Atmospheric Environment and Quality of Life Information Extraction from Twitter with the use of Self-Organizing Maps

ARTICLE *in* JOURNAL OF ENVIRONMENTAL INFORMATICS · JULY 2015

Impact Factor: 3.77 · DOI: 10.3808/jei.201500311

5 AUTHORS, INCLUDING:

Marina Riga
The Centre for Research and Technology, Hel…

Markus Stocker
University of Eastern Finland

Kostas D. Karatzas
Aristotle University of Thessaloniki

Mikko Kolehmainen
University of Eastern Finland

Authors

Marina Riga[1,*], Markus Stocker[2], Mauno Rönkkö[2], Kostas Karatzas[1] and Mikko Kolehmainen[2]

[1] Department of Mechanical Engineering, Aristotle University of Thessaloniki, P.O. Box 483, GR-54124, Thessaloniki, Greece

[2] Department of Environmental Science, University of Eastern Finland, P.O. Box 1627, FI-70211, Kuopio, Finland

[*] Corresponding author: Tel: +30 2310 994359; Fax: +30 2310 994176. E-mail address: mriga@isag.meng.auth.gr

Title

Atmospheric Environment and Quality of Life Information Extraction from Twitter with the use of Self-Organizing Maps

ABSTRACT

The emergence of Web 2.0 technologies has changed dramatically not only the way users perceive the Internet and interact on it but also the way they influence a community and act in real life aspects. With the rapid rise in use and popularity of social media, people tend to share opinions and observations for almost any subject or event in their everyday life. Consequently, microblogging websites have become a rich data source for user-generated information. The leading opportunity is to take advantage of the *wisdom of the crowd* and to benefit from collective intelligence in any applicable domain. Towards this direction, we focus on the problem of mining and extracting knowledge from unstructured textual content, for the atmospheric environment domain and its effect to quality of life. As the main contribution, we propose a combined methodology of unsupervised learning methods for analyzing posts from Twitter and clustering textual data into concepts with semantically similar context. By applying Self-Organizing Maps and k-means clustering, we identify possible inter-relationships and patterns of words used in tweets that can form upper concepts of atmospheric and health related topics of discussion. We achieve to group together tweets, from more generic to more specific description levels of their content, according to the selected number of clusters. Strong clusters with significant semantic relatedness among their content are revealed, and hidden relations between concepts and their related semantics are acquired. The results highlight the potential use of social media text streams as a highly-valued supplement source of environmental information and situation awareness.

Keywords

air quality, clustering, computational intelligence, k-means, semantic analysis, self-organizing maps, text mining, twitter

1. Introduction

The advent of Web 2.0 technologies introduced a new way of perceiving the Internet and interacting on it. Within the framework of social media platforms, microblogs and virtual communities, people share different types of content, such as text, images or video, provided either to the general public or towards specific user communities. A novel form of *push-push-pull* communication is established, where direct access, forwarding and further search for information takes places (Kaplan and Haenlein, 2011). Individual users easily communicate news, events, observations and personal opinions, while communities promote their services and information directly to groups of interest. These interaction processes can eventually lead to increased and shared awareness as well as to the discovery of knowledge on a personal level, among the active users.

In the current work, we focus on the atmospheric environment domain and its effect to humans' quality of life, where there is a trending involvement and participation of citizens related to environmental and quality of life aspects. With the empowerment of the general public and the provision of technology as a significant step in creating and distributing information, the idea of participatory sensing (Burke et al., 2006) has become more mature and intense nowadays. Ubiquitous data capture can be performed:

a) either explicitly, by taking advantage of the existing hardware components of commonly used devices in order to record and report directly measurements of a physical entity,
b) or implicitly, by prompting the user to share its own observations in particular domains of interest.

Applications like *PEIR* (Personal Environmental Impact Report) (Mun et al., 2009), *AsTEKa* (Skön et al., 2011) and *Air Quality Egg* (http://airqualityegg.com/) fall into the category of automated data capture, while *EnviObserver* (Kotovirta et al., 2012) and *Patient's Hay-fever Diary* (PHD) (https://www.pollendiary.com/Phd/) represent cases of direct reporting of personal observation and crowd knowledge from users. In both aforementioned ways of communication, individuals act as *soft sensors* of their surrounding environment with time and location stamp, replacing in this way the traditional, authority-centralized use of *hard sensors* and official monitoring stations (Hall et al., 2008). Sensing the environment through humans' participation can be done near real-time and on a broad scale, even in areas where static monitoring sensors do not operate. Involving citizens in quality of life aspects and making them potential contributors and users of information at the same time, will create a new, bi-directional way of communicating environmental data: from individuals to the community (practitioners) and vice versa. This is expected to transform citizens from passive receivers of information to actors and participants in environmental and quality of life related decision making processes (Karatzas, 2009).

By combining Web 2.0 and participatory sensing concepts, our motivation is to benefit from collective intelligence. We aim to mine content from users' activities and social media. Data in any form, either numeric or text, could be analyzed, structured and disseminated properly, according to applications' needs. A lot of effort has been already conducted in the manipulation of numerical air quality measurements (Niska et al., 2004; Voukantsis et al., 2010; Voukantsis et al., 2011) with significant results in mining and forecasting tasks. To the best of our knowledge, there is no published work that deals with the analysis and mining of massive user-generated text data for the atmospheric environment domain, with the aid of computational intelligence. There are authors that have recognized the importance of social media content and its impact to citizens in relation to air pollution (Cairns, 2013). There is the *AirTwitter* project that introduces the idea of potential use of social media to augment air quality event identification (Robinson, 2010). We aim to provide a straightforward methodology for analyzing and clustering text datathat are crawled and archived from social media, into classes of similar content. We do not make a real-time analysis; we are interested in analyzing individuals' personal observations expressed in free text and in mining knowledge that can be derived from massive collections of user-generated content, regarding the atmospheric environtment and quality of life related issues. The source of data in the current work is Twitter (http://twitter.com), one of the most popular microblogging services, which enables its users to post and read short text messages, known as *tweets*, of up to 140 characters. From its launch in 2006, Twitter's user base has been growing exponentially: it counts more than 500 million registered users, where 1/4 of them are considered as extremely active, while 460K accounts are created every day (TechCrunch, 2012). Tweets are publicly visible by default, while users can select to add geo-location information to their posts. Despite the controversial quality and bias of each individual tweet as a unit, the interaction and communication in social media oftenly reflects real-world events and dynamics, especially as the user base of social networks gets widen and more active in producing content about real-world events (Aiello et al., 2013).

Our work was motivated by the successful application of computational methods in mining knowledge from social media (Section 2). Still, Twitter is a challenging source of user generated text that differs from a typical document in length and structure. Usually sentences have syntactic or spelling problems, while abbreviations and semantic inconsistencies of words used may increase the complexity of text analysis (Kaufmann and Kalita, 2010). The effective adoption and adaption of text mining methodologies is considered as the main factor for successful results in analyzing the content of our collection.

The methodology followed is data-driven. We present an extensive feature selection of words and we create sets-of-words with semantically similar meaning that are used for further encoding of data into vectors. With this implementation we focus on the explicit description of the domain of interest, avoiding at the same time any redundant content that may cause irregularities or misleading results in the clustering process. We demonstrate the use of Self-Organizing Maps (SOMs) and k-means for clustering text into groups with similar patterns of used words and thus similar content and we investigate relationships between formed clusters.

The rest of the paper is structured as follows. In Section 2, we review the literature related to our work. In Section 3, we describe the technological background of our methodology as well as the main characteristics of the data (tweets) that have been collected and analyzed. In Section 4, we pinpoint and present the most significant results derived from the clustering process, at different levels of semantic description of clusters. In Section 5, we discuss the results and we highlight the information gained through this process. Finally, in Section 6, we conclude with future aspects and directions of our research.

2. Related work

Text mining, or also known as text data mining or text analysis, is the process of extracting non-trivial patterns or high-quality knowledge from unstructured text documents (Tan, 1999). It is an interdisciplinary field that involves different tasks, such as text structure, linguistic preprocessing, pattern recognition, information extraction and visualization.

In applied research, a lot of effort has been invested so as to benefit from the potential use and the commercial value of massive collections of textual data. Text mining is applied for business intelligence (Sullivan, 2001) and market trends (Zhang et al., 2011), for event detection (Li et al., 2012; Sadilek et al., 2012), political (Maynard and Funk, 2011) or commercial (Mostafa, 2013) opinion mining (Pang and Lee, 2008), security/crisis management (MacEachren et al., 2011) and decision making.

Many different methodologies have been developed for text mining tasks, ranging from traditional natural language processing (NLP) techniques and statistics to machine learning methods and their novel extensions. Statistics and probabilistic methods are used in (Hofmann 2001) for latent semantic text analysis of different corpus of documents. The authors map documents to a vector space called latent semantic space that has an order of approximately 130 dimensions. They encode the 'true' similarity hidden in the semantics of words with different terms, by forming *sets-of-words* with similar semantic meaning, but they do not address the problem of high dimensionality. Aiello et al. (2013) compare probabilistic, feature-pivot and NLP methods in order to evaluate their effectiveness in detecting events and sensing trending topics in Twitter. Dredze and Paul (2014) employ NLP methods to discover health related issues in social media, while Popescu and Pennacchiotti (2010) use supervised machine learning models for detecting controversial events from Twitter.

SOMs have been widely used to perform exploratory analysis of text data. The use of SOM for topic identification of large documents is presented in Yang and Lee (2010). Lagus et al. (2004) developed the WEBSOM method, a software system that extends the SOM principle, by using the so called 'document maps' for categorization of massive document collections.

