Information Fusion 40 (2018) 1–17

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A Social-aware online short-text feature selection technique for social media

Antonela Tommasel*, Daniela Godoy

ISISTAN, UNICEN-CONICET. Campus Universitario, Tandil (B7001BBO), Argentina

ARTICLE INFO

Article history:
Received 30 December 2016
Revised 4 May 2017
Accepted 16 May 2017
Available online 17 May 2017

Keywords:
Online feature selection
Classification
Micro-blogging communities

ABSTRACT

Large-scale text categorisation in social environments, characterised by the high dimensionality of feature spaces, is one of the most relevant problems in machine learning and data mining nowadays. Short-texts, which are posted at unprecedented rates, accentuate both the importance of learning tasks and the challenges posed by such large feature space. A collection of social media short-texts does not only provide textual information but also topological information given by the relationships between posts and their authors. The linked nature of social data causes new complementary data dimensions to be added to the feature space, which, at the same time, becomes sparser. Additionally, in the context of social media, posts usually arrive simultaneously in streams, which hinders the deployment of efficient traditional feature selection techniques that assume a feature space fully known in advance. Hence, efficient and scalable online feature selection becomes an important requirement in numerous large-scale social applications. This work presents an online feature selection technique for high-dimensional data based on the integration of two information sources, social and content-based, for the real-time classification of short-text streams coming from social media. It focuses on discovering implicit relations amongst new posts, already known ones and their corresponding authors to identify groups of socially related posts. Then, each discovered group is represented by a set of non-redundant and relevant textual features. Finally, such features are used to train different learning models for classifying newly arriving posts. Extensive experiments conducted on real-world short-texts demonstrate that the proposed approach helps to improve classification results when compared to state-of-the-art and traditional online feature selection techniques.

1. Introduction

Large-scale text categorisation in social environments is one of the most relevant problems in machine learning and data mining nowadays. With social media data growing at unprecedented rates, this problem becomes a matter of paramount importance for numerous real-world applications. For example, tweets could be classified aiming at discovering breaking news or events (such as natural disasters) helping to understand the impact of incidents, or assisting in emergency management and crisis coordination. Additionally, trending topics or social trends could be discovered by analysing clusters of related tweets.

The pervasive use of social media offers research opportunities for analysing user behaviour and how they interact with their friends. Unlike social connections formed by people in the physical world, social media users are free to connect with a wider number of people for a variety of reasons. The low cost of link formation might lead to networks with heterogeneous relationship origin or strength. For example, in *Twitter*, a user might follow others because they publish interesting information, they have the same interests, or even because they share some common friends, amongst other possible explanations. In addition to social information indicating friendship or simply user interaction, there are other information sources that might implicitly define connections between users in social media. For example, whether two users use the same terms, hashtags, or post about the same topic. Moreover, the social media experience of users is no longer limited to a unique site, as users use social media for different purposes [49]. As a result, each social media site provides heterogeneous and complementary information sources for describing a particular user, their interests and social relations.

A task that can greatly benefit from the integration of multiple information sources is text categorisation. Such task is characterised by the high dimensionality of their feature space where

* Corresponding author.
E-mail addresses: antonela.tommasel@isistan.unicen.edu.ar (A. Tommasel), daniela.godoy@isistan.unicen.edu.ar (D. Godoy).

http://dx.doi.org/10.1016/j.inffus.2017.05.003
1566-2535/© 2017 Elsevier B.V. All rights reserved.

most terms have low frequencies. This situation is commonly known as the curse of dimensionality, which refers to the increasing computational complexity of learning problems as the volume of data grows exponentially regarding the underlying space dimension. This problem worsens when considering short-texts, such as tweets, *Facebook* posts or even blogs' social annotations. Nonetheless, a collection of short-texts in social media does not only provide textual information but also topological information due to the relationships between posts and users. In turn, the linked nature of social media data causes new dimensions (such as friendship relations between users) to be added to the feature space [34]. The increasing amount of data does not only affect the computational complexity of algorithms, but also poses new challenges regarding how to represent and process new data, and how to effectively leverage on such data for improving the performance of text learning tasks [7].

Feature selection (FS) [3] is one of the most known and commonly used techniques to diminish the impact of the high-dimensional feature space by removing redundant and irrelevant features. The standard FS setting assumes the existence of instances, and therefore a feature space, fully known in advance. Thus, FS consists in finding a small subset of the most relevant features according to certain evaluation criterion. This setting is known as batch FS. However, in real-world applications, and particularly social media ones, such assumptions might not hold as either training examples could arrive sequentially, or it could be difficult to collect the full training set [44]. For example, in the context of social media data, posts usually arrive simultaneously in streams, hindering the deployment of efficient and scalable batch FS techniques. Thus, traditional batch FS techniques are not suited for emerging big data applications. In these situations, online feature selection (OFS) in which instances and their corresponding features arrive in a continuous stream, needs to be performed. This process involves choosing a subset of features and the corresponding learning model at different time frames. Thereby, OFS is particularly important in real-world systems in which traditional batch FS techniques cannot be applied.

Motivation

Although FS techniques have received considerable attention during the last decades, most studies focus on developing batch techniques instead of facing the challenging problem of OFS. The majority of FS techniques are designed for data containing uniform features, which are typically assumed to be independent and identically distributed. However, this assumption might not hold in social media since measuring the relevance of features in isolation possibly ignores dependencies amongst them given by the social context. Interestingly, most algorithms only focus on content-based information sources, even though social media content might be topically diverse and noisy, which hinders the effective identification of relevant and non-redundant features. It is worth noting, linked data has become ubiquitous in social networks, as in *Twitter* (in which not only tweets can be linked, but also their authors might be socially related) or *Facebook* (in which users share friendship relationships), providing additional information sources such as correlations between instances. For example, posts from the same user or two linked users are more likely to have similar topics. As the different information sources provide complementary views of data, when assessing them independently, algorithms may fail to account for important data characteristics. Instead, FS techniques should be capable of combining multiples information sources. In this context, the availability of link information enables advanced research in FS techniques, which needs to address two challenges: how to exploit relations amongst data instances, and how to leverage those relations for FS.

Efficient and scalable OFS is an important requirement for numerous large-scale social applications. Despite presenting significant advantages in efficiency and scalability, existing OFS techniques do not fully leverage on the multiple information sources available. Instead, they mainly focus on textual information. Potentially, the performance of such approaches could be improved by including additional information sources in social media data. Furthermore, most of the approaches that claim to be applicable in OFS, might fail when used in the context of social media data, due to the need of knowing either all data instances or features in advance, making them unsuitable for data streams. In consequence, novel approaches for efficiently selecting and updating the selected subset of features need to be developed.

Considering that different information sources in social media can provide multiple and possibly complementary views about data, this paper aims at addressing the OFS task for high-dimensional short-text data arriving in a stream. The hypothesis behind this work is that more accurate OFS techniques could be developed by effectively integrating multiple information sources. The main goal of this work is to define and evaluate a new intelligent technique for short-text mining to enhance the process of knowledge discovery in social media. To that end, an OFS technique for leveraging on social information to complement commonly used content-based information is presented. The technique is based on the integration of social network structures into the process of OFS [42].

Unlike other works found in the literature, the focus of the presented technique is to analyse different types of social relationships between posts and their authors. Particularly, this work aims at performing real-time classification of continuously generated short-texts in social networks by exploring the combination of multiple relations amongst data instances in the social environment and how to leverage such multiple relations for enhancing FS techniques. The goal is to discover implicit relations between new posts and already known ones, based on a network comprising the individual posts and the users who have written them. Then, the content in the discovered groups of socially related posts is analysed to select a set of non-redundant and relevant features to describe each group of related posts. Finally, such features are used for training different learning models to categorise newly arriving posts.

Contributions

The expected contributions of this work are described as follows. First, it tackles the problem of how to exploit social relations amongst data instances by studying the linked nature of social media data. Second, it proposes a technique for leveraging on those social relations. Third, it combines social information with the content of posts for effectively and efficiently performing FS. Fourth, the technique is scalable, and thus appropriate for real-time environments in which neither features nor instances are known in advance. Furthermore, it allows the process of data instances as they are generated in a reasonable amount of time. Finally, the presented technique could help in the development of new and more effective models for personalising and recommending content in social environments.

The rest of this paper is organised as follows. Section 2 discusses related research on OFS. Section 3 presents the proposed OFS technique combining two heterogeneous and complementary information sources: social and content-based information. Section 4 describes the experimental settings and results obtained for two social media datasets. Finally, Section 5 summarises the conclusions drawn from this study, and presents future lines of work.

2. Related work

Most OFS approaches, such as [27,45,46], assume that features arrive sequentially and individually, whilst all training instances are known in advance. Their goal is to build an appropriate learning model at each time frame given the full set of instances and the features known up to that moment. One simple OFS approach is to build the set of all discovered features that arrived in the continuous stream and then apply any traditional FS technique [27]. However, given that the size of the feature space can continuously increase, this approach might present scalability issues. For example, in online settings, FS approaches based on statistics would re-compute them each time a new feature is discovered, which might be time-consuming, even if new features are individually analysed. The problem worsens if the quality of features is assessed by learning algorithms. In an online setting, at each time-step, the feature set changes. In this context, evaluating the performance of the model considering each possible feature set would be inefficient and computationally complex. Particularly, in a real-world online environment, there is usually a limited amount of computational time in-between the arrival of new instances and features, thus the update time of the designed technique should not unlimitedly increment as more features or instances arrive. Ultimately, an OFS technique must allow performing efficient incremental updates.

Grafting and α-investing are traditionally used techniques for stream FS. Grafting (a stage-wise gradient descent feature testing) was proposed by Perkins and Theiler [27] for binary classification problems. The grafting technique involves ℓ_1-regularisation, un-regularised parameters and the definition of weight vectors for each potential feature. For each new feature and its associated weight vector, a gradient test is performed. If no weight passes the test, the feature is discarded. Conversely, if at least one weight passes the test, the feature and its highest weight is added to the model. After model updates, a re-optimisation step is applied and the tests are repeated. α-investing [50] is based on adaptively modifying the threshold α for adding new features. The threshold defines the probability of including spurious features, and it is adjusted by defining the acceptable rate of irrelevant features to select. Every time new features are not added to the model, α is decreased to avoid selecting more than a pre-defined proportion of spurious features. Similarly, when new features are added to the model, α is increased. Note that, the addition of new features does not trigger an analysis of the already selected features. Hence, it cannot be guaranteed that features are not redundant. Although this technique was reported to outperform Bag-of-Words (BOW) representations combined with Support Vector Machines (SVM), neural networks and decision trees, it needs prior knowledge of the feature space to heuristically control the selection of features [46]. As such information might be difficult to extract from the feature stream, the technique might not be applicable in truly online environments. Hence, more efforts would be needed to effectively analyse real-world streams in which the feature structure is unknown.

