Artif Intell Rev
DOI 10.1007/s10462-016-9528-0

Short-text feature construction and selection in social media data: a survey

Antonela Tommasel[1] · **Daniela Godoy**[1]

Abstract Social networking sites such as Facebook or Twitter attract millions of users, who everyday post an enormous amount of content in the form of tweets, comments and posts. Since social network texts are usually short, learning tasks have to deal with a very high dimensional and sparse feature space, in which most features have low frequencies. As a result, extracting useful knowledge from such noisy data is a challenging task, that converts large-scale short-text learning tasks in social environments into one of the most relevant problems in machine learning and data mining. Feature selection is one of the most known and commonly used techniques for reducing the impact of the high dimensional feature space in text learning. A wide variety of feature selection techniques can be found in the literature applied to traditional, long-texts and document collections. However, short-texts coming from the social Web pose new challenges to this well-studied problem as texts' shortness offers a limited context to extract enough statistical evidence about words relations (e.g. correlation), and instances usually arrive in continuous streams (e.g. Twitter timeline), so that the number of features and instances is unknown, among other problems. This paper surveys feature selection techniques for dealing with short texts in both offline and online settings. Then, open issues and research opportunities for performing online feature selection over social media data are discussed.

Keywords Feature selection · Short-text · Social media data · Text learning

1 Introduction

Since their beginning, social networking sites such as *MySpace*, *Facebook*, or *Twitter* have attracted millions of users, who have in turn integrated them into their daily life. As a consequence of the massive adoption and popularity of these sites, social media data is growing at

✉ Antonela Tommasel
antonela.tommasel@isistan.unicen.edu.ar

[1] ISISTAN, UNICEN-CONICET, Paraje Arroyo Seco, Campus Universitario, Tandil, Buenos Aires, Argentina

Published online: 15 November 2016

an unprecedented rate. For example, *Facebook* has grown from 1190 million active users in 2013 to 1317 million in July 2014, out of which only 802 million login daily. Every day, in average, 5 new profiles are created per second. As regards posting activity 293,000 statuses are updated, 136,000 photos are uploaded, and 510 comments and more than 3 million likes are posted per minute. This represents a 94% increase in the posting activity with respect to 2012. Meanwhile, *Twitter* has grown from 500 million users in 2013 to 645 million in 2014, out of which only 271 million are monthly active. Compared to *Facebook*, *Twitter* has a slower increase rate as only 2 new profiles are created per second. As regards posting activity, 500 million tweets are posted per day, representing a 47% increase regarding 2012.[1]

The continuous growing of social media changes the role of users, from traditional content consumers to both content creators and consumers, who generate an enormous number of short-texts (e.g. tweets or posts). The quality of these texts varies from actual valuable content to spam. Furthermore, social media texts are usually informally written and suffer from misspelling, grammatical and punctuation mistakes. In consequence, organising and extracting useful knowledge from such noisy data are challenging tasks, which convert large-scale short-text learning tasks in social environments into one of the most timely and relevant problems in machine learning and data mining.

Text learning tasks are characterised by the high dimensionality of their feature space, in which most terms have a low frequency, i.e. a long-tail distribution. Indeed, text learning is often susceptible to the problem known as the "curse of dimensionality", which refers to the increasing computational complexity of learning tasks as the data that needs to be accessed grows exponentially regarding the underlying space dimension. Furthermore, as data dimensionality increases, the size of the feature space rapidly grows and the available data becomes sparser. Additionally, the linked nature of social media data generates new information, such as who creates the posts (post authorship, i.e. user-post relations), and who is friend of whom (friendship, i.e. user-user relations), which can be added to the feature space (Tang and Liu 2012). Feature selection (Alelyani et al. 2013) 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 goal of feature selection techniques is to select a subset of features that could efficiently describe the input data, whist reducing the effect of irrelevant features on prediction results.

Although text feature selection techniques have been extensively studied in literature (Alelyani et al. 2013; Tang et al. 2014c; Yang and Pedersen 1997), short-text feature selection and its challenges have received comparatively less attention. For example, when considering short-texts, the feature space is sparser, making it difficult to fully exploit the correlation between features. In this context, novel research problems and applications emerge, demanding new feature selection techniques. Interestingly, the research opportunities on short-text feature selection have not been fully explored yet. For instance, the standard feature selection techniques assume the existence of a fixed set of instances. However, spam detection or trending topic detection in *Twitter*, among other tasks, impose additional restrictions to the feature selection technique as instances, such as e-mails or tweets, arrive sequentially and, thus, features incrementally appear. These tasks hinder the deployment of efficient and scalable standard feature selection techniques, hence creating novel environments for using the techniques.