Crooks et al. (2013) performed an analysis on Twitter feeds based on hashtags, proving that data streams from social media can be a unique source for rapid geo-located detection of earthquakes. These results motivated our research for analyzing air quality related tweets in order to identify topics of discussion and improve awareness concerning events as they occur, based on observations of users/citizens.

Costa et al. (2013) built a model that uses empirical meta-hashtag classes to group together hashtags with similar semantics, for better classification results. This approach is close to our definition of *unified concepts*, words that describe sets-of-words with similar semantic meaning.

3. Materials and Methods

We propose a straightforward methodology for mining text content, retrieved from Twitter, by combining a set of well-known computational methods that cover the following main tasks: information retrieval, lexical analysis and preprocessing, knowledge representation, clustering, information extraction and visualization. The ultimate aim of this process is to represent text data in an efficient way for further categorization in semantically similar groups of documents. The overall framework of the methodology is shown in Figure 1, while each part of it is described separately in the following subsections (3.1 to 3.3).

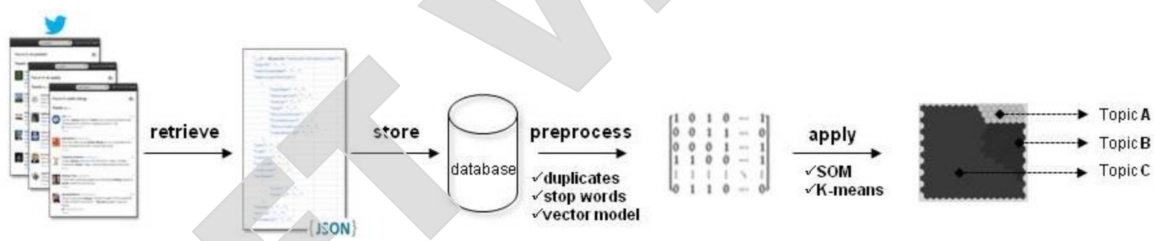


Figure 1: The framework that describes the text mining methodology.

3.1 Data retrieval via crawling

With the use of the Twitter Streaming API (https://dev.twitter.com/docs/streaming-apis), we developed a crawler that streams publicly available tweets, targeting at those tweets that are related to atmospheric environment and health related issues. The distinction of the topic was feasible by querying tweets that include specific keywords of interest, such as: air quality, air pollution, pollen allergies, air pollutants, medication, symptoms, etc. The collection period was from February till middle April 2013, and all tweets included time stamp and geo-location. We only collected tweets that were written in English language.

The data harvested through Twitter Streaming API are returned in JSON format. In order to keep the “structured” format of the information derived, we used MongoDB (http://www.mongodb.org/), a well known open source document-oriented database system. Data were parsed and then stored in a local database for further processing. An example of the JSON schema is shown in Figure 2.


```
{
  "_id": "511003ee56652adbda1c5d1a",
  "userID": 334068...,
  "userScreenName": "NameXSurnameY",
  "additionalUserInfo":
    {
      "userName": "someUsername",
      "description": "Sport, chocolate, laugh, etc",
      "status": null,
      "lang": "en",
      "followersCount": 220,
      "friendsCount": 216,
      "favoritesCount": 155,
      "isGeoEnabled": true
    },
  "text": "Love running but my asthma doesn't",
  "time": "4.2.2013 18:54:38",
  "location": "Billericay, Essex",
  "latitude": "51.63947377",
  "longitude": "0.4218109",
  "geoNames":
    {
      "city": "Essex",
      "country": "United Kingdom"
    },
  "retweet": 0,
  "streaming": "StatusJSONImpl{createdAt=Mon Feb 04 20:54:38 EET 2013, id=...}"
}
```


Figure 2: An example of the JSON schema used for storing tweets.

A total number of 52,500 tweets were collected from 34,502 unique users. The average number of words per tweet was approximately 11. The geographical distribution of all collected tweets is presented in Figure 3. Tweets are concentrated in UK and USA mostly due to language restrictions in the collection process.

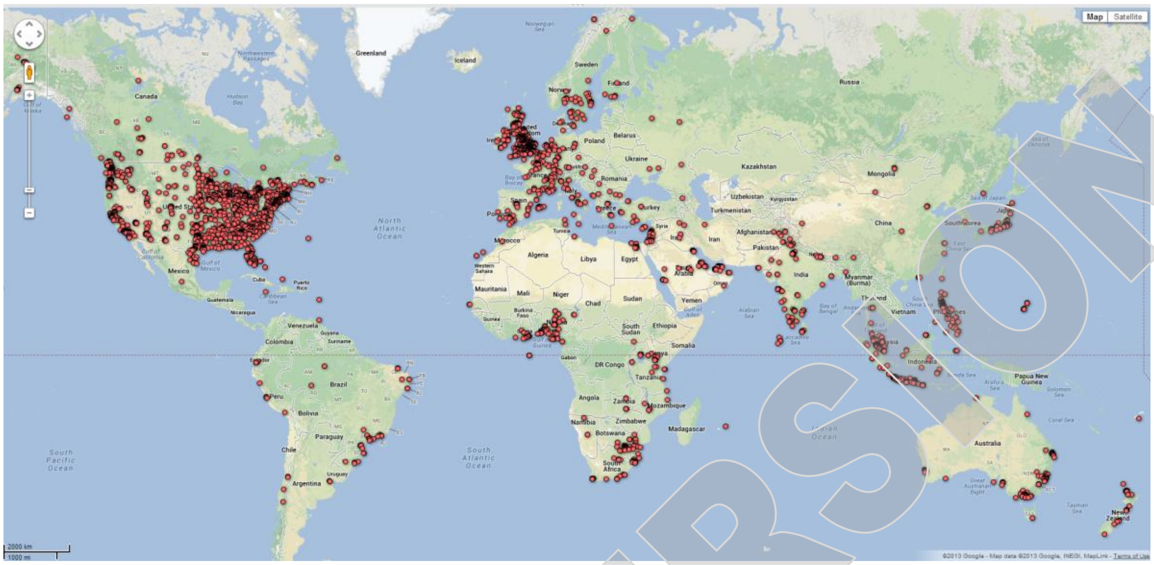


Figure 3: Tweets' geographical distribution, via Google Maps (https://maps.google.com/).

3.2 Preprocessing data

Tweets are often very peculiar in syntax and use of words, mainly due to the limitation of 140 characters and the way that users express themselves in informal text. It is common for Twitter users not to strictly follow the syntactic and grammatical rules of written language; they rather tend to use extensively abbreviations, acronyms, contracted words and emoticons, or even create new words in order to express their opinion in the most compact way.

We apply two levels of preprocessing: at first, we clean the text from unnecessary content. Then we turn the text into numerical data in an efficient way, in order to encode the represented information properly and feed it into selected computational methods for further analysis.

3.2.1 Removing noisy data

Text analysis refers to the process of deriving high-quality information from text. The process is directly related to the words used in text. For optimized results, it is important to reduce the unnecessary (noisy) content from each tweet, such as:

a) hyperlinks and re-tweet references (denoted with “http://” and “@” entities, correspondingly),
b) any punctuation or non-alphabet character,
c) words with one or two characters, and
d) English stop-words, meaning words that are frequently used in any text (like articles, pronouns, prepositions, etc.).

After the first level of preprocessing, the average number of words per tweet was 9, while the removal of duplicates resulted in a total number of 44,888 unique tweets, being available for further analysis.

3.2.2 Representing text documents with vector space model

A common methodology for the analysis and representation of text data in a structured way is to adopt the vector space model, also known as the *bag-of-words* model, introduced by Salton (1975). The general version of this methodology represents each text document d_i as a row vector of t_j features, each of which corresponds to a separate term (word), as shown in Equation (1).

$$d_i = [\, t_1,\ t_2\,, \dots, t_n] \tag{1}$$

If the term exists in the document, its value in the vector can be either one, corresponding to the occurrence in text, or any non-zero natural number, corresponding to the frequency of its occurrence in text. The whole database will be transformed into an *mxn* matrix, of *m* documents (tweets) and *n* terms, with zero and non-zero values, for the frequency of occurrence or the non-occurrence of each term, respectively (Berry, 2004).

Usually, the collection of words is unordered, thus no information about word position in the document is taken into account.

A bag-of-words may include all words used in the total collection of documents being investigated. Such an implementation would lead to a high-dimensional, sparse matrix with few combinations of words per document. The redundant information carried throughout the process, makes the task of extracting similar patterns from data more difficult for clustering algorithms. Dimensionality reduction is of great importance in order to manipulate efficiently massive amounts of text data and improve the solution to the problem at hand. A different implementation of a bag-of-words would include only the *n* most frequent words in the database, but this representation method would leave out words of high interest that might not appear often in tweets. In our approach, we select to include in the bag-of-words model the most frequent (above a threshold) words used in our collection of tweets, combined with some less frequent words that were empirically considered as important within the context of our domain of interest. We should prompt here that we have already removed any data considered as noisy (Section 3.2.1).

For dimensionality reduction reasons, it became evident that we had to take into consideration the meaning (semantics) of words, in the bag-of-words concept. For example, words like *flu*, *cold* and *sick* can be considered as one term, under the label of *medical condition* since these three words describe in general a medical condition which could be related to existing environmental conditions.

Hence, by having created the initial bag-of-words, we define sets of similar words that are not necessarily synonyms at the same time, but can be considered as having the same semantical meaning under the context of atmospheric environment and its effect to humans' quality of life. Each set of words *S* includes a different number of similar words that describe explicitly an upper semantic level of related content (unified concept). All words and sets formed define the *bag-of-sets of words*, as given in detail in Table 1. The selection of words and related unified concept has been done empirically, establishing thus implicitly the semantic context within which the analysis is going to be performed. We state that the current form gives a well-defined context of the most representative aspects that describe atmospheric environment and health related issues in an informal way. This context can be re-produced in any other combination of words, thus changing the interpretation of the analysis' results. Any change in the current definition of sets of words affects the semantic interpretation of results and the potential groups created from the clustering process.