Redundancy and relevance of stream features was analysed by Wu et al. [46] in the context of a supervised two-step approach. The first step analyses the relevance of the new feature. The second step is only performed when relevant features are found, and analyses the redundancy of the newly selected features regarding the already selected ones. The algorithm iterates through these two steps until a stopping criterion is satisfied. Both feature relevance and redundancy are analysed in terms of probabilistic conditional independence, which could be unreliable for small datasets. Aiming at reducing the computational complexity of the approach, the authors presented a variation in which the redundancy analysis of the set of selected features is only performed once the process of generating features is stopped instead of in every iteration. Although the described approaches were reported to achieve promising results in traditional classification tasks (i.e. long-text binary classification), their adaptation and performance for short-text classification has not yet been evaluated. Moreover, the presented approaches assume that the instance or feature set is fully known in advance, hindering their applicability in the context of social media data in which streams of data are continuously arriving, comprising already known or unknown features. Finally, as features are assumed to be unique and to individually arrive, techniques do not assess feature repetition. In social media settings, features arrive in groups, which might include both known and new unknown features. In this context, already known features might be re-evaluated, affecting the computational time of techniques. Contrasting with such approaches, this work considers a sequential arrival of instances, in which new features are analysed as they arrive. This approach aims at mimicking the social environment in which new content that needs to be categorised is permanently being generated.

The previously presented approaches dynamically evaluate individual features as they arrive. Nonetheless, they neglect the relationships between features, which might be of particular importance in certain environments. For example, in the context of social media, it might be useful to analyse both the content, and the social relations or friendship network of authors, as a unit. In this regard, Wang et al. [44,45] specifically designed approaches for analysing groups of features.

A variation of the two-step technique in [46], in which features are assumed to arrive in groups, was presented by Wang et al. [45]. This variation is also divided in two steps. The first step applies spectral analysis to select the most discriminative features of each new group by computing several matrix arithmetic operations, resulting computationally complex. The second step applies a linear regression model to select a global optimal subset from the newly selected discriminative features and the previously selected ones. As traditional spectral FS techniques rely on global information, which is not available for OFS, two criteria for selecting new features were defined. The first one selects features maximising the inter-class distances by analysing the differential discriminative power of the new feature regarding that of the already selected ones. The second criterion analyses the discriminative power of the individual features by performing a t-test. Features were selected if they satisfied any criterion. Both steps are iteratively performed until a stopping criterion was met: a pre-defined number of features are selected, there are no more features in the stream, or the predictive accuracy of the selected set of features is higher than a threshold. In real-time classification of continuously generated social short-texts, each newly arrived post could be regarded as a new group of features. However, in such environments features are repeated across posts. As the technique does not provide any mechanism for dealing with repeated features, features in social posts could be analysed several times. In addition, already selected features could be re-evaluated, negatively affecting the performance of the technique.

Replicating the setting of real-world applications, Wang et al. [44] considered a sequential arrival of instances. The authors used sparse online learning, and introduced a limit to the number of features that the linear learner is allowed to access for each instance. It is based on a greedy technique that randomly selects a subset of features only keeping those with non-zero values in the resulting linear classifier. The authors claimed that their approach outperformed state-of-the-art approaches such as the Minimum Redundancy Maximum Relevance Feature Selection [13] and the Forward Backward Splitting algorithm [14] for most of the evaluated datasets and tasks. However, as the approach was evaluated for long texts in a binary-class setting, its findings might not be generalised to social environments, which often involve multi-class classification of short-texts.

Most OFS approaches for short-text categorisation rely only on computing new features, such as aggregated statistics or N-grams, instead of selecting a subset of the original features. For example, Li et al. [20] proposed to detect real-world events by only extracting non-overlapping segments of one or more consecutive words contained in tweets, without further processing them. Tweet segments appearing in a large number of tweets were supposed to represent meaningful concepts or named entities, thus convening more specific information than individual uni-grams. Becker et al. [6] aimed at differentiating between real-world events and non-event messages in *Twitter* by defining features that considered statistics of the expected volume of messages, user interactions, topical coherence of tweet clusters and the usage of tags.

Zubiaga et al. [51] proposed an approach for real-time tweet classification into four categories: news, ongoing events, memes and commemoratives by defining a small set of language-independent features. Such features comprised aggregated statistics of lexical elements commonly found in tweets (hashtags, URLs, exclamation and questions marks), retweets and replies. Additionally, the authors included as features the Shannon diversity index of users, hashtag, language and vocabulary. This approach has four characteristics that make it suitable for real-time tweet classification: the required small feature set can be straightforwardly computed, it does not make use external data sources, it can improve the predictive power of content, it has a linear computational cost to the number of tweets to be analysed, and the number of features remains unchanged regardless the number of instances. However, as features were specifically designed for the four described classes, they might not be applicable to other domains.

Unlike the previously presented works, that only focus on one type of features (e.g. textual or statistical features), there are a few others that combine social information with textual features [12,21]. Textual and structural information provided by the relationship between tweets and users were combined by Cot [12] for classifying *Twitter* users according to their political opinion. Tweets were pre-processed and a forest of randomised trees was applied to select the most relevant content features. Structural information was used to build a bipartite graph in which nodes represented users and edges represented the Dice similarity between a user and each of his/her followees. Only edges with a similarity higher than a threshold were included in the graph. Then, a fuzzy community detection technique was applied. Once communities were found, the authors created a feature vector for each user indicating his/her affinity towards the communities, as the sum of the weights of the followees that belong to each community. Finally, the authors presented three alternatives for combining the content and structural features. First, concatenating both features into a single feature model. Second, feeding classifiers the complete set of features, and then combining their classification results. Third, using a meta-classifier for combining classifications based on each individual feature type. Results showed that combining textual and content information allowed improving classification results of the independent feature sets. The authors concluded that extracting knowledge and generating good feature models was harder for structural features than for content-based ones. Moreover, they stated that combining both types of features was not trivial, as the direct combination did not achieve the best results. Unlike Cot [12]'s technique, this work assumes a closer relation between structural and content-based features, and uses the information provided by the structural features to select the most relevant content features.

Finally, closely related to this work is the study carried out by Li et al. [21] that proposed an unsupervised FS technique for social media. First, the social latent factors for each instance are obtained based on the mixed membership stochastic blockmodel [2]. Then, the importance of each feature is measured as its ability to distinguish multiple social latent factors. The decision of whether to accept a new feature is defined as an optimisation problem involving the computation of several arithmetic operations between high-dimensional matrices. Each time a new feature arrives, a gradient test is performed to decide whether the feature is accepted. If the feature is accepted, the model is re-optimised, and there is also the possibility of removing already selected features. However, this approach presents different shortcomings that might affect its applicability. First, it is designed to be applied on specific streaming environments in which all data instances are required to be known in advance. Second, it assumes that link information is relatively stable. Thus, social links between instances are never updated to reflect changes in the social network. Third, its evaluation was performed in a batch setting. Fourth, solving the optimisation problem requires the computation of arithmetic operations between high-dimensional matrices.

In summary, all the presented techniques suffer from different shortcomings that might affect their applicability on real OFS settings [41]. First, they are intended to be applied on particular streaming environments in which features arrive one at the time, implying that the full set of instances is known in advance. Second, evaluations are mostly performed once all features are processed. Hence, performance is not assessed in real online environments. Third, techniques might not be scalable. Techniques that involve solving optimisation problems requiring the computation of arithmetic operations between matrices might not be applicable in online settings due to their high computational complexity and substantial memory consumption. In the same regard, techniques requiring to load the full dataset on memory might not be applicable on high-dimensional domains. Fourth, the majority of FS techniques are designed to analyse individual features assuming their independence. However, social media data does not follow that assumption, as data instances not only comprise groups of features but also are inherently linked through social relationships, which can provide extra information beyond the feature-value. As regards the techniques specifically proposed for the short-text domain, two of them are of low computational complexity. Nonetheless, features were specifically designed for the task to be performed, which might hinder their applicability on other domains. Moreover, those feature sets might not adequately adapt to the dynamically changing environment of social media data.

To conclude, the main shortcomings of the existent OFS techniques are related to whether considering features individually or grouped, the effect of neglecting the linked nature of social media data, and scalability. Such shortcomings lead to an imperious need of developing novel FS techniques to cope not only with an enormous amount of data that is continuously generated in social media networks, but also with the performance and computational complexity requirements.

3. Social-aware OFS technique

Efficient and scalable OFS becomes an important requirement for numerous large-scale social applications. As discussed in the previous Section, despite presenting significant advantages in efficiency and scalability, the existing OFS techniques do not fully leverage on the multiple information sources available. Instead, they mainly focus on textual information. The hypothesis behind this work is that more accurate OFS techniques could be developed by integrating multiple information sources. Potentially, the performance of such approaches could be improved if fully social media data. Furthermore, most of the techniques that claim to be applicable in OFS, might fail when used in the context of social media data, either due to the need of knowing all data instances or fea-

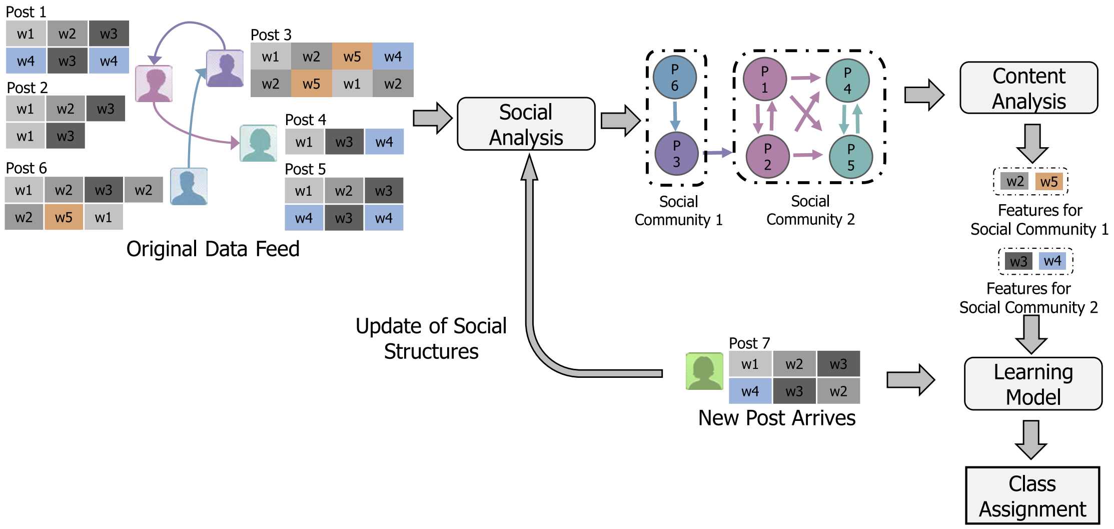


Fig. 1. Overview of the presented feature selection technique.

tures in advance, which makes them unsuitable for data streams. Moreover, most techniques have a high computational complexity and lack of updates of the selected set of features. In consequence, novel approaches for efficiently selecting and updating a subset of features need to be developed.

This work presents an OFS technique for high-dimensional data based on both social and content-based information. The technique aims at the real-time classification of social posts arriving in a continuous stream, i.e. neither the features nor the data instances are fully known in advance. Fig. 1 depicts the general methodology of the presented technique. The focus of the technique (Section 3.1) is to analyse the social relationships between posts and their authors to detect groups of socially related posts. In this regard, a social graph is created to define implicit relations between new posts and already known ones based on a network comprising the individual posts and the users who have written them. Then, the content in the discovered groups of socially related posts is analysed (Section 3.2) to select the set of non-redundant and relevant features describing each group of related posts. Finally, the selected features are used for training different learning models for classifying newly arriving posts (Section 3.3).