This survey reviews the field of short-text feature selection aiming at describing state-of-the-art techniques, illustrating how feature selection techniques have been applied to real-

[1] http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.

world social media applications, and discussing limitations and current unsolved issues of these techniques. A comprehensive view of state-of-the-art approaches for short-text feature selection is given, identifying their advantages and limitations when applied to learning tasks such as classification or clustering of social texts. To our knowledge, this is the only survey that specifically focuses on short-text feature selection approaches and their applications to learning tasks. A main contribution of this work is presenting open research questions and their relevance to the topic, which, in turn, might be useful to define new research lines.

The rest of this survey is organised as follows. Section 2 presents general concepts related to feature selection techniques. Most surveys organise reviewed techniques based on how the feature subset is selected, disregarding the assumptions made concerning the features' availability or the environmental settings in which techniques could be applied. Conversely, this survey primarily organises techniques based on the environment in which they can be used, considering the novel requirements for the feature selection task imposed by the continuous development of short-texts in social environments. In this regard, Sect. 3 describes in detail feature selection techniques specifically design for dealing with social media text in batch (or offline) settings, whereas Sect. 4 describes techniques for online (or streaming) settings. Then, the techniques in each section are organised based on whether the features are assessed individually or in groups. Considering the reviewed works, Sect. 5 analyses open issues and research opportunities for performing online feature selection over short-texts. Finally, Sect. 6 presents the conclusions of the survey.

2 An overview of feature selection

According to John et al. (1994), features can be classified into three disjoint categories, i.e. strongly relevant, weakly relevant, and irrelevant features. In turn, weakly relevant features can be further classified into two categories, weakly relevant and redundant features, and weakly relevant and non-redundant features. Strong relevant features are always necessary for defining an optimal subset, i.e. they cannot be removed without losing important information. Weakly relevant features are not always necessary, but might become necessary for defining an optimal subset of features. Irrelevant features can be disregarded as they are not necessary. Feature redundancy represents a type of data dependency that is normally defined in terms of feature correlation. It is accepted that two features are redundant if their values are completely correlated. If one feature is considered redundant, its removal does not cause a loss of information. Although the theoretical definition is simple, it might not be simple to determine feature redundancy when a feature is correlated (or partially correlated) with a set of features (Yu and Liu 2004). An optimal subset of features should include all strongly relevant features, none of the irrelevant features, and a subset of weakly relevant features. However, each time a feature is removed, there is the risk of also removing potentially useful information from the texts. Dimensionality reduction techniques have been studied to remove noisy and redundant features, and keep relevant features, i.e. to address the problem of the curse of dimensionality.

There are several potential benefits of applying dimensionality reduction techniques (Guyon and Elisseeff 2003), it reduces the measurement and storage requirements as well as the training times, and facilitates data visualisation and understanding. Furthermore, it also allows to improve prediction performance by avoiding overfitting (Sebastiani 2002). Overfitted learning models are tuned to the contingent characteristics of the training data, instead of inferring generalised data characteristics. They tend to achieve good performance when making predictions over the training data, but to fail when predicting over new or unseen data.

Dimensionality reduction techniques can be classified into Feature Extraction and Feature Selection (Tang et al. 2014c).

Feature extraction approaches project the original features into a new constructed feature space with lower dimensionality. In this case, as the original feature space is mapped into a new space by combining the original features, it might be difficult to link the original features to the new ones. As a result, further analysis of the new features results problematic since there is no physical meaning for the obtained transformed features. Conversely, feature selection approaches aim at selecting a small feature subset that minimises redundancy and maximises relevance. As features maintain their physical meaning, feature selection techniques are superior in terms of readability and interpretability. In turn, this property results of great importance in many practical applications, such as building a sentiment lexicon for sentiment analysis tasks, among other examples. Consequently, feature selection (Alelyani et al. 2013) is one of the most known and commonly used dimensionality reduction techniques.

Feature selection (FS) algorithms can be regarded as a combination of a search technique for finding feature subsets with an evaluation measure that scores the different feature subsets (Guyon and Elisseeff 2003), and are traditionally organised into four categories (Liu and Yu 2005; Alelyani et al. 2013; Saeys et al. 2007) filter, wrapper, hybrid and embedded depending on how the feature subset is selected. *Filter techniques* consider the intrinsic statistical characteristics of features independently of any classifier. *Wrapper techniques* select the feature subset with the most discriminative power regarding a specific learning algorithm, which makes them more computationally complex than filter techniques. *Hybrid techniques* first use statistical criteria to select candidate features subsets with a specific cardinality, and then choose the subset with the highest performance according to a learning algorithm. Finally, *embedded techniques* perform FS simultaneously to other data-mining tasks. As the search for the best feature subset is built into the construction of a learning model, embedded methods are also specific to a given algorithm.