Table 1: Sets of words, with similar context, forming the bag-of-sets of words for the analysis

#	Words in set	Unified concept
1	air, atmosphere, atmospheric	air
2	eye, nose, skin, mouth, throat, stomach, head, lungs, face, nasal, heart, respiratory, chest, body	body part/organ
3	pollution, pollute, pollutants	pollution
4	itch, itchiness	itch
5	sneeze, sneezing	sneeze
6	cough, coughing	cough
7	run, running, runny nose*	runny
8	flu, cold, sick, ill, headache, asthma, rhinitis, disease, fever	medical condition
9	city, town, urban, area, region, outdoor, outside, country	outdoor
10	alert, alarm, warning	warning
11	quality, condition	quality
12	breathe, breathing, breath	breathe
13	indoor, house, home, office, inside, school	indoor
14	bad, hard, severe, difficult, sore, suffer, poor, weak, risky	bad/poor
15	problem	problem
16	not bad*, good, great, well	good/well
17	today, now, tonight, morning, evening, night	time/now
18	seasonal, hourly, daily, weekly, monthly, yearly, hour, day, week, month, year	time/period
19	allergy, allergic, sensitive	allergies

20	food, feed, eat	food
21	pollen	pollen
22	hospital, doctor, medical, clinic	hospital
23	medication, medicine, pills	medication
24	high, huge, big, heavy	high
25	levels, standards, index	level
26	citizens, public, people, national, men, women, children, elder	people
27	cars, vehicles, motor, bike, bus, taxi	car
28	water, garbage, waste, soil	physical entity
29	smog, humidity, wind, dust, rain, weather	weather conditions
30	spring, summer, autumn, winter	season
31	pets, dogs, cats, birds	pets
32	particulates, particles, ozone	pollutants
33	horrible, hell, hate, crazy, killing, ugh	bad feelings
34	happy, funny, yeah	good feelings
35	wow, haha, lol	making fun

[*] In the preprocessing phase, the case *wordA followed by wordB* was converted to “wordAwordB” (words were collapsed without whitespace) in order to be distinguished from the cases where the same single words exist in free order.

Given the specific bag-of-sets of words, we can examine the occurrence (one) or non-occurrence (zero) of each unified concept in the text. A x^{th} tweet (T_x) can be represented as a vector of the form in Equation (2):

$$T_x = \left[Value_{S_1}\ \ Value_{S_2}\ \ Value_{S_3}\ \ \ldots\ \ Value_{S_n}\right] \tag{2}$$

where S_k is the k^{th} set of words that includes v distinct words and the corresponding value for this set will be: $Value_{Sk}$ = 0, if none or $Value_{Sk}$ = 1, if at least one of the v words in S_k exists/occurs in the examined tweet.

The overall occurrence matrix M (i.e the matrix including all information on the occurrence of the words in sets, as in Table 1) will be of the form:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & 1 & 1 & \cdots & 0 \end{bmatrix}$$

with *m* number of documents (tweets) per row and *n* number of sets of words per column.

By following the aforementioned methodology, we first analyze each tweet in its corresponding words and then we transform its textual representation into a binary vector with zero/one values according to whether a word from a predefined set of words (Table 1) belongs or not to the tweet. All tweets in the database are converted to array vectors with zero/one values and can thus be considered as vectors in a multidimensional space. Therefore, vectors that are close to each other in a topological manner (according to distance metrics applied in the learning phase) can be regarded as representatives of documents with similar word patterns. *Semantic similarity* can also be assigned to documents based on vector operations, due to the fact that each set of words has a general semantic interpretation. With the transformation of text into binary data and the use of a similarity metric, we move from lower levels of co-occurrence of words to upper levels of a realistic semantic interpretation of documents that have similar patterns of occurrence of words and thus analogous semantic content and similarity.

3.3 Learning, Mining and Evaluation Methods

The way text documents are encoded is of great importance for the performance of any learning and mining method. In the current paper, the key idea relies on a two-level approach, based on SOMs and k-means clustering. At the first level we apply SOM for its advantages of dimensionality reduction, data compression and topological relation properties. At the second level we feed the new vector space (weights of SOMs) to k-means clustering in order to produce clusters with similar characteristics. Clusters are formed by learning at the same time the structure of data from SOM weights and its segmentation using both distance and density information (k-means and Euclidean distance). The number of clusters formed is based on the Davies-Bouldin index, while additional methods, like Sammon mapping and silhouette metric, are taken into account to evaluate and validate the efficiency of the clustering results. The methodological background is described in the following subsections.

3.3.1 Self-Organizing Map

Kohonen's Self-Organizing Map (SOM) is a competitive, unsupervised learning method that maps high dimensional data into a low dimensional space by preserving their spatial correlation (Kohonen, 1990). SOMs consist of neurons, each of which is a set of weights (or prototype vectors) that correspond to the projection of the instances of a high-dimensional dataset onto a usually 2-dimensional grid. The neuron whose weight vector is most similar to the input is called the best matching unit (BMU) and this can be determined based on criteria like the Euclidean distance.

Since its first introduction, a lot of research has been conducted concerning SOMs and many different implementations have been proposed, for various application domains. Our research has been inspired by Ritter and Kohonen (1989) as well as Honkela (1997). The authors introduce a methodological background for SOMs that use word categories (so called self-organizing semantic maps) in order to preserve semantic relationships reflected in the data according to their relative distances in the map.

The value of SOMs lies in easily interpretable results through the investigation of topologically similar areas so as to discover relations between examined parameters. A matrix of occurrence or co-occurrence of words in documents can be visualized by SOMs as a topological grid that represents every correlation between sets of words in the same topological space. We take advantage of SOMs' potentials for dimensionality reduction and easily interpretable results, in order to form clusters of similar patterns of words' occurrence and extract topic categories from the underlying information.

3.3.2 K-means clustering

K-means is a well known non-hierarchical clustering technique which partitions a set of observations (data points) into *k* numbers of clusters (Hartigan and Wong, 1979). In the initial step of the algorithm, *k* "means" are randomly defined as centroids of clusters within the data domain. Then, an iterative refinement process is performed, where clusters are formed by associating each data point with the closest centroid (nearest mean based on Euclidean Distance) and centroids are updated by calculating the new mean values of data points belonging to each cluster. The iterative process continues until a convergence criterion (minimal decrease of total sum of squared errors, minimal reassignment of data points into different clusters, maximum number of iterations, etc.) is met.

The efficiency of k-means is highly depended on the determination of a suitable number (*k*) of clusters. This number should be somehow close to the number of "natural" clusters that are present in the data, otherwise:

a) if *k* is higher than needed, we might get "garbage cans" (clusters with no significant distinction between them) or empty clusters, due to the force of splitting, or
b) if *k* is lower, we will get more generic divisions.

Since there is no prior knowledge of what number to select, we adopt an internal evaluation criterion of clustering results: we make use of the Davies-Bouldin (DB) index for different numbers of clusters, which is calculated according to Equation (3) (Davies and Bouldin, 1979):

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j,j\neq i} \frac{S_i + S_j}{d_{ij}} \quad (3)$$

where k is the number of clusters. The within (S_i) and between (d_{ij}) cluster distances are calculated using the cluster centroids as follows:

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i}\|x - m_j\| \quad (4)$$

$$d_{ij} = \|m_i - m_j\| \quad (5)$$

where m_i is the centroid of cluster C_i, with $|C_i|$ being the number of data points that belong to the cluster. According to Equation (3), it becomes evident that the objective of DB-index is to minimize its value and, consequently, to have the minimum "within-cluster" dispersion and the maximum "between-clusters" separation (Davies and Bouldin, 1979). Here, k-means and DB-index is combined with SOM as a two-step approach for clustering, in order to define distinct areas of compact clusters in map representation.

3.3.3 Sammon mapping

Whereas the visualization of high dimensional data with the use of SOMs can reveal topological relations among data points and clusters, it doesn't give any interpretation of their actual closeness and inner distance between the neurons of the map. Usually, additional coloring schemes, like the unified distance matrix (U-matrix) (Ultsch and Siemon, 1990), can be used in the analysis to show the boundaries between the formed, and depict low or high distance estimation between them.

In order for a map to depict the data structure, the distance property must be retained along with the topology. Sammon mapping (Sammon, 1969), is a non-linear methodology that can preserve the topology of the input data as well as the distances between inter-points within the map. This projection of data can reveal cluster structure and tendency, showing how "strong" or "loose" the connections are between data points in each cluster and between clusters.

Sammon mapping is an iterative process with random initialization of the d-space configuration that performs pseudo-Newton minimization of the error function given in Equation (6), also known as Sammon's stress, by calculating the inter-point distances d^*_{ij} between the i^{th} and the j^{th} object in the original space, and the distance d_{ij} between their projections (Sammon, 1969):

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j}^{N} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{6}$$

We apply Sammon mapping on the actual clusters derived from SOM and k-means, by projecting the weights of neurons of the SOMs and visualizing their Euclidean distance in the 2-dimensional space.