3.1. Social analysis

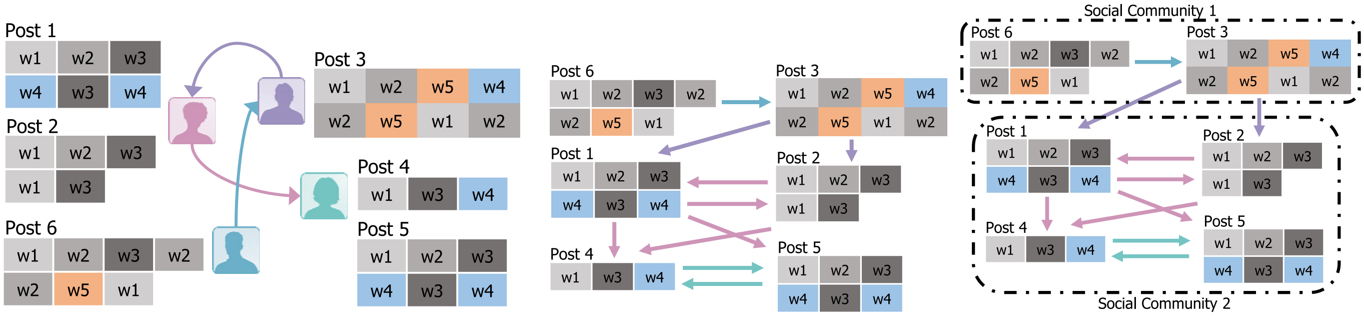
Social networks can be defined as a set of socially relevant nodes connected by one or more relations. Nodes can represent not only real people, but also diverse entities such as Web pages, journal articles, countries, neighbourhoods, or positions, amongst others [25]. The nature and nomenclature of connections amongst users might differ from site to site. For example, links could represent followee relations as in *Twitter*, or friendship relations as in *Facebook*. This work focuses on social networks composed by individual posts and the users who have written them. Although the technological features of the different social networking sites are similar, the cultures that emerge around them are diverse [9]. For instance, most sites encourage the maintenance of pre-existing social networks, whilst others help strangers to create new connections based on shared interests, which could result in connections between individuals that would not otherwise be made.

Following the same data feed used in Fig. 1 to depict the general methodology of the presented technique, Fig. 2 focuses on the process that social data undergoes during the social analysis step, which covers the transformation of the original data feed into a graph and the posterior community discovery. Considering the social and content-based relationships described, the original data feed shown in Fig. 2a is transformed into a social graph. In the derived graph (Fig. 2b), each node represents a social post, and the edges between them, the different types of relations between the involved posts. In the example, an edge between two posts exists if a relationship existed between their authors, i.e. the example graph only depicts the topological relationship of users. Nonetheless, additional types of relationships could be defined, which are presented in Section 3.1.1. Once the graph is derived, it can be analysed to discover communities of related posts (Fig. 2c), as described in Section 3.1.2.

3.1.1. Modelling the social graph

In the context of social media data, both the graph topological structure (i.e. social relations between users) and the vertex properties (i.e. posts characteristics) are important as they offer complementary views of data. Several studies have analysed the different types of relations that can be defined amongst users and their corresponding posts [34] based on theories such as homophily [26]. As a result, besides the relations between posts that could be derived from the social relations between their authors, there could be additional information sources available for each post [40]. The resemblance of content or posts categories (when available) could be the source of new relations. Moreover, each microblogging site has specific characteristics and metadata that could be exploited for establishing meaningful relations between posts. For example, *Twitter, Instagram* and *Facebook* promote the usage of hashtags, which represent a type of label or metadata to aid in the search of a specific theme or content. Additionally, *Facebook* allows to search for posts sharing specific activities, for example “’listening Aerosmith” or “’reading Oscar Wilde”. Fig. 3 exemplifies different types of relations that could be established between two nodes. Interestingly, social information and content-based relations offer complementary views of the social posts. Hence, each individual data view might not be sufficient for accurately describe the relations between posts. For example, content-based information could be irrelevant or redundant, whilst social information might be sparse or noisy. As a result, in order to accurately describe posts it is important to combine both types of relations.

The defined content-based relations could be used either to reinforce the social relations already found amongst posts or to establish new relations between posts that are not socially related. Note that, in both cases the content information of nodes is transferred to edges to characterise the specific relations between the



(a) Original data feed

(b) Modelling the data feed as a social graph

(c) Detecting communities over the social graph

Fig. 2. Social analysis step of the social-aware OFS technique.

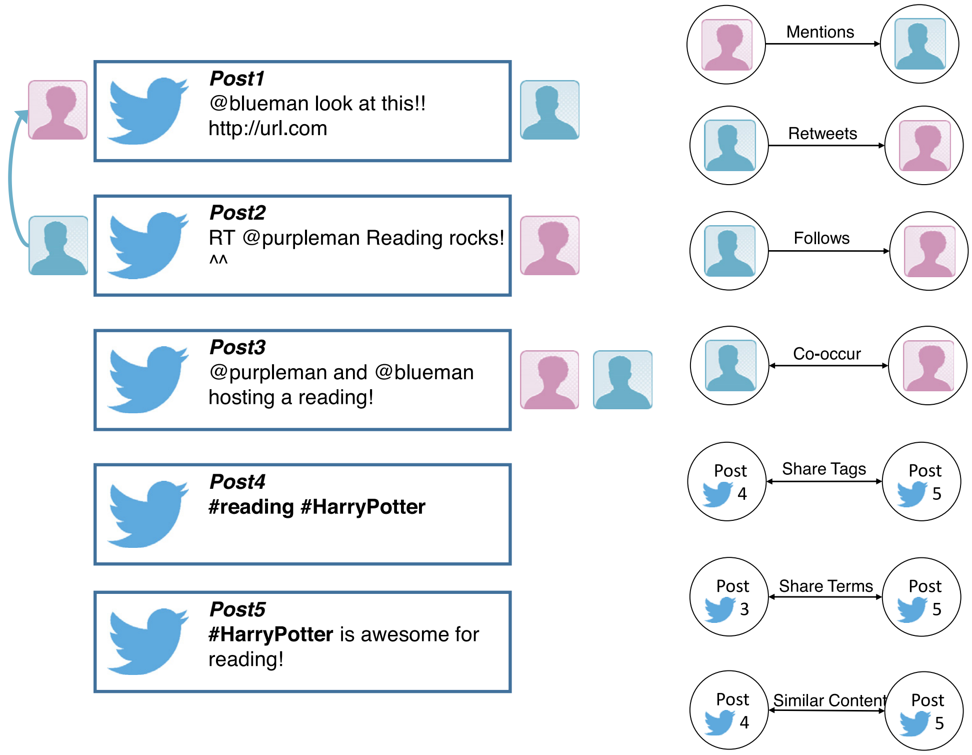


Fig. 3. Examples of possible content-based relationships between social posts.

linked nodes. In the former case (referred as *Weighted* derivation), the graph only includes edges representing the social relations between nodes, whose relevance is given by the content features. For example, the content similarity between two posts could be used for determining the strength of the existing social relation between them. As a result, in this case, the quality of the social relations depends on the adequate assessment of the content-based features, which should allow to fully exploit the social media data information. In the latter case (referred as *Independent* derivation), social and content-based relations are assumed to be unrelated, i.e. edges in the graph do not only represent social links, but also content ones.

For the purpose of this work, a social relation between two posts (named *Social*) was established in two cases. First, between posts written by the same author. Second, between posts whose authors are friends in the social network. Note that these social relations, depending on the network under analysis might not be symmetrical. In such cases, as most community detection algorithms are based on the analysis of undirected graphs, a symmetrisation technique must be applied to accurately capture the semantics of the asymmetric relationships conveyed by the edges of a directed network. The adopted symmetrisation technique defined the new adjacency matrix of the graph as $U = A + A^T$, where A represents the adjacency matrix of the original graph. This strategy is similar to ignoring edge directionality, except that in the case a pair of nodes is connected with edges in both directions, the weight of the edge in the symmetrised graph will correspond to the sum of the weight of both edges.

In addition to the existing *Social* relation, several content-based relationships between nodes can be computed as defined [40]. Note that, all content-based relations are symmetric, i.e. they do not have directionality. Moreover, each relation might have an individual scale-factor representing the importance of such relation in the final graph. Considering the information available on social networking sites, the content-based relations were defined as follows:

- *Similar Content.* Measures the content resemblance of two nodes. A minimum similarity threshold can be imposed to avoid creating a complete dense graph. Thus, only edges with similarity above a certain threshold would be added to the graph. Diverse text similarity metrics can be adopted to define the nature and strength of the similarity. For example, similarity could be expressed by simply computing the percentage of shared terms between the two nodes or by computing the Cosine Similarity metric.

- *Shared Class*. Two nodes are said to be related if they belong to the same class. All edges have a weight of 1. In those cases in which categories are organised in hierarchies or taxonomies (as in the *Open Directory Project*[1]), the weight of edges could be computed as the distance between both categories.

Once all relations between nodes are defined, they are integrated into a unique social graph. Such unique graph collapses multiple (and possibly heterogeneous) relations between two nodes into a unique edge. The weight of such edge would be equal to the sum of all edges' weights. Then, the social graph is analysed in search of communities.

3.1.2. Discovering communities from the social graph

Communities refer to potentially overlapping groups of nodes that have dense connections between the nodes within the community, but sparse connections with nodes belonging to other communities. In this context, the goal of community detection techniques (also known as graph clustering techniques) is to divide the nodes into communities (or clusters), such that the nodes of a particular community are similar or connected in some predefined sense [32]. For example, in some cases it might be desirable to obtain communities of similar order and/or density. Interestingly, not all graphs present a structure of natural communities. In the case of a uniform graph in which the edges are evenly distributed over the set of vertices, the resulting clustering will be rather arbitrary. It is important to emphasise that due to the immense scale and evolving nature of social media data, it is infeasible to estimate the number of actual communities.

Community detection has proven to be valuable in a diverse set of domains [15] such as biology, social sciences and bibliometrics, amongst others. In the context of recommendation, identifying communities of costumers with similar interests in the purchase network of customers and products of online retailers could improve the quality of product recommendations, thus better guiding customers in their shopping activities. As regards biology, Lusseau [22] analysed a network of bottlenose dolphins living in New Zealand. Due to the natural separation of dolphins in groups, with few vertices joining the different communities, such network is often used to test the performance of algorithms. Additionally, community detection can be applied to protein interaction networks to group proteins having similar functions and interactions, which are fundamental for the inter-cell processes, and hence improve the protein function prediction [19].

Although social graphs can comprise direct links, most techniques found in literature disregard such directionality, failing to accurately capture the semantics of the asymmetric relationships implied by the edges of a directed network [24]. Regardless of the adopted detection technique, the social structure of posts should be updated when new posts are discovered. In other words, as data arrives in a stream, the social graph has to be periodically updated for coping with the continuous evolution of topics and the newly discovered posts, and thus be updated with the new data.