Feature selection techniques can be used in combination with both supervised and unsupervised learning models. Unsupervised FS is performed when the class labels of instances are unknown. In contrast, supervised FS techniques select a subset of highly discriminant and relevant features guided by class information. Such feature subset should be useful for discriminating the instances belonging to different classes. However, obtaining such labelled instances could be time consuming. The problem worsens in online environments in which short-texts are constantly generated. Whilst unsupervised FS considers unlabelled data, it could be difficult to accurately assess the relevance of features. Particularly, when considering social media texts, it is common to have an enormous volume of high-dimensional data, but a small volume of labelled instances. In consequence, FS techniques have to be carefully designed to cope with these characteristics. Feature selection techniques can be classified according to the setting in which they are applied in batch or online techniques.

Batch feature selection techniques assume the existence of a fixed set of instances, and therefore a feature space fully known in advance. However, in real-world applications, such assumptions might not hold as training examples could sequentially arrive, features might incrementally appear or it could be difficult to collect a full training set (Wang et al. 2014). For example, in the context of spam detection tasks, e-mails usually arrive sequentially, hindering the deployment of efficient and scalable batch FS techniques. Another example is the classification of newly arriving social posts, which could be used for event or trending topic detection, among other possibilities. Furthermore, batch approaches often require the

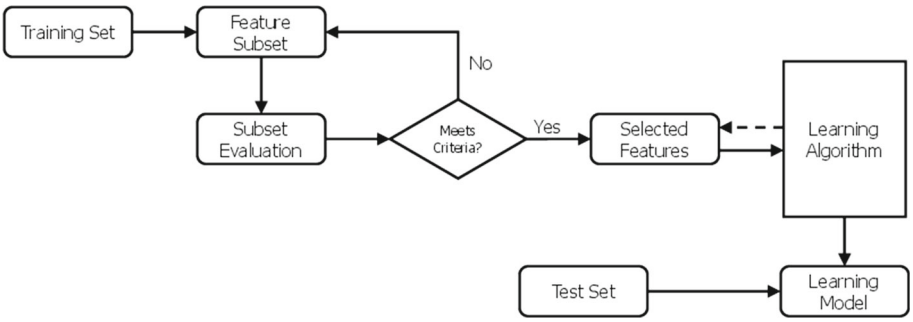

Fig. 1 A general framework for batch feature selection

entire training dataset to be loaded into memory, which is non-scalable and impractical for real-world applications involving large-scale datasets that exceed memory capacity. As a result, traditional batch FS techniques are not suited for emerging big data applications.

Online feature selection (OFS) techniques, on the contrary, assume that instances and their corresponding features arrive in a continuous stream. OFS techniques involve choosing a subset of features and the corresponding learning model at different time frames. At each moment, it is possible to not only select the most recently arrived features, but also remove already selected features, or even include previously rejected ones (Perkins and Theiler 2003). Consequently, OFS is particularly important in real-world systems in which traditional batch FS techniques cannot be directly applied (Wang et al. 2014).

3 Batch feature selection

The standard batch 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. In general, standard FS techniques process features individually, assuming that they are independent and identically distributed. Considering some specific applications where the feature space comes with prior knowledge of group structures, some standard methods have been developed accordingly.

The batch FS process can be divided into four subtasks, as Fig. 1 shows:

1. A search process is performed in which a feature subset is chosen as a candidate subset for representing the resource. The search might start with an empty set to which features are added, with a full set from which features are successively removed, or with a randomly selected subset.
2. Each new feature subset is evaluated and compared to the previous best one according to an evaluation criterion. Evaluation criteria can be either dependent or independent from learning algorithms.
3. The criteria determinate whether the process of feature selection should be stopped.
4. Once the stop criterion is met, the subset that best fits such criterion is validated. Then, the selected features are used in combination with a learning algorithm to build the learning model that will be used to classify or cluster new instances.

3.1 Based on individual features

These techniques involve ranking features according to a statistical score and then selecting the subset that achieved the highest scores, enriching the feature space with semantic information, or even replacing the existent features with a new feature set derived from the original ones.

3.1.1 Ranking techniques

Ranking techniques score each feature according to a particular metric and then select the best k features. Commonly-used ranking metrics include (Forman 2003): Term Frequency (*TF*), Document Frequency (*DF*), Inverse Document Frequency (*IDF*), Term Strength (*TS*), Information Gain (*IG*), Mutual Information (*MI*), Chi-Square (χ^2) and Odds Ratio (*OR*). Several works (Rosa and Ellen 2009; Saif et al. 2014) have evaluated the performance of such metrics in the context of short-texts. Rosa and Ellen (2009) experimentally evaluated the performance of text classification techniques over short-text military chat. Four traditional FS techniques were considered: DF, χ^2, IG and *MI*. The selected dataset comprised a medium-scale number of US military chat lines, each containing a text message and the timestamp. Noise was introduced to the dataset by including unrelated chat messages, and by altering irrelevant chat messages to include terms appearing in the relevant categories. According to the authors, DF obtained better results than χ^2 in all cases.