3.3.4 Silhouette metric

Silhouette values can act as a metric of validity of the already assigned clusters, with respect to factors like: (a) cohesion (how compact each cluster is), and (b) separation (how clearly clusters are separated from each other). A silhouette value *s*(*i*) for an object *i* is calculated according to the Equation (7):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad -1 \le s(i) \le 1 \tag{7}$$

where *a*(*i*) is the average dissimilarity of object *i* to all the other objects within the cluster in which *i* falls, and *b*(*i*) is the minimum average dissimilarity of object *i* to all objects of each assigned cluster, thus defining the neighboring cluster of *i* (Rousseeuw, 1987). Silhouette values can be used as an indicator of how well objects have been classified in a cluster, or which objects are misclassified. By plotting the silhouette values of objects per cluster, we can distinguish compact and clearly separated clusters from non-tight ones. If a cluster contains many objects with low silhouette values, or even negative ones, this indicates that the cluster is not well-separated from other clusters. On the contrary, the more close to 1 the mean of silhouette values of all objects in a cluster, the better separation from its neighboring clusters has been achieved. Similarly, we can follow the mean silhouette value per cluster and conclude to corresponding results.

4. Results

In the preprocessing phase, tweets are encoded as a high dimensional matrix of 44888x35 zero/one values in relation to the non-occurrence/occurrence of 35 different sets of words (Table 1) in posts. We apply the SOM algorithm by using the SOM Toolbox 2.0 (http://www.cis.hut.fi/somtoolbox/) for Matlab. SOM neurons were positioned in a 16x16 orthogonal grid arranged on a hexagonal lattice of a 2-dimensional map. With SOM representation, we manage to reduce the dimensionality of the initial encoded data up to 99.98%. The overall results of the mapping process are shown in Figure 4.

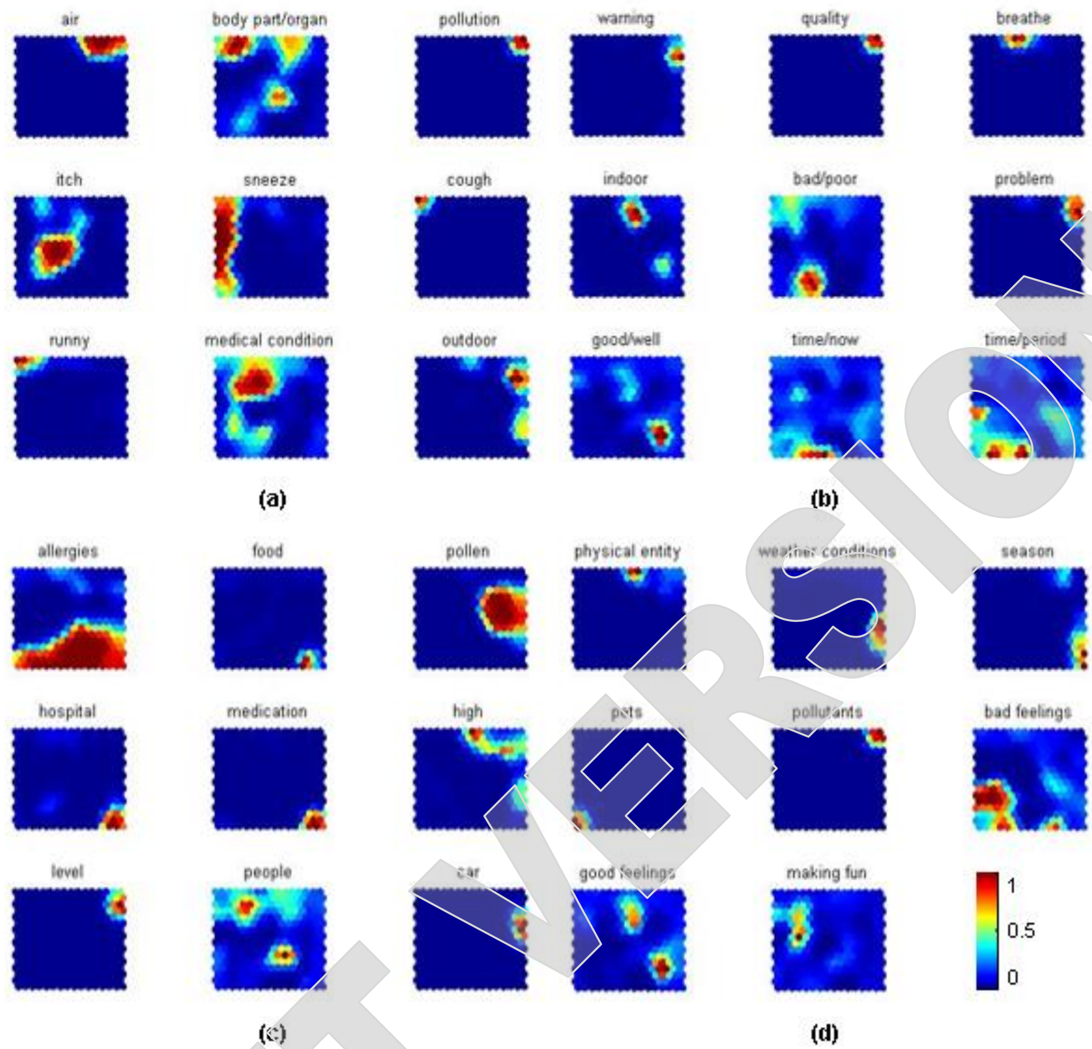


Figure 4: Four sets (a, b, c, d) of SOMs for 35 sets of words in total, derived from the analysis of feature vector values of tweets with the use of SOM algorithm.

Each map has 256 (16x16) colored neurons (cells) that correspond to the weight values, calculated by SOM, for the corresponding examined variable. According to the given scale in Figure 4, dark blue color corresponds to the lowest value (zero) of the variable, meaning the non-occurrence of a set of words in text. On the other hand, dark red color corresponds to the highest value (one) of the variable of interest, meaning the occurrence of a set of words in text. Other colors apart from the aforementioned two correspond to intermediate weight values between the range [0,1]. The interpretation behind intermediate values is that such neurons do not clearly belong to any of the two defined cases that zero/one values represent. Instead, the closer the values are to zero/one, the most certainly they belong to the corresponding cases.

The visual inspection of SOMs in Figure 4 and the topology characteristics of each map can reveal qualitative information about the use of each variable and the inter-correlation between them. The latter can be examined by the co-occurrence (values equal to one) of two or more sets-of-words in posts. Thus, similar areas per variable should be investigated per case. For example, we can imply that *allergies* are related to *pets* within the tweets being analyzed. This can be supported by inspecting the bottom-left corner of the corresponding maps in Figure 4. This area for both variables, *allergies* and *pets*, has red colored cells, meaning that there are a number of tweets that use words from both aforementioned sets in the same sentence. On the other hand, we can spot negative relation, like for example between *pets* and *pollutants*. According to Figure 4, there are no data with high values in both variables' visualizations. As expected, concepts like pets and pollutants are not related to each other.

Through this process, we can state similar possible correlations (positive/negative) among words and unified concepts, by investigating topologically related areas of SOM for different sets of words. Though, it is hard to form manually groups of data (clusters) with similar behavior from these distinct maps. The U-matrix in Figure 7a can give a rough representation of the "natural" clusters that exist in the dataset. It represents the Euclidean distance between the high dimensional data and the SOM weight vectors. The colored scale next to it corresponds to the distance values. Cells with blue color in the U-matrix can be considered as clusters (minimum distance between neurons), while cells with colors related to higher values of the scale show the boundaries between clusters (maximum distance between them). According to the formed U-matrix we can stand out approximately 5-7 clearly defined clusters but there are also some additional clusters characterized by fuzzy districts.

As we do not have clear knowledge on the number of clusters within the data, it is important to experiment with a range of values for *k*. In order to select the number of clusters that sufficiently and discretely divide the area of SOMs with the use of k-means, we made use of the DB-index within a big range of *k* values (from 1 to 50), as presented in Figure 5. It is acknowledged that the lower the DB-index value the better the clustering configuration, in terms of compactness and separation (Kovács et al., 2006). Here, the global minimum of DB-index is for $k = 22$ but there are three local minima that appear before it, for $k = 3$, 10 and 16, which are of special interest. The DB-index is not highly varying from $k = 10$ to $k = 27$, while within the range that these two values define, the global minimum occurs. After $k = 27$, sudden changes of DB-index appear, which can be used as a criterion that the situation from this point is unstable. The clustering algorithm is forced to do splitting of already well-defined clusters, thus producing “artificial” clusters with loose connections between data points. Thus, the global minimum can be considered as the upper limit of *k* that provide well defined clusters in the examined case. In the following subsections, we summarize clustering results for the aforementioned local minima and the global one, and we describe the evaluation process.