In this work, one of the most efficient and traditionally-used community detection algorithms was used, the Louvain algorithm [8]. This algorithm implements a greedy method based on the local optimisation of modularity, and the aggregation of nodes of the same community to build a new network whose nodes are such communities. Although the output and execution time of the algorithm depend on the order of the analysed nodes, the authors stated that it did not have a significant effect on the final network modularity. After community structures have been discovered, the content analysis step takes place.

3.2. Content analysis

Once communities are discovered, they are individually analysed to find the non-redundant and relevant features, i.e. the most important features, describing such particular group of posts. According to [18], features can be classified into three disjoint categories, i.e. strongly relevant, weakly relevant, and irrelevant features. Strongly relevant features are always necessary for defining an optimal subset, i.e. they cannot be removed without losing important information. An optimal subset of features should include all strongly relevant features, none of the irrelevant features, and a subset of weakly relevant features [48]. However, it is unknown which of the weakly relevant features should be selected and which of them removed. In other words, a feature should be selected if it is relevant but is not redundant to any other relevant feature. Sections 3.2.1 and 3.2.2 describe the redundancy and relevance analysis of features, respectively. Once the feature set per each community is found, communities are represented following the traditional vector space model proposed by Salton et al. [31], in which each vector dimension corresponds to an individual term weighted by its frequency of appearance.

Finally, the feature sets corresponding to each community are used for representing the posts in it, and then for training the learning models that will be used for classifying the newly arriving posts. A model is trained for each community using only the posts that belong to its associated community. Note that first, posts are grouped into communities according to their social and content similarities, and then, the characteristics of each particular community are distinguished by means of these specialised learning models. Each model aims at accurately reflecting the particularities of its training posts. The characteristics of posts belonging to other communities are disregarded in order to avoid introducing noise to the model, allowing an accurate classification of new posts.

3.2.1. Redundancy analysis

The goal of the redundancy analysis is to find all possible redundancies and identify each of the redundant features to be removed. Traditionally, the focus of FS techniques has been on the identification of relevant features, disregarding the explicit analysis of feature redundancy. However, only assessing feature relevance might not identify redundant features, as they are likely to have similar rankings. As long as features are deemed relevant, they will all be selected, regardless their correlations.

For removing redundant features, most techniques rely on subset evaluations, which implicitly handle feature redundancy through feature relevance. Although they can achieve better results than not handling feature redundancy at all, they might be inefficient when analysing high-dimensional data.

The calculation of correlation between two random variables is usually based on either linear correlations or information theory [23,47]. Analysing feature redundancy by means of feature correlations have several benefits [47]. First, it helps to remove features with zero correlation. Second, it helps to effectively detect the redundancy between any pair of features. If data is linearly separable in the original representation, it is still linearly separable if all but one of a group of linearly dependent features are removed. However, these techniques also have limitations as it cannot be assumed that in real-world environments features will be linearly separable.

[1] http://www.dmoz.org/.

For the purpose of this work, the feature redundancy analysis is performed by computing the Pearson correlation. Features that are highly and positively correlated with a certain percentage of features are removed. Then, redundant features are mapped to their most correlated feature, i.e. the feature with which they share the highest correlation value. If redundant features were not mapped to their non-redundant equivalent, there would be no available information for classifying those instances that only contain redundant features that were found to be redundant.

3.2.2. *Relevance analysis*

Relevance techniques are based on assigning weights to the individual features according to their degree of relevance. Then, a subset of features is often selected from the top of a ranking list, aiming at approximating the set of relevant features. However, these techniques are not capable of removing redundant features as they are likely to have similar rankings. Thus, they will be all selected regardless whether they are highly correlated with each other. For high-dimensional data that might contain a large number of redundant features, this situation might define subsets of features far from the optimal [48]. As a result, the relevance analysis should be performed after redundant features have been removed, or at least replaced by their non-redundant equivalent. For the purpose of this work, features inside each community are ranked according to their *TF-IDF* score, and a pre-defined percentage of the highest ranked features is selected for representing such community.

3.3. *Classification of newly arriving posts*

Fig. 4 depicts the processing of newly arriving posts, once the social graph is derived and the textual representation of each community is obtained. Particularly, based on the Original Data Feed and the community structures presented in Figs. 1 and 2, Fig. 4 exemplifies the arrival and posterior processing of a new post, denoted as Post 7 (shown in Step 1 of the Figure).

When a new post to be classified arrives, the community it belongs to is first determined. Such community defines the features for representing the new post, as well as the trained classifier to be used for prediction. As the community detection algorithm does not allow to find the community of a post without affecting the existing community structure, vertex similarity strategies are applied for finding the community the new post belongs to. Both the social and content-based relations of the newly arriving post with the other posts in the existing graph are established (shown in Step 2 in the Figure). The same node relationships that were used for deriving the social graph are now used to establish the relationships between the new node and the existing communities. Note that, as new posts are not yet assigned to any class, relationships including an assessment of posts' class are disregarded. To determine the community a post should be assigned to, i.e. its most similar community, the built graph structure is used for computing the similarity between a node and each community, as the average similarity between the new node and each of the nodes in the community (Step 3 in the Figure).

Most vertex similarity approaches rely on structural and connectivity characteristics. For example, the Sørensen similarity metric (which measures the overlapping between the neighbourhoods of posts, penalising them by the sizes of such neighbourhoods) or the Pearson correlation [32] (which can be computed over the adjacency matrix of the graph). Eqs. (1) and (2) show the definition of both similarity metrics, where p_i and p_j denote the post for which the similarity score is computed, $\Gamma(p_i)$ denotes the set of neighbours of p_i, $|\Gamma(p_i)|$ denotes the degree of post p_i, A denotes the adjacency matrix, and n denotes the number of posts contained in the graph. In order to the similarity scores to be comparable, the Pearson correlation score was normalised to the range [0; 1].

$$Sørensen\left(p_i, p_j\right) = \frac{2\left|\Gamma(p_i) \cap \Gamma\left(p_j\right)\right|}{|\Gamma(p_i)| + \left|\Gamma\left(p_j\right)\right|} \tag{1}$$

$$\begin{aligned} &PearsonCorrelation\left(p_i, p_j\right) \\ &= \frac{n\left(\sum_{k=1}^{n} A_{p_i,p_k} * A_{p_j,p_j}\right) - |\Gamma(p_i)| * \left|\Gamma\left(p_j\right)\right|}{\sqrt{|\Gamma(p_i)| * \left|\Gamma\left(p_j\right)\right| * (n - |\Gamma(p_i)|) * \left(n - \left|\Gamma\left(p_j\right)\right|\right)}} \end{aligned} \tag{2}$$

Preliminary studies over several similarity metrics showed the performance differences amongst the vertex similarity alternatives. They stated the impact of performing an adequate analysis of similarity amongst posts, and thus selecting an adequate vertex similarity alternative for maximising classification results. In all cases, the best results were achieved for the harmonic mean of the two presented metrics. Thus, experimental evaluation considers such metric combination. In Fig. 4, the reported values of similarity correspond to the harmonic mean of the two presented metrics. The computed similarities determined that Post 7 should be assigned to Community 2.

Once the most similar communities and their corresponding classifiers are identified, three alternatives for assigning the post to a category are defined. The first alternative (*Single)* selects the most similar community, and assigns the post to the category with the highest probability within this community. The second (*Average*) and third (*Voting*) alternatives select the *N* most similar communities and classify the post according to their corresponding classifiers. The alternatives differ in how the prediction of the different classifiers are combined for assigning the post to a unique category. The *Average* alternative averages the probabilities of a post belonging to each category, assigning the post to the category with the highest averaged probability. Finally, the *Voting* alternative selects the three categories with the highest probability per classifier and applies a voting scheme, assigning the post to the most voted category. In those cases in which two or more categories share the same number of votes, the post would be assigned to the category with the highest probability. In the particular example, once Post 7 is assigned to Community 2, its textual representation is built and then the learning model corresponding to Community 2 is used for classifying the new Post 7 into class *A* (Step 4 in the Figure). As noted, the *Single* alternative was applied, i.e. the new post was assigned to only one community.

After new posts are classified, they are added to the social graph to update the underlying community structure (shown as Step 5). Thus, the feature space describing communities is periodically updated. Due to the computational complexity of the content analysis step, such updates could be trigger by each new post or after a minimum number of posts have been classified. This restriction tries to prevent a degradation of the technique's performance due to frequent updates.

4. Experimental evaluation

This section presents the experimental evaluation performed to assess the effectiveness of the proposed technique. Section 4.1 details the implementation and experimental settings. Section 4.2 describes the data collection employed. Section 4.3 presents the baselines used for assessing the performance of the presented technique. Finally, Section 4.4 analyses the results achieved from the performed experimental evaluation.