Saif et al. (2014) studied the effect of dynamically removing stopwords for sentiment classification of tweets by applying traditional FS methods. The authors considered five techniques: *TF*, *TF1* (only considers words appearing more than once), *IDF*, Term-Based Random Sampling, and *MI*. Experimental evaluation was based on five small-scale *Twitter* datasets, belonging to different domains. Two classifiers were selected for the task: Maximum Entropy and Naïve Bayes (*NB*). FS techniques were compared with the performance achieved when no stopwords were removed. The best classification results were achieved with *TF1* and *MI*.

Even though ranking techniques are widely used in text learning tasks, they present some limitations (Peng et al. 2009). First, it might be difficult to define the optimal number of features to select. Second, as the metrics evaluate the features individually, implicit relations might go undetected, and thus redundant features could be selected. Third, as most features in short-texts are unlikely to repeat, they present a long-tail distribution. As a result, the selected feature set might not be optimal due to the difficulty of detecting noisy features with frequencies at the end of the tail. Furthermore, supervised techniques applied to multi-class environments are susceptible to class distribution as they can be misled by strongly predictive features in the most populated classes, thereby preventing the selection of useful features appearing in the least populated classes. This problem worsens in the context of social-media data as, not only it might be difficult to collect a complete labelled set, but also most features have extremely low frequencies.

3.1.2 Enrichment techniques

Considering texts' shortness, the sparsity of the feature space and the low term frequencies, the traditional bag-of-words (BOW) representation might not be the most appropriate model for managing and analysing short-texts (Chen 2011; Rafeeque and Sendhilkumar 2011), as it may not preserve the semantic meaning of the original texts. Furthermore, in social-media,

abbreviations are widely used and new terms are being incessantly created, exacerbating the problems of synonymy and polysemy. One possible solution for handling sparsity is to expand short-texts by appending new features based on semantic information extracted from Web searches, lexical databases or provided by machine translations. Techniques based on expanding short-texts by using Web searches might have efficiency problems when analysing a high number of short-texts. Additionally, their performance strongly depends on the quality of the search engine. To address these problems, explicit concept taxonomies or implicit topics are used to enrich short-text representations (Chen 2011). Lexical databases such as *WordNet* or web-site directories such as the *Open Directory Project* (ODP),[2] have rich predefined taxonomies and human labellers who assign Web pages to each node in the taxonomy. Such corpora and other similar pre-defined taxonomies have been extensively used for enriching short-text representations (Wang et al. 2012; Liu et al. 2010). The articles presented in this section are organised according to the source of the enrichment, i.e. topic modelling, external data resources or other knowledge sources.

Topic modelling

Wang et al. (2012) proposed to combine domain knowledge, provided by Web pages, with statistical methods to alleviate the sparseness of labelled short-texts. The approach extracted topics from the domain knowledge by means of the Latent Dirichlet Allocation (LDA) model based on Gibbs sampling, and selected their most probable terms. Then, given a category and a term set, the category contribution and the *IG* score were computed to filter terms belonging to different topics. Although the results seemed promising, the actual precision improvements were insignificant (lower than a 2%), regarding the simple BOW approach and its modifications. Thus, further evaluations are needed to confirm the actual benefits of applying the approach to large-scale datasets.

Saif et al. (2012) proposed an approach for alleviating tweets' sparseness for performing sentiment analysis by considering semantically hidden concepts, and latent topics in combination with the sentiment topic of tweets. Alchemy API[3] was used for extracting tweets' semantics. Three alternatives were analysed for adding the semantics into the tweet representation. First, terms were replaced with their corresponding concepts (semantic replacement). Second, the original tweets were augmented with the obtained concepts (semantic augmentation). Third, concepts were included in the *NB* classifier by means of a new smoothing function that interpolated the original unigram features with the obtained concepts. The set of semantic-topic features was obtained by using the joint sentiment topic (JST) model (Lin and He 2009). JST simultaneously detects both sentiment and topics from text. First, a sentiment label is assigned to the text under analysis. Second, a topic associated to the chosen label is selected. Finally, words conditioned by both the sentiment and the topic are selected. The technique did not relied on labelled documents for training, instead it required the polarity information of words, which was extracted from the MPQA subjectivity lexicon.[4] In other words, JST clustered different words sharing similar sentiment and topics. This clustering could help to reduce the sparseness of *Twitter* data.