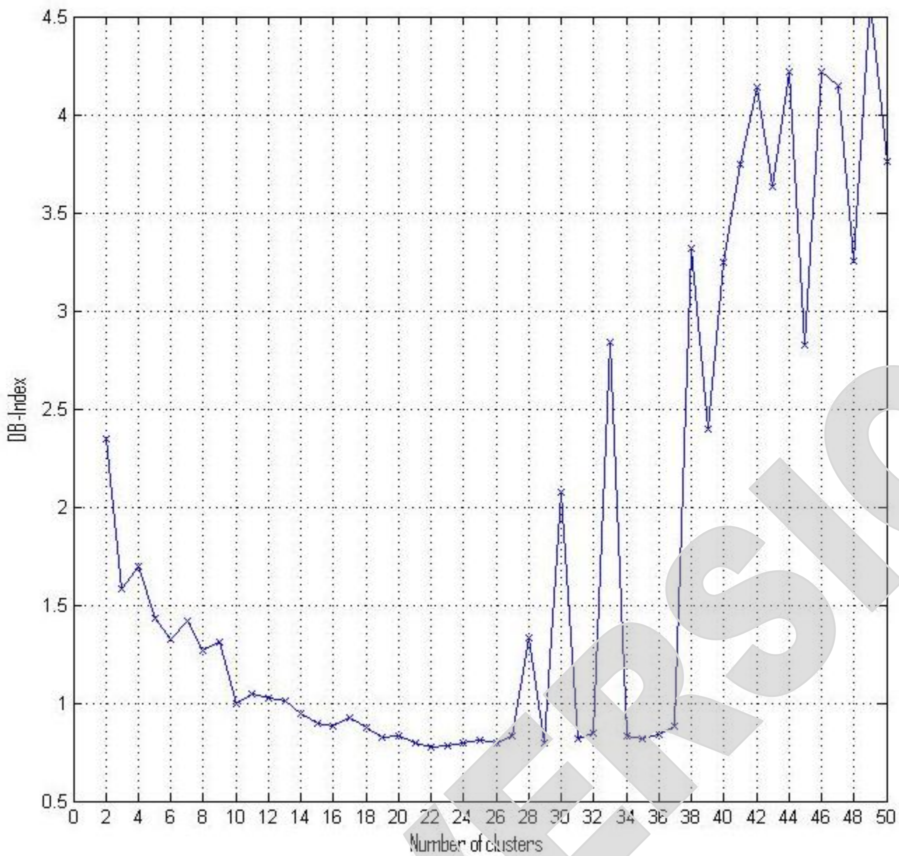


Figure 5: DB-index values for different numbers of *k* clusters.

4.1 Clustering results for different numbers of clusters

By applying k-means clustering to the weight vectors of the SOM we can form *k* number of clusters, based on distance metrics. Each neuron of the map will correspond to a specific cluster. Since neurons are linked to high dimensional data, we can easily define the actual data behind the clustering results.

For each C_{ifromk}, meaning the i^{th} cluster of *k* in total, we can identify the instances (tweets) it includes. We can thus calculate the total number of instances per cluster, as well as the percentage value of occurrence of each set of words per cluster. The latter values are depicted in Figure 6.

According to Figure 6, each cluster has a peak value for a specific set of words. For example, almost 90% of tweets that are included in cluster C_{2from3} include in their text one or more words from the set of words that corresponds to the unified concept *air*. Clusters C_{1from3} and C_{3from3} include tweets with words related to *pollen* and *allergies* respectively. Their percentage of occurrence of corresponding words is lower than the one calculated for cluster C_{2from3}, but still of significant level (73% and 46% approximately). The name of the dominant concept can clearly define the label of the formed clusters of tweets, and can thus provide with the general topic of discussion per case. In other words, through this clustering process, we map text data into concepts and their semantics.

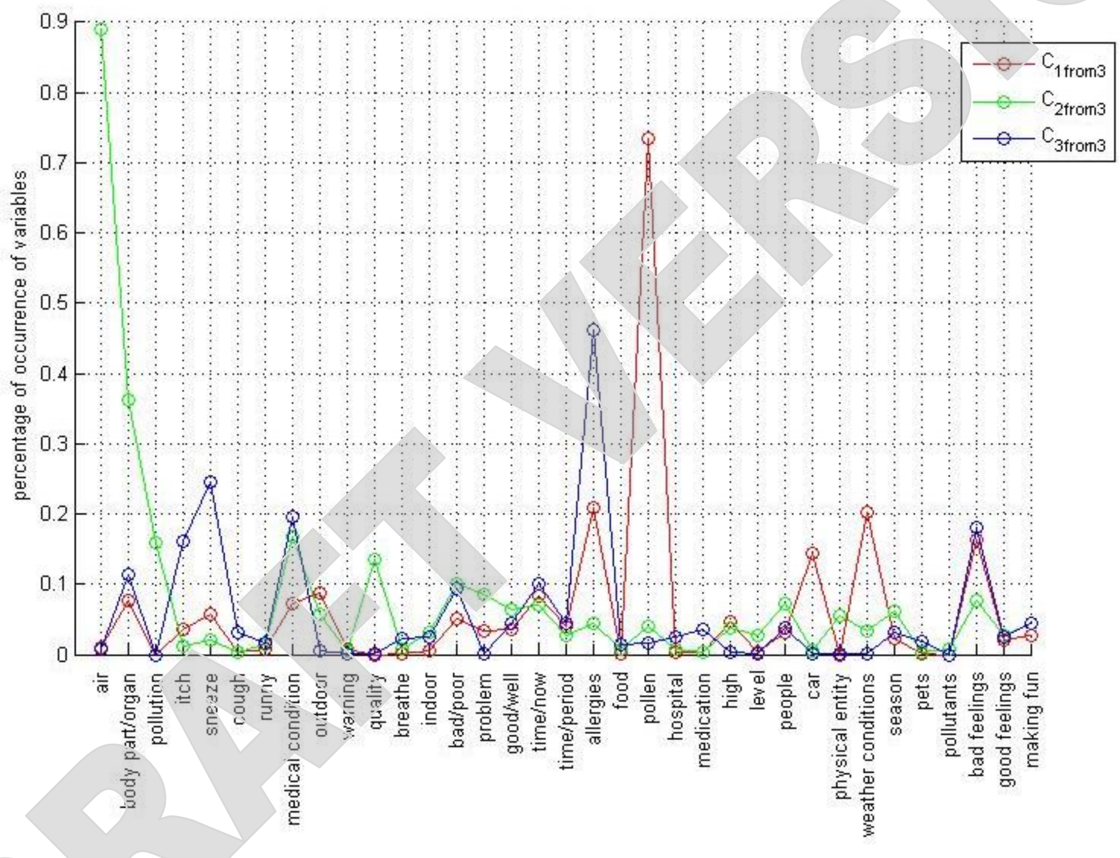
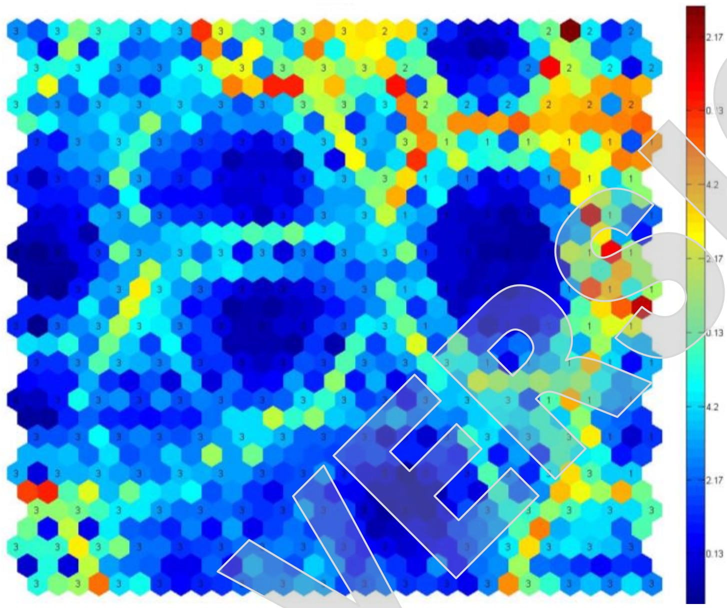


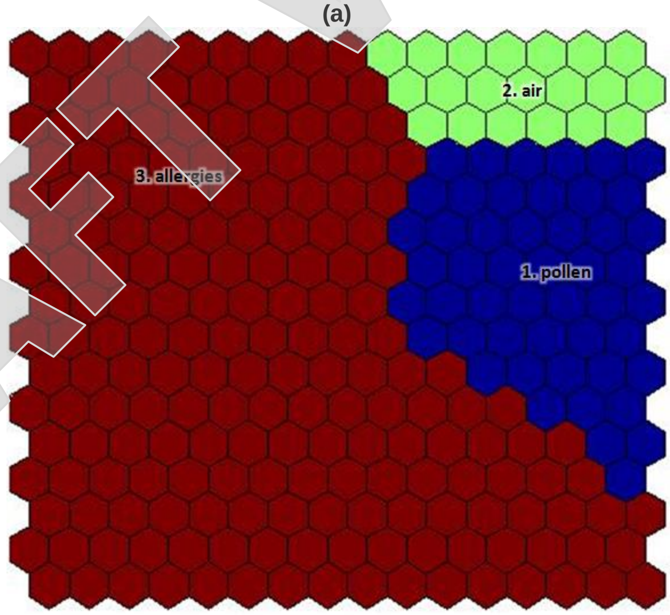
Figure 6: Percentage of occurrence of each concept (set of words) for 3 formed clusters.

By using the U-Matrix along with the results from k-means clustering, we can draw the identification number of each cluster per node in the map, as shown in Figure 7a. An integrated area per cluster is formed and we can get the visualization of clusters' content and coverage by adding the corresponding labels of description on the map (Figure 7b).

The cluster named *allergies* covers the biggest area in the map, as it includes 80.35% of tweets, while *pollen* is the second biggest cluster with 14.19% of tweets.



(a)



(b)

Figure 7: (a) U-Matrix and clusters per neuron, (b) areas and labels of 3 formed clusters on 2-dimensional map.

Up to this point, the three distinct concepts named *allergies*, *pollen* and *air* describe at a general level the content of the database. We thus gained a conceptual level from unstructured data, with semantic interpretation as well. By dividing the areas into more clusters, as the DB-index suggests, we can obtain a more detailed overview of the concepts that are discussed in tweets.

When applying k-means clustering of SOMs for k = 10, different clusters with different content are formed, which is depicted in form of percentage values of occurrences of sets of words per cluster, as shown in Figure 8. Concepts per cluster are also summarized in Table 2.