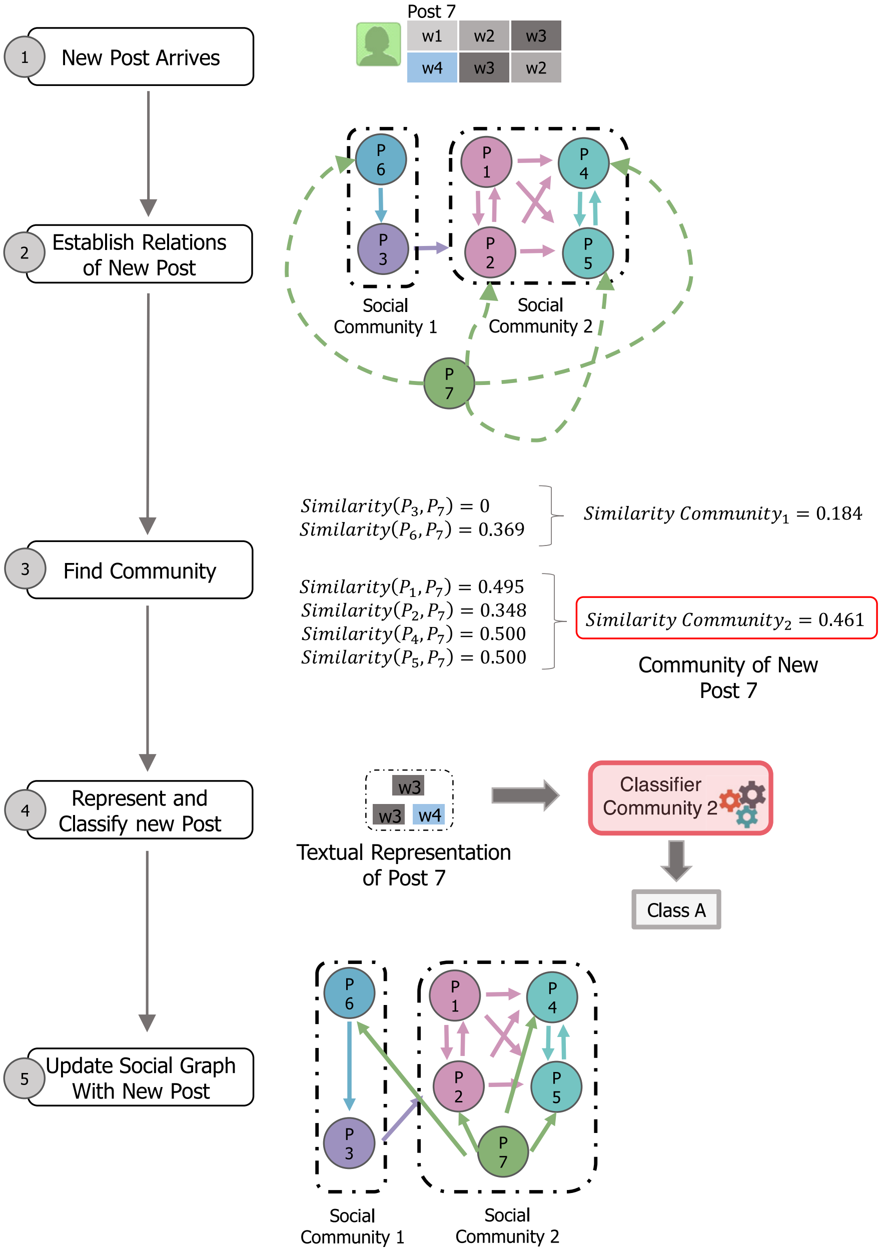

1 New Post Arrives
Post 7
w1 w2 w3
w4 w3 w2
2 Establish Relations of New Post
P 6
P 3
P 1
P 4
P 2
P 5
P 7
Social Community 1
Social Community 2
3 Find Community
$Similarity(P_3, P_7) = 0$
$Similarity(P_6, P_7) = 0.369$
$Similarity\ Community_1 = 0.184$
$Similarity(P_1, P_7) = 0.495$
$Similarity(P_2, P_7) = 0.348$
$Similarity(P_4, P_7) = 0.500$
$Similarity(P_5, P_7) = 0.500$
$Similarity\ Community_2 = 0.461$
Community of New Post 7
4 Represent and Classify new Post
w3
w3 w4
Textual Representation of Post 7
Classifier Community 2
Class A
5 Update Social Graph With New Post
P 6
P 3
P 1
P 4
P 2
P 5
P 7
Social Community 1
Social Community 2

Fig. 4. A schematic view of the arrival and classification of new posts.

4.1. Experimental settings

The Java programming language was chosen for implementing the technique [33]. The social graph was derived based on the social and content-based relations defined in Section 3.1. A minimum similarity of 0.6 was imposed in order to connect two nodes (named *SimilarContent-0.6*). Both the *Independent* (each relation created an edge between two nodes) and the *Weighted* (the content-based features were used for weighting the found social relations) graph derivations were evaluated. Communities were discovered using the graph partition and community detection algorithms implemented in the *Gephi Tookit*[2] [5], an open source library for exploring, filtering, navigating, manipulating and clustering networks. Posts were classified using the WEKA[3] implementation of the Sequential Minimal Optimisation [28] classifier, which is an optimisation of the Support Vector Machines (SVM) classifier [11]. Although the experimental evaluation used the WEKA implementation of the algorithm, it could be change to make the technique more computationally efficient in a real-world setting. Moreover, the presented technique could be instantiated with other learning algorithms.

For all the evaluated cases, sets of different sizes of instances were randomly generated (ranging between 100 to 1000 or 2000 instances, according to the dataset used). For each set, several training and test splits were analysed. In turn, for each given set, five training-test splits were created by gradually incrementing the training set from 10% to 60% of the total number of instances. For each combination of number of instances and training split, five random partitions into training and test set were generated, thus the mean results are reported. Only the best results of the different combinations of training and test set splits are reported. Classification performance was evaluated by means of accuracy, and the

[2] http://gephi.github.io/.
[3] http://www.cs.waikato.ac.nz/ml/weka/.

standard precision and recall measures [4] summarised by the F-Measure.

The selection of redundancy and relevance thresholds was guided by the characteristics of the dataset under analysis, and the statistical distribution of features in the dataset. The selection of the statistical metric to guide such selection is also important as it depends on the type of distribution. Assuming the existence of outliers in the dataset, average measures of data cannot be used as they do not give any indication if the data dispersion. Instead, statistics that are not based on the supposition of a symmetric distribution of data, such as the interquartile range, are needed. Regarding redundancy, an analysis showed that the redundancy scores corresponding to pairs of features (considering a confidence value of 0.01) were homogeneously distributed over the full range of possible values, as the first and third quartiles corresponding to the scores obtained for pairs of features were as similar to the median as the standard deviation was from the mean. In this context, the chosen threshold corresponded to the value associated to the median value, i.e. 0.6. This means that for a potential pair of features to be deemed as redundant, its correlation score should be higher to that of the 50% of the correlation distribution. The percentage of features to which a feature must be highly correlated in order to be considered redundant was set to 60%. A similar analysis was performed for selecting the number of relevant features, i.e. the relevance threshold. In this case, the statistical analysis was based on the mode of the relevance score population and its relation with the quartile distributions. The mode of the relevance scores of features matched the score of the first quartile. This implies that the first 25% of the features shared the same relevance value, and hence do not contribute to the relevance analysis. Consequently, the threshold for feature relevance was set to the 75% of features, i.e. all features excepting the ones with scores equal or lower to the mode were kept. As exposed, the selection of the thresholds responded to the characteristics of the feature distribution in the datasets, and, then, they cannot be directly generalised to different datasets. In case of analysing another dataset, the particular thresholds can be computed by the proposed methods. Note that the statistical properties of the defined threshold could be further explored aiming at optimising their selection.

4.2. Data collections

The performance of the approach was evaluated on two data collections. The first collection comprised data extracted from *Twitter*[4] [51][5], including the content of more than 500,000 tweets belonging to 1,036 trending topics, which were manually assigned to one of 4 broad categories: news, ongoing events, memes (trending topics triggered by viral ideas) and commemoratives (the commemoration of a certain person or event that is being remembered in a given day, for example birthdays or memorials). Tweets were tokenised by means of the tool defined in [16][6], which was specifically designed to process and tag social content. Then, stemming was applied to the obtained tokens. The second collection comprised data extracted from *BlogCatalog*[7] [43][8]. *BlogCatalog* is a blog directory in which users can register their blogs under predefined categories. In addition, users are also encouraged to establish social relationships by following the activity of other users. The data collection includes a summary of the content, categories and tags of each blog, ownership information and non-reciprocal social relationship between users. The selected *BlogCatalog* dataset was used for evaluating the performance of feature selection techniques [36–38], community detection [39], and learning collective behaviour [35], amongst others. For the purpose of the experimental evaluation, content-based features were defined as the tags assigned to each blog. Tags were pre-processed by removing stopwords and applying the Porter Stemmer algorithm [29] to the remaining words. Table 1 summarises the main characteristics of both datasets.

Table 1
Statistical characteristics of the data collections used in the evaluation.

	Twitter	*BlogCatalog*
Number of Instances	1,036	111,648
Number of Features	226,043	60,411
Number of Classes	4	319
Number of Following Relations	251,522,840	3,348,554
Average number of Followees	816	47
Average number of Features per Instance	1084	7
Average number of Instances per Class	259	376

4.3. Baselines for comparison

The proposed OFS technique was compared against three traditional baselines. First, the classification based on a traditional batch FS technique in which features are known in advance and weighted according to their relative frequency (named *Traditional-Batch*). Second, the classification applying the filter FS technique Information Gain (IG) retaining the 75% of the features (named *InfoGain-75*). Third, the classification results achieved by an incremental classification algorithm such as IBL [1], a variation of the traditional *k-NN* in which features were weighted according to their relative frequency (named *Updatable-KNN*).

Additionally, our approach was compared to three state-of-the-art baselines: OFSs [17], OGFS [45] and OFSp [44]. In all cases, the approaches were implemented in Java following the algorithms, details and parameters defined in the papers. OFSs and OFSp were originally conceived to be applied on a binary classification problem. Hence, they were extended to support multi-class classification by using the technique known as one-vs.-rest, in which a classifier is trained per class with the instances of that class as positive instances and all other instances as negatives.

On the other hand, several considerations were introduced to OGFS. First, as the library for solving LASSO (Least Absolute Shrinkage and Selection Operator) used by the authors was only available for C++ and Mathlab, a different Java implementation named “Statistical Machine Intelligence & Learning Engine” (SmileMiner)[9] was used. Second, the authors exposed the concept of “groups of features” that arrive together to the system. In the context of real-time classification of continuously generated short-texts in social networks, each newly arrived post can be regarded as a new group of features that arrives to the system. For the purpose of the experimental evaluation, each new post was considered as the group of features to be analysed. Third, as different posts might comprise overlapping sets of features, features might be analysed several times and also, already selected features could be analysed once again. In this regard, a modification was introduced to analyse each feature only once. Fourth, as the authors proposed three stopping criteria for the FS process but they did not specified which was the one they experimented with, the FS process was stopped when there were no more features to process, i.e. when were no more posts available for training. The classification of new instances was based on the WEKA implementation of the C4.5 algorithm [30], i.e J48, as it was one of the algorithms originally selected by the au-

[4] http://www.twitter.com/.

[5] http://nlp.uned.es/~damiano/datasets/TT-classification.html.

[6] http://www.ark.cs.cmu.edu/TweetNLP.

[7] http://www.blogcatalog.com/.

[8] http://leitang.net/social_dimension.html.

[9] https://github.com/haifengl/smile.

thors for evaluating the approach. Finally, for the purpose of classification selected features were weighted according to their relative frequency.

Results for the *Twitter* dataset were also compared to that of [6] (Becker et al.) and [51] (Zubiaga et al.). In the case of Zubiaga et al., the evaluation considered both the features and the experimental settings defined by the authors. In the case of Becker et al., each trending topic was treated as a cluster from which all features were computed and then used to classify the trending topics. For the experimental results based on the *Twitter* dataset to be comparable to those of the baselines, each trending topic was regarded as a data instance, i.e. each instance comprised a tweet set associated to such trending topic. Thus, the feature vectors of each instance included all terms appearing in such tweet set.

Finally, two additional classification settings for further assessing the importance of integrating both social and content features for performing FS are proposed. First, a batch setting (*Batch*) in which the classifier of each community is only trained once, i.e. classifiers are never updated. For the purpose of the experimental evaluation, the *Batch* strategy was only combined with the *Single* strategy, i.e. only the most similar community was considered. Second, a setting in which a unique classifier (*Unique*) is trained and updated joining all the features belonging to each community in a single feature space. As only one classifier is built, this strategy does not need to find the most similar community for each newly arrived post, but exploits the formed communities that independently select their own non-redundant and relevant set of features.

4.4. Experimental results

This section presents the results of the experimental evaluation performed to assess the effectiveness of the presented OFS technique combining social and content-based information extracted from social media data. Section 4.4.1 presents the results obtained for the *Twitter* dataset and Section 4.4.2 those obtained for the *BlogCatalog* dataset. In both cases, the performance of the two proposed derivations of the social graph are analysed.