Experimental evaluation was based on the Stanford *Twitter* Sentiment Dataset.[5] The performance of the proposed sets of features was compared to that of the BOW approach. When

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

[3] http://www.alchemyapi.com/.

[4] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.

[5] http://help.sentiment140.com/.

considering the semantically hidden concepts, results showed that semantic replacement reduced both the size of the dataset, and the classification accuracy regarding the baseline. The authors hypothesised that the performed semantic replacement caused a loss of information, which in turn, affected performance. Furthermore, augmenting the feature space with semantic information achieved better results than replacing concepts. However, results were also lower than the baseline. On the other hand, adding the semantic information to the classifier instead of the tweets' representation helped to improve the baseline. When considering latent topics and their associated semantics, results were higher than the baseline, but lower than when considering the hidden topics. The authors claimed that semantic topics are preferable over semantic features as their results with fewer selected features are comparatively better than those achieved with semantic features.

External data resources

Liu et al. (2010) presented an approach based on part-of-speech (POS) tagging and *HowNet*[6] for semantically enriching short-texts. Short-text representation was built by replacing the original features (only nouns and verbs) with the most semantically related concepts retrieved from *HowNet*. The approach was reported to decrease its performance as the number of added concepts for each original term increased. This could mean that, as stated in (Gabrilovich and Markovitch 2006), semantic enrichment based on lexical databases does not necessarily improve classification results for short-texts. Additionally, the computational complexity involved in enriching features could not be worthy as simpler approaches have obtained similar results. Although these results might seem contradictory with the results reported by the other techniques in this section, it is important to highlight that as the different techniques were evaluated in diverse settings and considering different datasets, their results are not comparable. Hence, the conclusions stated for each of the presented techniques might not be generalisable to all techniques.

One possible drawback of using pre-defined taxonomies is their lack of up-to-date descriptions and coverage. For example, hierarchical taxonomies, such as ODP, define only one relation between nodes, ignoring other relations such as relatedness, chronology, homonyms and meronyms, among others (Gabrilovich and Markovitch 2006). Though large text corpus can be collected, the resulting taxonomies might not be appropriate for certain learning tasks in highly dynamic domains such as social media. Furthermore, pre-defined taxonomies might not be available for some languages (Chen 2011). Hence, it is not possible to enrich new terms or expressions, or consider social variations of standard language (slang), as well as abbreviations or acronyms. To overcome this problem, several approaches (Ferragina and Scaiella 2012; Tang et al. 2014b; Perez-Tellez et al. 2010) proposed using *Wikipedia* for topic extraction as an alternative to static taxonomies such as *WordNet* or ODP. In all cases, the authors stated the importance of using encyclopaedic knowledge for enriching short-texts, which have been proved to be inadequately represented by the BOW model.

Ferragina and Scaiella (2012) proposed the tool *TagMe*[7] to semantically tag short-texts with *Wikipedia* pages. Each input text was enriched with the semantically closest *Wikipedia* anchors and pages. Similarity between pages was computed by the Mendelyan method (Medelyan et al. 2009), which considers the intersection between the incoming links of two pages, and weights the similarity according to the commonness of each anchor. Then, annotations are assigned a score according to the probability of a certain term to appear in

[6] www.keenage.com/html/e_index.html.

[7] http://tagme.di.unipi.it/.

Wikipedia as an anchor, and the similarity value computed in the previous phase. The approach was able to improve results achieved with similar techniques developed for long-texts.

Tang et al. (2014b) also proposed to enrich short-texts (particularly tweets) with concepts extracted from *Wikipedia*. Additionally, the authors presented strategies for alleviating the sparse data problem, and solving both the polysemy and synonymy problems. Each tweet was pre-processed to select the noun and verb phrases to be enriched. Then, each phrase was assigned a unique *Wikipedia* concept (addressing polysemy). Equivalent concepts were grouped using *Wikipedia*'s redirection links (addressing synonymia). To alleviate the sparse data problem, the authors proposed three alternative representations: adding all the out-links in the article belonging to the target concept, the out-links in the article belonging to the target concept in the same semantic category as the target article, and the out-links found in the first sentence of the corresponding article (i.e. the definition) of the target concept. Representations were weighted by means of *TF–IDF*, and their similarity was assessed by the cosine similarity. Experimental evaluation was based on a low-scale number of tweets containing one out of six pre-defined hashtags. *TreeTagger* (Schmid 1994) was used for POS tagging tweets. Only nouns and verbs were selected. Three clustering algorithms were used: Hierarchical Agglomerative Clustering (HAC), Bisecting K-Means, and a graph based algorithm. Results showed that considering both concept disambiguation and synonymy redirection improved clustering based only on the original tweets' features. As regards tweet representation strategies, the worst results were achieved when considering all the out-links belonging to the target concept, as unrelated semantic information and noise were added. On the contrary, the best results were achieved when considering only the out-links belonging to the definition of the target concept. Finally, as regards the clustering algorithms, K-Means achieved the best results, whereas the HAC achieved the worst ones. Additionally, the *Wikipedia* based approach was able to outperform a traditional LDA approach. The authors concluded that *Wikipedia* redirections were useful for resolving synonyms as concept definitions were sufficient for solving the polysemy problem.

