns From Geotagged Photos. *ACM Trans. Intell. Syst. Technol.* **2012**, *3*, 56:1–56:18.
30. Hou, J.; Gao, H.; Li, X. Dsets-Dbscan: A Parameter-Free Clustering Algorithm. *IEEE Transactions on Image Processing* **2016**, *25*, 3182–3193.
31. Lee, I.; Cai, G.; Lee, K. Exploration of Geo-Tagged Photos Through Data Mining Approaches. *Expert Systems with Applications* **2014**, *41*, 397 – 405.
32. Belcastro, L.; Kechadi, M.T.; Marozzo, F.; Pastore, L.; Talia, D.; Trunfio, P. Parallel Extraction of Regions-Of-Interest From Social Media Data. *Concurrency and Computation: Practice and Experience* **2021**, *33*, e5638.
33. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings; Bengio, Y.; LeCun, Y., Eds., 2013.
34. Otsuka, E.; Wallace, S.A.; Chiu, D. A Hashtag Recommendation System for Twitter Data Streams. *Computational Social Networks* **2016**, *3*, 1–26.
35. Ben-Lhachemi, N.; Nfaoui, E.H. Using Tweets Embeddings For Hashtag Recommendation in Twitter. *Procedia Computer Science* **2018**, *127*, 7–15. Proceedings of the First International Conference on Intelligent Computing in Data Sciences, ICDS2017.
36. Godin, F.; Slavkovikj, V.; De Neve, W.; Schrauwen, B.; Van de Walle, R. Using Topic Models for Twitter Hashtag Recommendation. Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 593–596.

37. Li, Y.; Liu, T.; Hu, J.; Jiang, J. Topical Co-attention Networks for Hashtag Recommendation on Microblogs. *Neurocomputing* **2019**, *331*, 356–365.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.