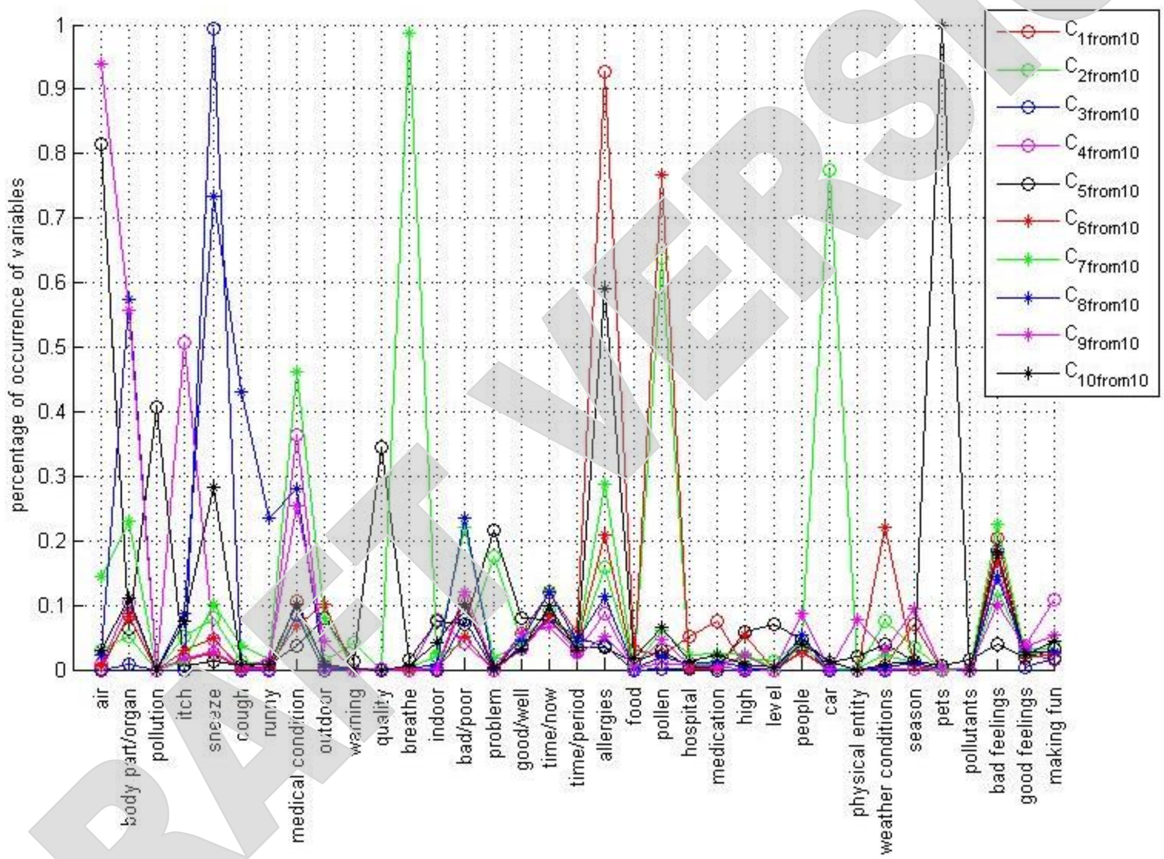


Figure 8: Percentage of occurrence of each concept (set of words) for 10 formed clusters.

Table 2: Defining concepts that describe 10 clusters, based on the percentage of occurrence of sets of words in each cluster

Cluster ID	Instances (%)	Set of words based on their occurrence in clusters (% values)			
		(90, 100]	(70, 90]	(50, 70]	(30, 50]
$C_{1from10}$	35.34%	allergies	-	-	-
$C_{2from10}$	2.63%	-	car	pollen	-
$C_{3from10}$	13.36%	sneeze	-	-	-
$C_{4from10}$	23.20%	-	-	itch	medical condition
$C_{5from10}$	2.15%	-	air	-	pollution
$C_{6from10}$	11.38%	-	pollen	-	-
$C_{7from10}$	1.87%	breathe	-	-	medical condition
$C_{8from10}$	5.33%	-	sneeze	body part/organ	cough
$C_{9from10}$	3.31%	air	-	body part/organ	-
$C_{10from10}$	1.44%	pets	-	allergies	-

Following a similar way in the interpretation of results, we can conclude that there are clusters with more than one dominant concept that cover at least 50% of the corresponding tweets. In other words, clusters' content can be defined by a combined label of both dominant concepts, where applicable. With the use of the combined label we get a more detailed semantic interpretation of the content and the tweets included in each cluster.

Through this refinement, from 3 to 10 numbers of clusters, cluster named *allergies* ($C_{1from10}$) remains the biggest one, including approximately 35% of total tweets, while it was reduced compared to the results obtained for k = 3. From its division, new clusters arose, namely *sneeze* ($C_{3from10}$), *itch & medical condition* ($C_{4from10}$), *breathe & medical condition* ($C_{7from10}$), *sneeze & body part/organ* ($C_{8from10}$), and *pets & allergies* ($C_{10from10}$), forming clusters with more detailed description of the content. Similarly, cluster *air* from the previous clustering results was split into two sub-clusters: *air & pollution* ($C_{5from10}$) and *air & body part/organ* ($C_{9from10}$). Finally, from cluster *pollen* a new sub-cluster was created, named *car & pollen* ($C_{2from10}$). Labels and areas of clusters in the topological space are shown in Figure 9.

Figure 9: Areas and labels of 10 formed clusters on 2-dimensional map.

From the so far analysis, it becomes evident that by increasing the number of clusters, new sub-clusters arise with lower number of instances. In addition, labels derived give more detailed description of clusters' content. Based on the dominant sets of words presented in groups of tweets, we arrive to more detailed results for 16 and 22 numbers of clusters, as shown in Figures 10 and 11 respectively.

Figure 10: Areas and labels of 16 formed clusters on 2-dimensional SOM.

Figure 11: Areas and labels of 22 formed clusters on 2-dimensional SOM.

4.2 Evaluation of clustering results

As a next step, we quantitatively evaluate the clustering results, by using the silhouette metric. We can investigate how strong or loose is the relationship between the instances that form each cluster, by following the mean as well as the minimum silhouette value per cluster. The optimum value of the metric equals to 1. All calculated values for cases with 3 and 10 numbers of clusters are summarized in Table 3.

Table 3: Mean and minimum silhouette value per cluster, for different values of k clusters

Cluster ID	Name/ concept	Mean silhouette value	Min silhouette value
(a) SOM and k-means for *k* = 3			
Cluster ID	**Name/ concept**	**Mean silhouette value**	**Min silhouette value**
C_{1from3}	pollen	0.3680	0.0349
C_{2from3}	air	0.4563	0.1154
C_{3from3}	allergies	0.3559	-0.0085
(b) SOM and k-means for *k* = 10			
Cluster ID	**Name/ concept**	**Mean silhouette value**	**Min silhouette value**
$C_{1from10}$	allergies	0.3384	-0.2542
$C_{2from10}$	car & pollen	0.4403	-0.1371
$C_{3from10}$	sneeze	0.6402	0.2009
$C_{4from10}$	itch & medical condition	0.3143	-0.1522
$C_{5from10}$	air & pollution	0.7499	0.5789
$C_{6from10}$	pollen	0.5532	0.1742
$C_{7from10}$	breathe & medical condition	0.5922	-0.0280
$C_{8from10}$	sneeze & body part/organ	0.4144	0.1427
$C_{9from10}$	air & body part/organ	0.7803	0.4646
$C_{10from10}$	pets & allergies	0.5941	-0.1097

In the first clustering analysis (where $k = 3$), all three clusters have low mean silhouette values (Table 3a). Also, their minimum silhouette value is close to zero. These results denote high distances between neurons within clusters and the connections between instances of clusters tent to be loose. Such clusters are unstable and more likely to split in a next clustering process, with increased number of clusters.

When we increase the number of clusters ($k = 10$), the situation is different (Table 3b): we still get some weakly connected clusters, like $C_{4from10}$ (*itch & medical condition*) which has the lowest mean silhouette value (0.3143) and it is more likely to be split in a next division. On the other hand, there are clusters, like $C_{5from10}$ (*air & pollution*) or $C_{9from10}$ (*air & body part/organ*) that have high mean silhouette value (closer to 1) and thus these clusters are considered to be tight and their instances are strongly connected. We prove that such clusters remain stable in the next clustering process.

By increasing once more the number of clusters from 10 to 16, we get new clustering results. Clusters with low mean silhouette value in the previous step were split or transformed. For example, cluster *itch & medical condition* ($C_{4from10}$) became two separate clusters with one dominant concept per each: *itch* ($C_{16from16}$) and *medical condition* ($C_{1from16}$). On the contrary, clusters with high mean silhouette value, like *air & pollution* ($C_{5from10}$) and *air & body part/organ* ($C_{9from10}$) remain intact (now as $C_{8from16}$ and $C_{14from16}$ correspondingly) from the current clustering process, including almost the same number of instances in both divisions.

By plotting clusters in two dimensions using Sammon mapping, we obtain the distribution of data points per cluster and we can visualize the distance properties among them. We demonstrate Sammon mapping for the case where k = 16 clusters (results are presented in Figure 12). Following this visualization, we can distinguish three main categories of clusters:

a) clusters that are clearly separated, like $C_{8from16}$ (*air & pollution*), $C_{13from16}$ (*pollen*) and $C_{14from16}$ (*air & body part/organ*),
b) clusters that are overlapping, like $C_{15from16}$ (*bad feelings & medical condition & allergies*), and
c) clusters whose data points are spread over the map, even though they belong to the same cluster, like $C_{2from16}$ (*allergies*) and $C_{4from16}$ (*indoor & allergies*).