4.4.1. Results for the Twitter dataset

Fig. 5 shows the results for the *Independent* derivation of the social graph applied to the *Twitter* dataset. As regards accuracy, the best baseline scores were obtained by the *Twitter*-specific techniques (Becker et al. and Zubiaga et al.). Interestingly, those alternatives relaying only on a traditional assessment of content-based features obtained the worst results in all cases. Moreover, using IG as a filter for selecting features in a batch scenario obtained worse results than using all features weighted according to their relative frequency. This further exposes the limitations of traditional batch content-based feature selection approaches for performing short-text classification. Similarly, considering a more dynamic classification setting in which features are updated when new posts appear (represented by *Updatable-KNN)* was not sufficient for achieving good performance.

All baselines were outperformed by every of our OFS strategies, stating the importance of leveraging not only on content-based features, but also on social information. Note that, the best results were obtained when learning from a small number of instances. As the number of posts increased, results did not further improve, but instead, they remained unchanged or even decreased. This situation reinforces the necessity of updating the community structure by not only adding newly classified instances, but also removing old instances to adapt to new tendencies or topics emerging on social media. This is further evidenced by the results of the *Batch* strategy, which outperformed the results only based on content, but achieved worse results than other alternatives including updates of the social structures.

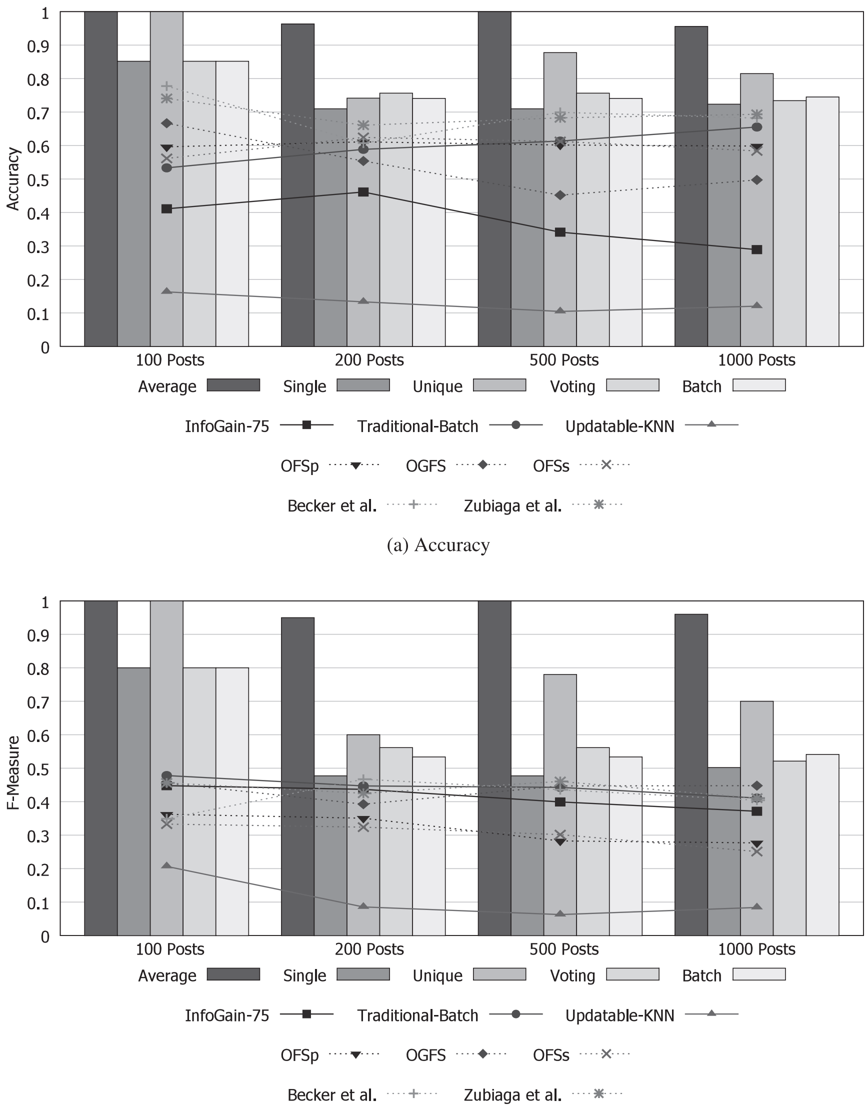
F-Measure results were alike. The highest improvements of our OFS strategies were obtained when analysing and classifying low numbers of posts. Note that in such case, both *Average* and *Unique* were able to obtain perfect precision and recall, demonstrating the benefits of combining social and content-based information. Interestingly, for this evaluation metric, the results of the *Twitter*-specific baselines were similar to those of the *Traditional-batch* and *InfoGain-0.75* and even lower than those of *OGFS* when analysing more than 1000 posts. These results imply that even when the baselines seemed to be accurate, they were not able to perform predictions with high precision.

Results obtained for the *Weighted* derivation of the social graph are depicted in Fig. 6, which are higher than all of the baselines. This graph derivation obtained similar results than the *Independent* one, excepting for low numbers of instances. In that case, the *Independent* derivation outperformed the *Weighted* one. The highest differences were obtained for the *Unique* and *Voting* strategies, which obtained the best results for the *Independent* derivation. These results highlight the importance of adequately leveraging the information extracted from the social networking site in order to improve the quality of classifications. The *Average* and *Unique* strategies results continue to demonstrate the importance of accurately determining social and content-based relations between posts, and thus, finding the most similar social matches to new ones. The differences between *Average* and *Voting* emphasise the importance that finding the most similar community has on the subsequent classification of new instances. Regarding our OFS strategies, the worst results were obtained with *Single*. These results might have different explanations. First, instances might belong to more than one community, i.e. communities might overlap, meaning that an individual community might not be sufficient for accurately describing the content of new instances. Second, when including multiple communities in the analysis, it is important to choose how to combine their results to maximise classification performance. Finally, as the results of the *Batch* strategy were lower to those of *Average* and *Unique*, they demonstrate the importance of not only including social information in the selection of textual features, but also of adapting the set of selected features according to the continuous appearance of new topics and posts in the social media stream.

4.4.2. Results for the BlogCatalog dataset

Fig. 7 shows the results for the *Independent* derivation of the social graph when applied to the *BlogCatalog* dataset. As regards accuracy, the best baseline results were obtained by OFSp. In spite of knowing all the feature space in advance, traditional baselines were outperformed by our OFS strategies. Interestingly, those alternatives relaying only on a traditional assessment of content-based features obtained the worst results in all cases. When analysing few instances, selecting features with *InfoGain-75* obtained worse results that including all features. Similarly to the *Twitter* dataset, *Updatable-KNN* obtained worse results than the traditional baselines. These results show the shortcomings of classifying social posts solely based on their content, and the importance of also including social information. In all cases, our approach achieved better results that the state-of-the-art baselines.

Unlike the results obtained for the *Twitter* dataset, as the number of analysed instances increased, the accuracy results of our OFS strategies improved, whilst those of the baselines decreased. These results further emphasise the importance of including social information for accurately selecting features. Nonetheless, they also implied that the rate of new topic generation is not as fast as in *Twitter*, as even when old topics are not removed, results continue to improve. Similarly to the results for the *Twitter* dataset, of the



(a) Accuracy

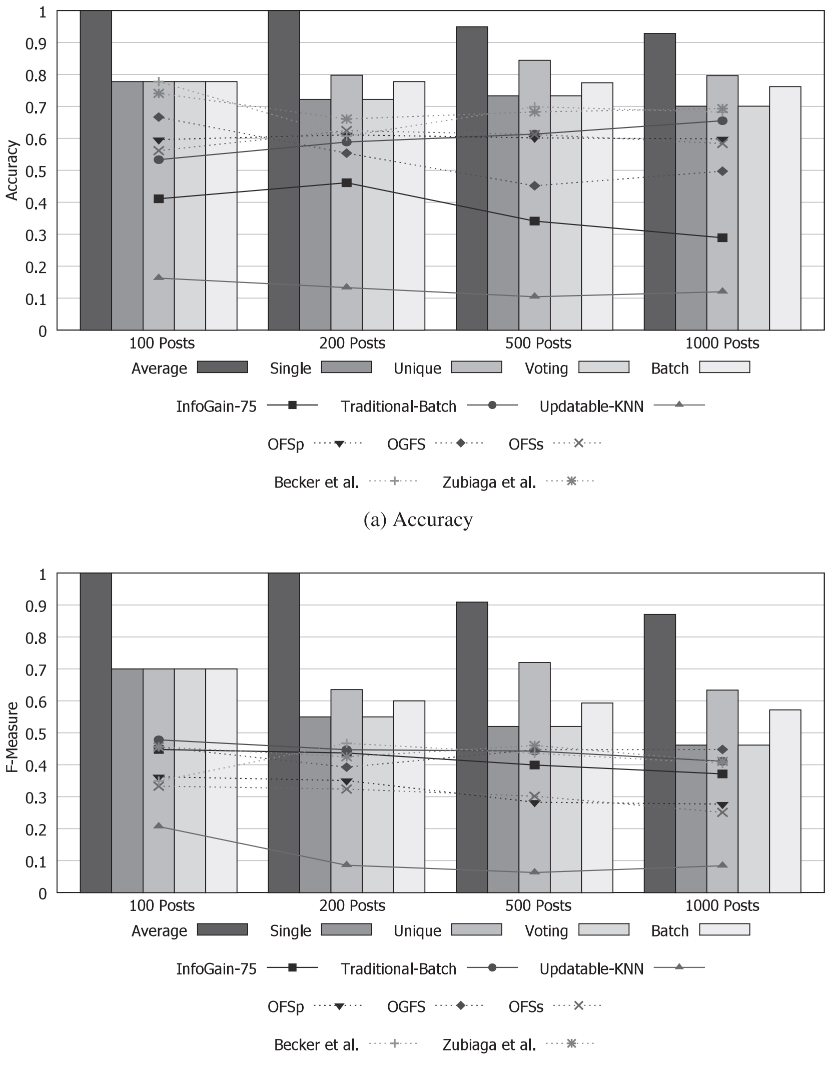
(b) F-Measure

Fig. 5. Results of classification using the *Independent* graph derivation for the *Twitter* dataset.

presented OFS strategies, the best results were obtained with *Average*, closely followed by the other strategies. However, the differences amongst the other alternatives are smaller than the differences found in the *Twitter* dataset. This could imply that joining multiple communities in the class assessment of instances does not enrich the representation of such instances. Moreover, as the number of instances increased, the differences in accuracy tended to decreased.

F-Measure results did not follow the same tendencies as accuracy, adopting a similar distribution than those for the *Twitter* dataset. Even though all baselines were outperformed by our strategies regardless the number of posts, the highest improvements were obtained for low numbers of posts. In this case, the best baselines results were obtained for OFSs, followed by OFSp. Note that the best performing baselines differ from those found for the *Twitter* dataset, highlighting the intrinsic differences amongst the datasets. For this metric, it is possible to observe a performance difference between the presented OFS strategies. In this case, *Average* obtained the best results, followed by *Unique* and *Single*. These results imply that even when the techniques seemed to be accurate, they were not able to perform high precision predictions. Furthermore, they validate the importance of social information for obtaining high quality results. The decrement in F-Measure results could indicate (as for the *Twitter* dataset) the necessity of not only frequently updating the community structure and the sets of relevant features, but also performing a selection of the posts to include in the community analysis.