Perez-Tellez et al. (2010) presented different semantic enrichment strategies for tweets aiming at improving clustering performance. Particularly, the authors aimed at clustering tweets corresponding to companies having ambiguous names. For that purpose, they proposed four methodologies for semantic enrichment:

- *Self-Term Expansion Methodology* (*S-TEM*) Terms appearing in tweets were replaced with a set of co-related terms found by means of Pointwise Mutual Information (PMI), which computes the degree of relationship between two features.
- *Term Expansion Methodology-Wiki* (*TEM-Wiki*) Tweets were enriched not only with the list of co-occurrence terms extracted from the same corpus, but also with the information about the company provided by *Wikipedia*.
- *Term Expansion Methodology with Positive examples - Wiki* (*TEM-Positive-Wiki*) Only those tweets actually referring to a company were enriched with information extracted from *Wikipedia*. The authors acknowledged the limitations of this strategy related to the need of positive examples, which could be difficult to obtain.
- *Full Term Expansion Methodology* (*TEM-Full*) Only the company name was enriched with all the terms that co-occur with it in the same class of tweets, without imposing restrictions on the degree of the relationship. The authors considered that adding all the information that co-occur in tweets belonging to a company was always beneficial for the enriching process.

In all cases, after feature enrichment, only the most relevant features according to their *DF* score were selected. Experimental evaluation was based on a small number of companies

whose tweets were written in English. The companies were divided into two groups. The first one comprised companies with generic names, i.e. names that were expected to be ambiguous as they have a separated meaning in English and thus appear in the dictionary. On the contrary, the second group comprised companies whose names are not considered ambiguous, i.e. the names can be used only in a limited number of context and do not appear in the dictionary. K-means was the selected clustering algorithm. The baseline was the *BOW* tweet representation. Results showed that *TEM-Full* achieved the best results for the group of generic company names. On the other hand, *TEM-Wiki* achieved the best results for the group of specific company names. Although the authors supposed that clustering the second group of companies would be a simpler task, not every methodology improved its results when clustering such group. Interestingly, the methodologies that achieved the best results are opposites in the sense of whether considering external enrichment or only information available in the dataset. The authors concluded that the proposed enrichment techniques were effective for tweet clustering, as they were able to outperform the baseline representation in most cases.

In addition, there are approaches that combined both taxonomies and encyclopaedias for enriching short-texts. For example, Hu et al. (2009) combined *WordNet* and *Wikipedia* knowledge. In this case, short-texts were enriched with titles and links extracted from selected *Wikipedia* articles and *WordNet* synsets. To avoid the negative impact of the huge number of features obtained through the semantic enrichment, the authors ranked all features (original and the new concepts) according to their TF-IDF score to select the top-ranked ones. According to the authors, results proved the benefits of integrating semantic information from multiple sources.

Other knowledge sources

Although the presented approaches were reported to outperform simple BOW representations, Jin et al. (2011) stated that in highly dynamic domains as social media, where novel topics and trends constantly emerge, it is not always possible to naïvely find strongly related long-texts, such as Web snippets or *Wikipedia* articles. The implicit assumption that the auxiliary data is semantically related to the input short-texts is hardly true in practice due to the noisy nature of the long-text enrichment task. Furthermore, semantically enriching short-texts might incur in prohibitively increasing computational complexity in both time and memory space. There are additional computational times involved not only in actually enriching the representations (due to the sophisticated natural language processing analysis involved), but also in periodically updating the semantic databases. In this scenario, it might not be worthy to apply the presented semantic enrichment techniques based on real-world knowledge in real-time environments. Consequently, other approaches have intended to semantically enrich representations by adding machine translations.

Tang et al. (2012) proposed to represent short-texts by adding machine translations to the BOW representation. According to the authors, term translations help to: alleviate synonymity as multiple words that are synonyms in one language may be translated into a unique term in another language, involve word sense disambiguation in the translation process as it is based on context information, and deal with abbreviations and new words. However, results showed that integrating multiple languages had a negative impact on categorisation performance as the feature space was expanded. Furthermore, the performance depended on the quality of translations.