Clusters of categories (b) and (c) are more likely to be split in the next clustering phase. This fact is also confirmed by the corresponding low silhouette values of these clusters.

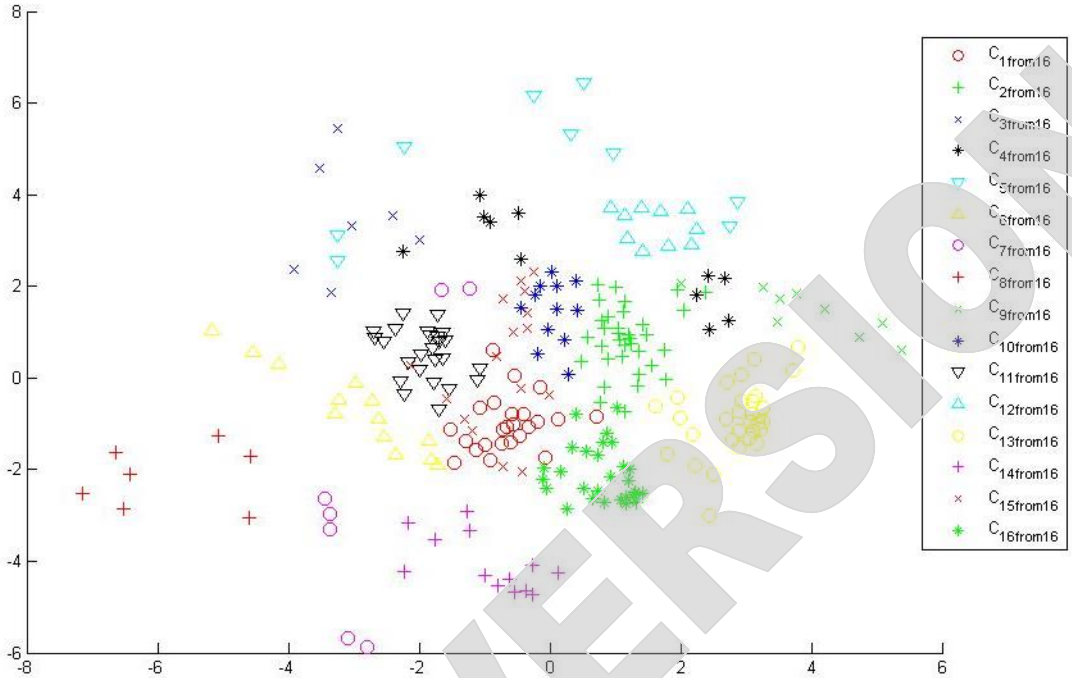


Figure 12: Sammon mapping of SOM weights of 16 clusters in 2-dimensional space.

Similarly, by incrementing the number *k* for k-means from 16 to 22, those clusters that were more likely to split, generated new, smaller and more detailed ones in the next level. Overall, the Silhouette Coefficient (SC), calculated as the mean of all silhouette values per clustering level, increases (see Table 4), meaning that we achieve to get more stable clusters in each clustering step.

Table 4: Silhouette coefficient values per different total number of clusters

Total number of clusters	Silhouette coefficient
3	0.3661
10	0.4512
16	0.5448
22	0.5669

5. Discussion

Our results demonstrate that SOM is an efficient and convenient method to structure, analyze and interpret user generated text from social media. The representation of text data with vectors, based on the occurrence of specified sets-of-words in text, can transform original text into a binary matrix. Sets-of-words are assigned with a unified concept that maps them into a general semantic meaning. Clusters may appear with one or more dominant concepts that describe their content. We assign topics of discussion in groups of tweets and we enrich formed clusters with conceptual semantics, based on their dominant concepts.

Through the performed analysis, we derive to a clear visual representation of the actual content of the text database (see Figure 11). Clusters' labeling corresponds to the real content of the text data included and to the semantic knowledge implicitly assigned to each concept. The size of each cluster reflects the number of instances included on it. We identified big clusters, like *pollen*, *allergies*, *itch* and *sneeze*, where general concepts uniquely describe their content. We also identified smaller, more descriptive clusters, such as *pets & allergies*, *car & pollen*, *air & pollution* and *body part/organ & sneeze* that combine more than one concept in order to characterize their content. Such clusters acquire labels that state clearly and in more detail their content (topic of discussion). The potential of determining the relationships between words as well as correlations among concepts is a highly-valued prospect of this method. Not only logically apparent concepts, but also new, non-typical relations can be discovered. A good example of unexpected interrelation learned is the cluster *car & pollen*. Its name defines two dominant concepts that describe the content of the cluster. However, such a correlation does not seem to make "sense". A thorough investigation of the actual data reveals that pollen is related to cars in a functional way, i.e. people are expressing their disappointment or frustration because their vehicle is covered with pollen during the pollen season, and requires cleaning. We should also note here that both concepts were introduced into the bag-of-sets-of-words due to the increased frequency of use in the collection of documents, even though *car* seems to be unrelated to the context of our research.

Our results also highlight the ability of SOM and k-means to organize practically related topics of discussion into neighboring clusters. As an example, the main cluster *allergies* adjoins with concepts of similar content, like *allergies & body part/organ*, *allergies & bad/poor & medical condition*, *food & allergies*, while *air* related concepts are in a different area of the map.

The number *k* of clusters that divide adequately and efficiently the actual data is based on the DB-index value. By investigating different numbers for *k* and not only the one that corresponds to the global minimum of DB-index, we derive different levels of description of clusters' content. It is characteristic that the lower the number of clusters the more generic the resulting concepts. With increasing number of clusters, these generic clusters are subdivided into more concrete ones. Hence, given the content of a database, we can build a *hierarchy of concepts*, as shown in Figure 13. The derived hierarchy can be considered as a tree diagram, where upper nodes are the generic description concepts and leaf nodes are the detailed concepts. Concepts of low levels describe the content by combining concept nodes from upper levels of the tree. Here, the hierarchy is given only for the three out of four levels of the clustering process that we performed, due to clarity reasons.

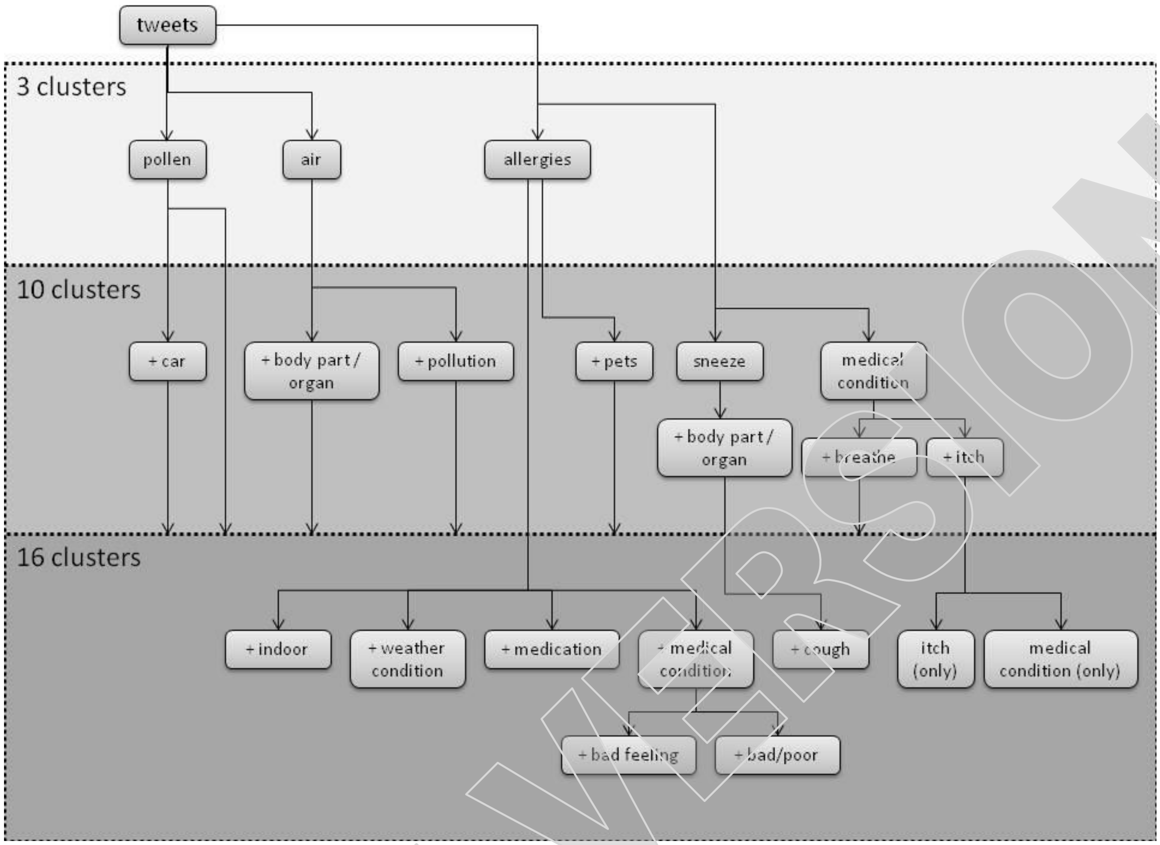


Figure 13: The hierarchy of concepts derived after three levels of clustering.

With Sammon mapping we can visualize the relationship between individual neurons and defined clusters, in terms of their relative distance. Clusters with low distances between their individual points are more tight and concentrated in an area of 2-dimensional space, while disperse clusters with high distances among their instances are more likely to be split in an expanded clustering process. Results through the evaluation phase confirm the aforementioned statement.