Results obtained for the *Weighted* derivation of the social graph are depicted in Fig. 8. This graph derivation obtained similar accuracy results than the *Independent* one. As regards F-Measure, results were higher than for the *Independent* derivation, and accentuate even more the superiority of *Average*. Such differences continue to state the importance of adequately leveraging the information extracted from the social networking site to improve the quality and precision of classifications. From these differences, it could be inferred that treating social and content relations independently adds noise to the processing, resulting in lower classification performance. Hence, it could be stated that, even though content relations add useful community information, for this dataset, the social component is more relevant for determining the community structure of posts.



(a) Accuracy

(b) F-Measure

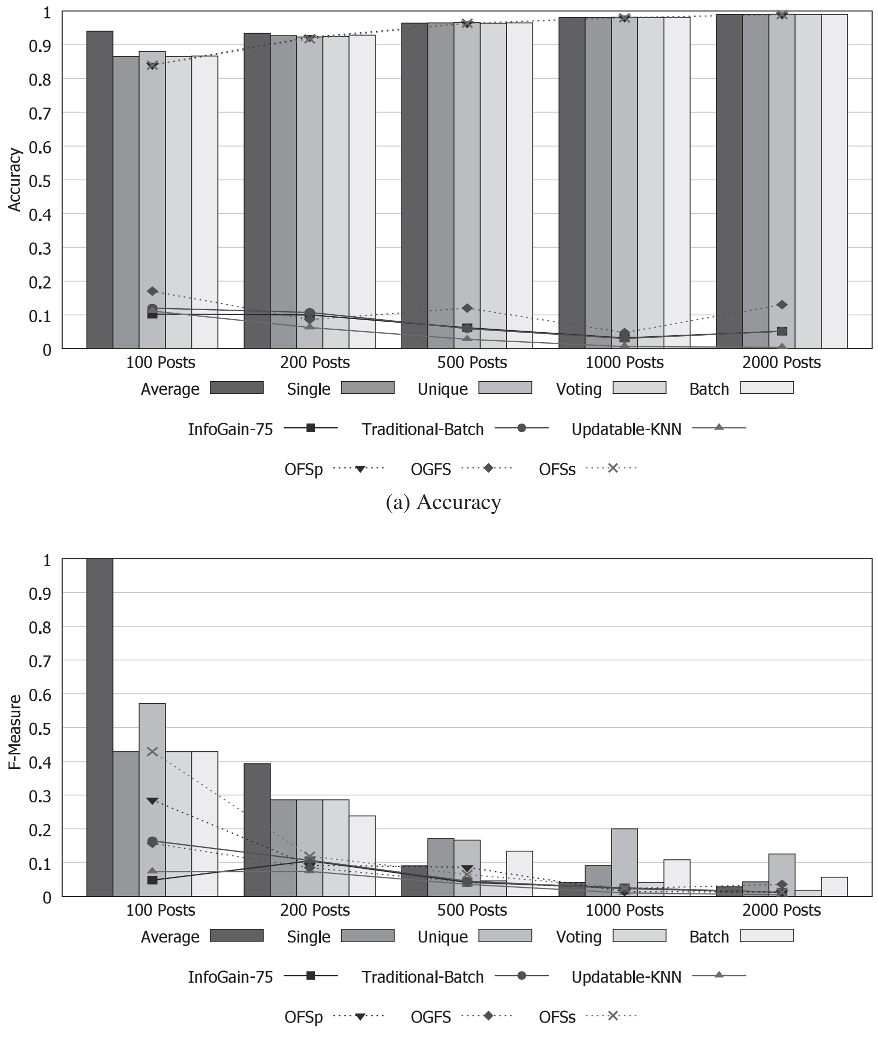
Fig. 6. Results of classification using the *Weighted* graph derivation for the *Twitter* dataset.

The differences between *Average* and *Voting* remark the importance that finding the most similar community has on the subsequent classification of new instances, strengthening the importance of choosing how to combine the results of such communities for maximising classification performance. As for the previous dataset, the worst results were obtained for *Batch*, showing that it is not sufficient to consider social information for performing feature selection, but it is also necessary to adapt the set of selected features according to changes in the discovered instances. Finally, the fact that *Single* obtained worst results than *Average* might indicate that posts could belong to overlapping communities, meaning that an individual community is not sufficient for accurately describing the content of new instances.

4.4.3. Summary of results

Table 2 summarises the F-Measure improvements of the best strategies of our OFS technique for the *Independent* social graph derivation, i.e. *Unique, Average*, and *Voting* or *Single* (depending on the analysed dataset), over the best traditional baseline (*Traditional-Batch*) and the best state-of-the-art baseline (OGFS or OFSs for the *Twitter* and *BlogCatalog* datasets, respectively). Note that the best performing state-of-the-art baseline varies according to the dataset under analysis. Additionally, in the case of the *Twitter* dataset, the improvements over the best *Twitter*-specific baseline (Zubiaga et al.) are reported. For every number of evaluated posts and both datasets, the best strategies outperformed almost every baseline results. This shows the relevance and superiority of the presented OFS technique over both traditional techniques based solely on content, and state-of-the-art OFS techniques. For the *Twitter* dataset, baselines results were improved up to a 143%. Interestingly, the highest improvements for the *Twitter* dataset were obtained regarding the OFS techniques. In the case of the *BlogCatalog* dataset, baselines results were improved up to a 23,531,145.13 % regarding traditional baselines and 695% regarding the best state-of-the-art baseline. These results reinforce the suitability of our OFS technique in the context of social media data.

When comparing the results obtained for the two graph derivations of the social graph, it can be observed that they varied according to the chosen derivation. For the *Twitter* dataset, in overall, the *Independent* derivation obtained slightly higher results than the *Weighted* one. On the contrary, for the *BlogCatalog* dataset, the *Weighted* derivation allowed to obtain the highest results. As exposed in [40], the intrinsic characteristics of the network under analysis influence the quality of communities achieved for the diverse combinations of relations and graph derivation. Hence, the results differences might account for the intrinsic characteristics



(a) Accuracy

(b) F-Measure

Fig. 7. Results of classification using the *Independent* graph derivation for the *BlogCatalog* dataset.

Table 2
Summary of F-Measure improvements over traditional and state-of-the-art baselines (%).

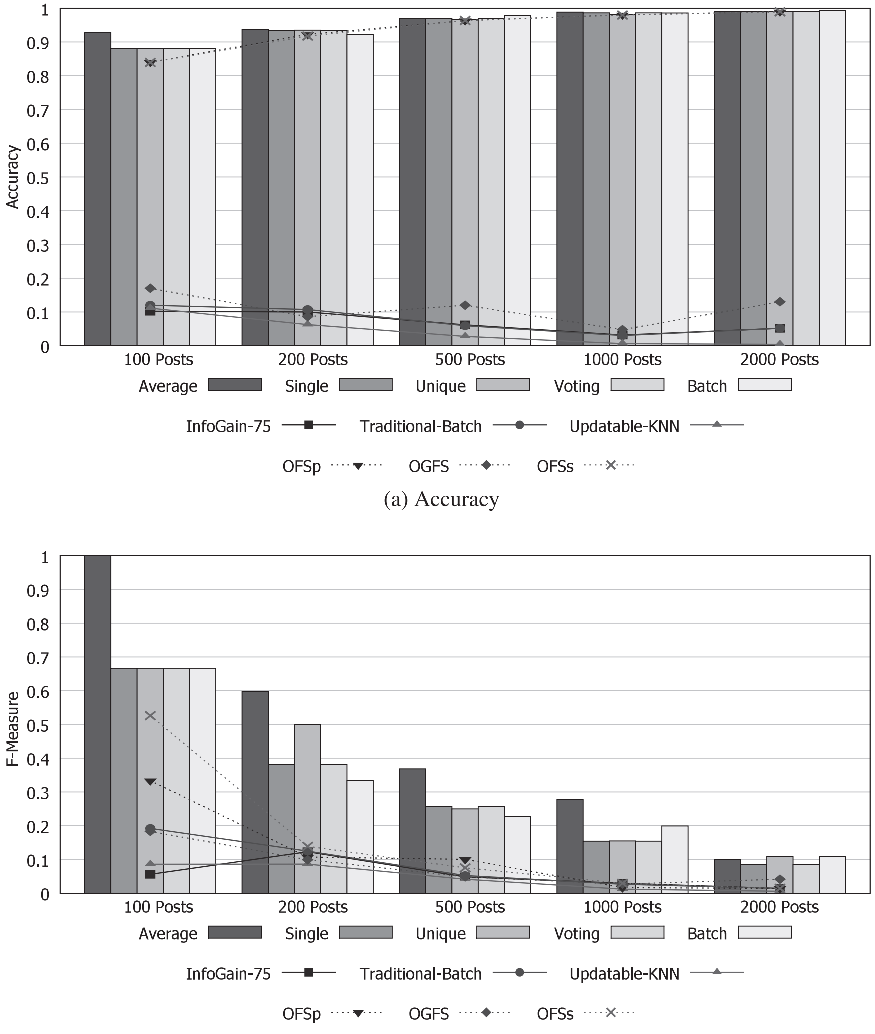
(a) Improvements for the *Twitter* dataset

	100 posts			200 posts			500 posts			1000 posts		
	Trad.	OGFS	Zub	Trad.	OGFS	Zub	Trad.	OGFS	Zub	Trad.	OGFS	Zub.
Unique	109.09	119.41	119.15	34.14	52.74	41.04	76.04	74.39	69.30	70.19	56.26	70.40
Average	109.09	119.41	119.15	112.39	141.83	123.31	125.69	123.58	117.05	133.41	114.30	133.69
Voting	67.27	75.53	75.32	25.60	43.01	32.06	26.79	25.60	21.94	26.81	16.43	26.96

(b) Improvements for the *BlogCatalog* dataset

	100 posts		200 posts		500 posts		1000 posts		2000 posts	
	Trad.	OFSs	Trad.	OFSs	Trad.	OFSs	Trad.	OFSs	Trad.	OFSs
Unique	86739.03	33.33	266954.77	140.00	1586839.15	157.75	6281951.28	695.34	23531145.13	923.09
Average	507.87	133.33	267.20	230.00	101.01	40.04	82.62	65.21	133.31	131.02
Single	160.52	0.00	167.05	140.00	150.00	165.33	301.76	263.47	253.02	249.55

of *Twitter* and *BlogCatalog*, which affect the community detection process, and the posterior processing of new posts. For example, in information oriented networks as *Twitter* is, it is expected that content-based relations would be more important than social ones, implying that independently assessing both types of relations (i.e. the *Independent* graph derivation) would allow to achieve higher community quality than weighting the social relation with the content ones (i.e. the *Weighted* graph derivation). Conversely, *BlogCatalog's* goal is to foster the connection amongst bloggers highlighting the social aspect instead of simple being a blog directory. As a result, it is expected that the *Social* relation would be more relevant whilst content-based relations might introduce noise, thus favouring the *Weighted* relation. In summary, these differences accentuate not only the importance of adequately leveraging on the



(a) Accuracy

(b) F-Measure

Fig. 8. Results of classification using the *Weighted* graph derivation for the *BlogCatalog* dataset.

information extracted from the social networking site, but also on their intrinsic characteristics to optimise the quality of communities, and thereby improve the overall OFS performance.