3.1.3 Replacement of features

Unlike the previously presented techniques that aim at refining the available sets of features by selecting the optimal subset that reduces noise and ambiguity, the techniques presented in this section aim at transforming the original feature set (i.e. terms appearing in short-texts) into a new feature set by inferring or creating new and distinct features. Such transformation could include the computation of statistics based on lexical or syntactical characteristics of data (for example, number of capitalised words, number of hashtags used or number of adjectives used), the definition of specific features according to the task to be performed (for example, features regarding the writing style), of inferring new features by embedding methods. These techniques might improve data representation as they can help to find hidden relationships between features. The articles presented in this Section are organised according whether they consider features that can be extracted directly from the text or they aim at inferring underlying patterns from the input texts.

Categorical features

Verma et al. (2011) proposed to represent texts based on a combination of hand-annotated and automatically extracted linguistic features aiming at automatically identifying tweets that contribute to situational awareness during mass emergencies. According to a data preliminary analysis, the authors claimed that those tweets that contribute to situational awareness are likely to be written in a particular objective, impersonal and formal style. Hence, the identification of subjectivity, personal style and register (formal/informal) could provide useful features for the classification task. The set of defined features included:

- *Lexical features* Uni-grams and Bi-grams, and their raw frequency as well as POS tags.
- *Tweet Subjectivity* (whether tweets are written in an objective or subjective style). A preliminary analysis showed a correlation between tweets written in an objective style and those that contribute to situational awareness (e.g. the location of hazards or the state of recovery efforts). On the other hand, tweets written in a subjective style were not correlated to situational awareness tweets.
- *Tweets Register* (whether tweets are written in a formal or informal style). Formal tweets are grammatically coherent and express complete thoughts. On the contrary, informal tweets tend to be fragmented, to lack of content, to include slang and to have grammatical mistakes. Although the higher correlations to situational awareness tweets were found for tweets written in a formal style, informal tweets were also found to contribute to situational awareness.
- *Tweets Tone* (whether tweets are written in a personal or impersonal style). Impersonal tweets show a sense of emotional distance from the event by the author. Tweets that convey personal characteristics are generally subjective. However, tweets might convey subjective content without being written from a personal point of view. Moreover, a tweet might be written in a personal style without conveying subjective content.

The values of the last three types of features were determined in two different ways: manually and predicted by a classifier. Furthermore, the subjectivity value was also determined by means of the *OpinionFinder*.[8] Experimental evaluation was based on several low-scale datasets comprising tweets related to four mass emergency events. Two classifiers were selected for the task: NB and Maximum Entropy. The baseline was the BOW

[8] http://mpqa.cs.pitt.edu/opinionfinder/.

tweet representation. Finally, to determine the stability and suitability of the approach, each classifier trained with tweets belonging to a specific event was tested with data belonging to a different event. Lexical features were paired with every other type of features, and then all features were combined together. The Maximum Entropy classifier consistently outperformed *NB* for every dataset and combination of features tested. The baseline results were improved for all but one dataset. Interestingly, the combination of features that achieved the best results varied according to the considered dataset. The authors concluded that their feature set, which was based on low-level linguistic features, could improve the performance of traditional lexical features without increasing the computational complexity. However, the approach was not completely automatised as it considered manually annotated features, hindering its usefulness in real-time systems.

Ma et al. (2013) aimed at predicting the popularity of *Twitter* hashtags by identifying and evaluating the effectiveness of content and contextual features. Content features were lexically derived from both hashtags and the set of tweets annotated with those hashtags. Examples of such features are: whether the hashtag contains digits, number of manually segmented words contained in a hashtag, fraction of tweets containing URLs, fraction of neutral, positive and negative sentiment tweets, 20 topics inferred from the tweets containing each hashtag, topical cohesiveness of tweets containing a hashtag, and topically cohesiveness of the segmented words. Contextual features were specific to a particular time interval, and were extracted from the social graph formed by users who have used the hashtag, and users who have not used it but are related to users who have. Examples of such features are: number of users who have used the hashtag, number of tweets containing the hashtag, retweets of tweets containing the hashtag, replies to tweets containing the hashtag, influential level of the users using the hashtag, strength of ties among users, ratio of the number of edges and possible edges in the user graph (graph density), overall degree of user interaction (average edge strength), ratio of the number of disconnected components and users, and number of users who have not used the hashtag but are connected to al least one user who have.