The efficiency of the methodology is evaluated by using the silhouette metric, which gives quantitative information for how strong or loose the connections between the neurons within a cluster are. We showed that, by performing the optimum clustering as the DB-index denotes, clusters derived are more stable, having at the same time higher mean silhouette values than the ones of less extensive clustering processes.

On the other hand, the qualitative efficiency of the methodology can be evaluated subjectively by: (a) the clustering results derived, (b) the accuracy of the actual data selected per cluster, and (c) the correlation between the content and the semantic annotation. Based on our extended investigation, all three factors are achieved in the current analysis.

The results of the proposed analysis are of interest for several reasons. Clustering results can be used to perform efficient classification of newly retrieved tweets into different description groups of content. They can also be used as an awareness or decision factor. The semantic interpretation of clusters can play a key role in different applications, such as recommendation systems and event detection services. By exploiting the massive content from social media and mining their text streams, practitioners can monitor events like dispersion of allergies or bad quality conditions, even in areas where there is no other means of monitoring air quality.

6. Conclusions

We focused on the task of social media analysis and we investigated the use of Twitter as a source of user-generated information concerning the atmospheric environment domain and its effect to humans' quality of life. We proposed a straightforward methodology for encoding and analyzing tweets based on the occurrence or non-occurrence of selected words in text. The ultimate goal that we achieved was to cluster text into groups with similar content, giving also a semantic prospect to the interpretation of the results. We demonstrated the use of SOM and k-means, along with the DB-index to perform the clustering process efficiently. We were able to extract well-defined clusters and derive topics of discussion from the massive collection of posts.

Relationships among concepts that describe the content can further form hierarchical structures. Both interpretation of visualizations and evaluation of clustering results can underscore the efficiency and the consistency of the methodology. By structuring the knowledge derived in an easily interpretable way, we manage to provide a rich content for individuals and practitioners for quality of life issues.

In future work, we aim to overcome the limitation of language used in text collection and add spatiotemporal dimensions to the related extracted concepts. Profiling concepts over a longer time interval, i.e. within a calendar year, or analyzing the content near real-time will create a reference of event detection or supplementary serve as an observation/alarm service (in our example of air quality issues), even in areas where monitoring data for air quality are not available. The hierarchy of concepts as well as additional semantic description can be used to form a data-driven ontology for the air quality domain. Finally, we are interested to research the applicability of the methodology in other domains of interest.

Acknowledgements

We would like to thank the Centre for International Mobility CIMO (http://www.cimo.fi/) for the CIMO Fellowship funding which enabled the work described in this article.

References

Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I. and Jaimes, A. (2013). Sensing trending topics in Twitter. IEEE Transactions on Multimedia. 15 (6), 1268-1282. doi: 10.1109/TMM.2013.2265080

Berry, M.W. (ed.) (2004). Survey of Text Mining: Clustering, Classification and Retrieval. Springer Science + Business Media, Inc.

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S. and Srivastava, M.B. (2006). Participatory Sensing, Proc. of the Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications, Colorando, USA, 117-134.

Cairns, C. (2013). Air Pollution, Social Media, and Responsive Authoritarianism in China, UCLA COMPASS Conference in Comparative Politics.

Costa, J., Silva, C., Antunes, M. and Ribeiro, B. (2013). Defining Semantic Meta-hashtags for Twitter Classification. LNCS Adaptive and Natural Computing Algorithms. 7824, 226-235. doi: 10.1007/978-3-642-37213-1_24.

Crooks, A., Croitoru, A., Stefanidis, A. and Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. Transactions in GIS. 17 (1). 124-147. doi: 10.1111/j.1467-9671.2012.01359.x.

Davies, D. and Bouldin, D.W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2), 224-227. doi: 10.1109/TPAMI.1979.4766909.

Dredze, M. and Paul M.J. (2014). Natural Language Processing for Health and Social Media, 64-67.

Hall, D.L., Llinas, J., McNeese, M. and Mullen, T. (2008). A Framework for Dynamic Hard/Soft Fusion, Proc. of the Eleventh International Conference on Information Fusion, Cologne, Germany.

Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics. 28(1), 100-108. doi: 10.2307/2346830.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning. 42, 177-196. doi: 10.1023/A:1007617005950.

Honkela, T. (1997). Self-Organizing Maps of words for natural language processing applications. Ph.D. Dissertation, Neural Networks Research Center, Helsinki University of Technology, Espoo, Finland.

Kaplan, M. A. and Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. Business Horizons. 54(2), 105-113. doi: 10.1016/j.bushor.2010.09.004.

Karatzas, K. (2009). Informing the public about atmospheric quality: air pollution and pollen. Allergo Journal. 18 (3/09), 212-217.

Kaufmann, M. and Kalita, J. (2010). Syntactic normalization of twitter messages, Proc. of the International Conference on Natural Language Processing, Kharagpur, India.

Kohonen, T. (1990). The self-organizing map, Proc. of the IEEE, 78 (9), 1464-1480. doi: 10.1109/5.58325.

Kotovirta, V., Toivanen, T., Tergujeff, R. and Huttunen, M. (2012). Participatory Sensing in Environmental Monitoring – Experiences, Proc. of the Sixth Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Palermo, 155-162. doi: 10.1109/IMIS.2012.70.

Kovács, F., Legány, C. and Babos A. (2006). Cluster Validity Measurement Techniques, Proc. of the Fifth WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, 388-393.

Lagus, K., Kaski, S. and Kohonen, T. (2004). Mining massive documents collections by the WEBSOM method. Journal of Information Sciences – Special issue: Soft computing data mining. 163 (1-3), 135-156. doi: 10.1016/j.ins.2003.03.017.

Li, R., Lei, H. K., Khadiwala, R. and Chen-Chuan Chang, K. (2012). TEDAS: a Twitter-based Event detection and Analysis System, Proc. of the IEEE 28th International Conference on Data Engineering (ICDE), Arlington, Virginia USA, 1273-1276. doi: 10.1109/ICDE.2012.125.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J. and Mitra, P. (2011). Geo-Twitter Analytics: Applications in Crisis Management, Proc. of the 25th International Cartographic Conference, Paris, France, 1-8.

Maynard, D. and Funk, A. (2011). Automatic detection of political opinions in tweets, Proc. of the Eighth International Conference on The Semantic Web: ESWC 2011, Heraklion, Greece, LNCS 7117, 88-99.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. Expert Systems and Applications. 40 (10), 4241-4251. doi: 10.1016/j.eswa.2013.01.019.

Mun, M., Reddy, S., Shilton, K., Yau, N., Burke, J., Estrin, D., Hansen, M., Howard, E., West, R. and Boda, P. (2009). PEIR, the personal environmental impact report, as a platform for participatory sensing systems research, Proc. of the Seventh International Conference on Mobile systems, applications and services, Kraków, Poland, 55-68. doi: 10.1145/1555816.1555823.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J. and Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence. 17 (2), 159-167. doi: 10.1016/j.engappai.2004.02.002.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2 (1-2), 1-135. doi: 10.1561/1500000011.

Popescu, A. M. and Pennacchiotti, M. (2010). Detecting controversial events from Twitter, Proc. of the 19th ACM International Conference on Information and Knowledge Management, 1873-1876. doi: 10.1145/1871437.1871751.

Ritter, H. and Kohonen, T. (1989). Self-Organizing Semantic Maps. Bioclinical Cybernetics. 61 (4), 241-254. doi: 10.1007/BF00203171.

Robinson, E. (2010). Integration of multi-sensory earth observations for characterization of air quality events. Ph.D. Dissertation, School of Engineering and Applied Science, Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, USA.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20, 53-65. doi: 10.1016/0377-0427(87)90125-7.

Sadilek, A., Kautz, H. and Silenzio, V. (2012). Predicting Disease Transmission from Geo-Tagged Micro-Blog Data, Proc. of the 26th AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada, AAAI Press 2012.

Salton, G., Wong, A. and Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM. 18(11), 613-620. doi: 10.1145/361219.361220.

Sammon, J.W. (1969). A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers. C-18 (5), 401-409. doi: 10.1109/T-C.1969.222678.

Skön, J.P., Kauhanen, O. and Kolehmainen M. (2011). Energy Consumption and Air Quality Monitoring System, Proc. of the Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Adelaide, Australia, 163-167. doi: 10.1109/ISSNIP.2011.6146606.

Sullivan, D. (2001). Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales. John Wiley & Sons.

Tan, A.H. (1999). Text Mining: The state of the art and the challenges, Proc. of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, Beijing, China, 65-70.

TechCrunch (2012). Twitter May have 500M+ Users but only 170M Are Active, 75% on Twitter's Own Clients. http://techcrunch.com/2012/07/31/twitter-may-have-500m-users-but-only-170m-are-active-75-on-twitters-own-clients/ (accessed Apr 09, 2013).

Ultsch, A. and Siemon, H.P. (1990). Kohonen's Self Organizing Maps for Exploratory Data Analysis, Proc. of the International Neural Networks Conference (INNC), Paris, France, 305-308.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A. and Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Science of the Total Environment. 409 (7), 1266-1276. doi: 10.1016/j.scitotenv.2010.12.039.

Voukantsis, D., Niska, H., Karatzas, K., Riga, M., Damialis, A. and Vokou, D. (2010). Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. Atmospheric Environment. 44 (39), 5101-5111. doi: 10.1016/j.atmosenv.2010.09.006.

Yang, H.C. and Lee, C.H. (2010). A novel self-organizing map algorithm for text mining, Proc. of International Conference on System Science and Engineering, 417-420. doi: 10.1109/ICSSE.2010.5551734.

Zhang, X., Fuehres, H. and Gloor, A. P. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not bad as I fear". Procedia - Social and Behavioral Sciences. 26, 55-62. doi: 10.1016/j.sbspro.2011.10.562.