A statistical analysis based on [10] was performed to determine whether the differences amongst results were significant. As data was shown not to be normal, the Friedman test for related samples was applied to the results obtained for each baselines and our OFS technique. Particularly, the results obtained for each of the generated training-test set splits of the defined random partitions was regarded as a sample. To perform the test, two hypotheses were defined: the null and the alternative hypothesis. The null hypothesis stated that no difference existed amongst the results of the different samples, i.e. all the evaluated FS techniques performed similarly. On the contrary, the alternative hypothesis stated that the differences amongst the FS techniques were significant and non-incidental. The statistical test showed that (with a confidence of 0.01) the differences amongst results were statistically significant (in all presented cases the obtained value was higher than the corresponding critical value). Then, to specifically test whether the differences amongst the results obtained for our OFS technique and the selected baselines were statistically significant, the Wilcoxon test was applied. The same hypotheses were defined. The statistical test showed that for the alternatives summarised in Table 2, the null hypothesis could be rejected with a confidence of 0.01, meaning that the results differences were significant and not due to chance. Moreover, the test showed that the results of the presented OFS technique in all cases were statistically higher than those of the evaluated baselines and the state-of-the-art techniques. As regards the remaining results, in most cases the statistical superiority of the presented OFS technique was maintained. The exception was found for the *Voting* alternative in the *Weighted* derivation and the F-Measure evaluation metric.

In summary, our technique can be applied in real-world problems in which batch FS approaches might not be suited, achieving even better performance than state-of-the-art techniques specifically design for social media. Moreover, purely content-based strategies might not be sufficient for classifying social media texts, due to the limited number of available features, confirming the importance of the social relations between users. Thus, leveraging on social information becomes crucial to OFS techniques. It is worth noting that the intrinsic characteristics of both datasets showed to have an effect on the presented technique as the proposed social graph derivation alternatives obtained diverse results for both datasets. Finally, the performed statistical analysis strengthened the superiority of the presented OFS technique over the analysed baselines and state-of-the-art techniques.

5. Conclusions

Feature selection is one of the most known and commonly used techniques for diminishing the impact of the high-dimensional feature space, which is reduced by removing redundant and irrelevant features. The standard FS setting assumes the existence of a fixed set of instances, and therefore a feature space fully known in advance. In real-world applications, such assumptions might not hold as either training examples could arrive sequentially, features might appear incrementally or it could be difficult to collect a full training set [44]. In these situations, OFS techniques which consider the arrival of instances and their corresponding features in a continuous stream should be used. OFS techniques involve choosing a subset of features and its corresponding learning model at different time frames. However, the challenges posed by OFS remain open as most studies in the literature are focused on developing batch techniques instead of online ones. As a result, in order to mine big data in real-world applications, new online techniques for efficiently identifying a number of relevant features and then build accurate predictions models have to be developed [17].

This work aimed at assessing both social and content-related factors for real-time classification of continuously generated short-texts in social networks. The proposed technique tackled the challenging problem of OFS, which is an important requirement in numerous large-scale social applications. Although the presented technique is applied to multi-class classification of socially generated posts, it can be also used in both binary and multi-class settings, and even for other learning tasks, such as clustering.

Experimental evaluation conducted on real-world social media datasets demonstrate that the proposed technique helps to improve classification results when compared to traditional and state-of-the-art FS techniques in both batch and online settings, exposing the limitations of pure content-based techniques for social text classification. The obtained results evidence the importance of the social relations amongst users for classifying short-texts in social media, and its advantages for selecting the most relevant set of features. Although the preliminary implementation significantly improved precision results of state-of-the-art techniques found in literature, for one dataset, precision results are still lower than those achieved for traditional tasks of text classification. This situation highlights the difficulty of the task and the need of continuing to develop, improve and evaluate new techniques.

As regards future work, an extensive experimental evaluation of all the parameters involved in the OFS technique must be performed. Moreover, the performance of other methods for determining the redundancy and relevance of features should be analysed. For example, mutual information could be used for simultaneously selecting non-redundant and relevant features. Additionally, new graph representations for further exploiting the topology structure of social relations and communities could be defined. For example, the chosen graph representation collapses possibly heterogeneous information into a unique and homogeneous space, ignoring the possible differences amongst such relations. Hence, a multi-graph representation in which each relation is represented as a separated dimension could be devised, which would also allow to optimise the community partition at each dimension individually. Regarding community detection, the possibility of analysing overlapping communities will be explored. Finally, other applications to the technique will be analysed. For example, the technique could be inserted in the context of a followee recommendation system, under the hypothesis that information regarding the existence of communities of users can help to improve followee prediction quality.

References

[1] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, in: Machine Learning, 1991, pp. 37–66.

[2] E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008) 1981–2014.

[3] S. Alelyani, J. Tang, H. Liu, Feature selection for clustering: a review, in: Data Clustering: Algorithms and Applications, 2013, pp. 29–60.

[4] R.A. Baeza-Yates, B.A. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., MA, USA, 1999.

[5] M. Bastian, S. Heymann, M. Jacomy, 2009, ICWSM'09. Proceedings of the Third International Conference on Web and Social Media

[6] H. Becker, M. Naaman, L. Gravano, Beyond trending topics: real-world event identification on twitter, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17–21, 2011, 2011.

[7] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.

[8] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech: Theory Exp. 2008 (10) (2008) P10008.

[9] D.M. Boyd, N.B. Ellison, Social network sites: definition, history, and scholarship, Journal of Computer-Mediated Communication 13 (1) (2007). Article 11

[10] G.W. Corder, D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, New Jersey: Wiley, 2009.

[11] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[12] Tweet categorization by combining content and structural knowledge, Inf. Fusion 31 (2016) 54–64.

[13] C.H.Q. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (2) (2005) 185–206.

[14] J.C. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (2009) 2899–2934.

[15] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (35) (2010) 75–174.

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N.A. Smith, Part-of-speech tagging for twitter: annotation, features, and experiments, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, Association for Computational Linguistics, 2011. HLT '11, pages 42–47, Stroudsburg, PA, USA

[17] S.C.H. Hoi, J. Wang, P. Zhao, R. Jin, Online feature selection for mining big data, in: Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, ACM, New York, NY, USA, 2012, pp. 93–100. BigMine '12

[18] G.H. John, R. Kohavi, K. Pfleger, et al., Irrelevant features and the subset selection problem, in: Machine Learning: Proceedings of the Eleventh International Conference, 1994, pp. 121–129.

[19] J. Lee, J. Lee, Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction, PLoS ONE 8 (4) (2013) 1–11. 04

[20] C. Li, A. Sun, A. Datta, Twevent: segment-based event detection from tweets, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2012, pp. 155–164. CIKM '12

[21] J. Li, X. Hu, J. Tang, H. Liu, Unsupervised streaming feature selection in social media, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2015, pp. 1041–1050. CIKM '15

[22] D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. London B: Biol. Sci. 270 (2003) S186–S188. (Suppl 2)

[23] M. Makrehchi, M.S. Kamel, Aggressive feature selection by feature ranking, in: H. Liu, H. Motoda (Eds.), Computational Methods of Feature Selection, Chapman and Hall/CRC Press, 2007.

[24] F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: a survey, Phys. Rep. 533 (4) (2013) 95–142. Clustering and Community Detection in Directed Networks: A Survey

[25] A. Marin, B. Wellman, Social network analysis: an introduction, in: J. Scott, P.J. Carrington (Eds.), The SAGE handbook of social network analysis, SAGE Publications Ltd, London, 2014, pp. 11–25, doi:10.4135/9781446294413.n2.

[26] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Sociol 27 (1) (2001) 415–444.

[27] S. Perkins, J. Theiler, Online feature selection using grafting, in: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, AAAI Press, 2003, pp. 592–599. ICML'03

[28] J.C. Platt, Fast Ttraining of Support Vector Machines using Sequential Minimal Optimization, MIT Press, Cambridge, MA, USA, 1999. pages 185–208

[29] M.F. Porter, An algorithm for suffix stripping, in: Readings in Information Retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 313–316.

[30] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[31] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620.

[32] S.E. Schaeffer, Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64.

[33] G.L. Taboada, S. Ramos, R.R. ExpÃşsito, J. TouriÃśo, R. Doallo, Java in the high performance computing arena: research, practice and experience, Sci. Comput. Program. 78 (5) (2013) 425–444.

[34] J. Tang, H. Liu, Feature selection with linked data in social media, 2012,

[35] L. Tang, H. Liu, Scalable learning of collective behavior based on sparse social dimensions, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2009, pp. 1107–1116. CIKM '09

[36] J. Tang, H. Liu, Feature selection with linked data in social media, in: SDM, SIAM / Omnipress, 2012a, pp. 118–128.

[37] J. Tang, H. Liu, Unsupervised feature selection for linked social media data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2012b, pp. 904–912. KDD '12

[38] J. Tang, H. Liu, An unsupervised feature selection framework for social media data, IEEE Trans. Knowl. Data Eng. 26 (12) (2014) 2914–2927.

[39] J. Tang, X. Wang, H. Liu, Integrating social media data for community detection, in: Proceedings of the 2011 International Conference on Modeling and Mining Ubiquitous Social Media, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 1–20. MSM'11

[40] A. Tommasel, D. Godoy, Integrating heterogeneous information from social networks into community detection, 4th IJCAI Workshop on Heterogeneous Information Network Analysis (HINA), 2016a. New York, NY, USA

[41] A. Tommasel, D. Godoy, Short-text feature construction and selection in social media data: a survey, Artif. Intell. Rev. (2016b) 1–38.

[42] A. Tommasel, Integrating social network structure into online feature selection, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, 2016, pp. 4032–4033.

[43] X. Wang, L. Tang, H. Gao, H. Liu, Discovering Overlapping Groups in Social Media, in: Proceedings of the 2010 IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2010, pp. 569–578. ICDM '10

[44] J. Wang, P. Zhao, S.C.H. Hoi, R. Jin, Online feature selection and its applications, IEEE Trans. Knowl. Data Eng. 26 (3) (2014) 698–710.

[45] J. Wang, M. Wang, P.-P. Li, L. Liu, Z.-Q. Zhao, X. Hu, X. Wu, Online feature selection with group structure analysis, 2015, Volume 27, pages 3029–3041.

[46] X. Wu, K. Yu, H. Wang, W. Ding, Online streaming feature selection, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21–24, 2010, Haifa, Israel, Omnipress, 2010, pp. 1159–1166.

[47] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of the Twentieth International Conference on Machine Learning, AAAI Press, 2003, pp. 856–863.

[48] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.

[49] R. Zafarani, H. Liu, Users joining multiple sites: friendship and popularity variations across sites, Inf. Fusion 28 (2016) 83–89.

[50] J. Zhou, D.P. Foster, R.A. Stine, L.H. Ungar, Streamwise feature selection, J. Mach. Learn. Res. 7 (2006) 1861–1885.

[51] A. Zubiaga, D. Spina, R. Martínez, V. Fresno, Real-time classification of twitter trends, J. Assoc. Inf. Sci. Technol. 66 (3) (2015) 462–473.