Experimental evaluation was based on more than 31 million tweets belonging to more than 2 million Singaporean users published during January and August 2011. Approximately only the 9% of tweets contained hashtags. The performance of five traditional classification algorithms (NB, k-NN, decision trees, SVM and logistic regression) using the defined features was compared to three baselines (Random, Lazy, PriorDist) ignoring content and lexical features. The Random baseline randomly predicted the popularity of a hashtag. The Lazy baseline predicted the popularity of a hashtag as its popularity in the previous time interval. The PriorDist baseline randomly predicted the popularity of a hashtag given a prior probability distribution of five popularity ranges. Results showed that using the generated features significantly outperformed the three baseline methods. The logistic regression classifier achieved the best results with respect to the Random and PriorDist baselines. SVM was the second best performing classifier. Contextual features were more effective than content features for the classification task. This could imply that community properties played a dominant role in the information diffusion process. However, the best results were achieved when combining both types of features. The most effective contextual features were user count, number of users who will potentially adopt a hashtag, and number of tweets containing a hashtag. On the other hand, the most effective content feature was hashtag clarity.

Embedded features

In recent years, several approaches (Amir et al. 2014; Tang et al. 2014a; Severyn and Moschitti 2015) have leveraged on neural networks and deep learning techniques for finding a feature set to represent tweets in the context of sentiment analysis. In the field of text mining and natural language processing, deep learning is used for deriving word-vector representations, part-of-speech tagging, semantic role labelling and named entity recognition, among others. Deep learning presents an advantage over existing approaches for sentiment analysis (Severyn and Moschitti 2015) as they can automate the feature construction and replacement phase to learn more general representations, instead of relying on extensive feature replacement and building techniques. Hence, such approaches are more flexible and stable when applied in constantly changing environments, such as social media data. Moreover, methods such as (Mikolov et al. 2013; Pennington et al. 2014) allow learning instance representations that capture fine-grained semantic and syntactic regularities. Interestingly, although widely used in tasks including short and noisy social texts, such models were created and evaluated using Google's and Reuters' news articles. One possible drawback of these techniques is their heavy dependence on matrix arithmetic operations between high dimensional matrices, which might hinder their applicability on real-time applications where tweets arrive in a continuous stream.

The quality of features obtained by general purpose deep learning embedding models (Mikolov et al. 2013) for the sentiment analysis task was assessed by Amir et al. (2014). Particularly, the authors compared the performance of embedded features to that of traditional word features (unigrams, Brown word clusters, euclidean distance between the word and class vector), lexicon features (vector representations were enriched with features considering the presence of words with known polarity such as happy or sad, and features considering the sum of sentiment scores of all uni-grams and bi-grams) and syntactic features (the use of punctuation, emoticons and character repetitions, and the number of capitalised words, among others). Experimental evaluation was based on using L_2-regularised logistic regression on the *SemEval2014* [9] tweet dataset. According to the authors, the best results were obtained when considering the combination of embedded features with the Brown word clusters. On the contrary, including uni-grams or syntactic features decreased the performance of classification. These results showed the potential of using general purpose embedded features for sentiment analysis tasks.

On the other hand, Tang et al. (2014a) and, Severyn and Moschitti (2015) tuned deep convolutional neural networks to consider sentiment-specific word embeddings. Tang et al. (2014a) developed three neural networks to include sentiment polarity in the loss functions. The first and second networks aimed at predicting sentiment distribution based on input n-grams. The distributed representation of the higher layer of the neural network was interpreted as the features describing the input. The third neural network not only considered the sentiment polarity of the input, but also the syntactic context of terms. Each neural network was trained with a set of tweets, which were labelled according to whether they contained positive or negative emoticons. The training process considered the derivative of the loss through back propagation. Then, the continuous representation of words and phrases was used as the tweet's features. Experimental evaluation based on the *SemEval2013* [10] dataset showed that the presented approach was able to outperform state-of-the-art techniques. The best results were obtained with the neural network combining both the polarity and the syntactical con-

[9] http://alt.qcri.org/semeval2014/task9/.

[10] https://www.cs.york.ac.uk/semeval-2013/task2/.

text of terms. Results also showed that general purpose embedded models (Mikolov et al. 2013; Pennington et al. 2014) were not effective enough for sentiment analysis in *Twitter*, as they only model the context information of words being unable of distinguishing words with similar contexts but opposite sentiment polarities. Then, when such word embeddings are used as the feature set, the discriminative ability of sentiment words is weakened, affecting classification performance.

Severyn and Moschitti (2015) built a deep convolutional neural network on top of general purpose word embedding models (Mikolov et al. 2013; Pennington et al. 2014) and the approach in (Tang et al. 2014a). Interestingly, the authors updated the pre-trained general purpose models by backprogration during the training phase with *Twitter* data to adapt them for the sentiment analysis task. Experimental evaluation showed the effectiveness of considering pre-trained word vectors and the advantages of leveraging on *Twitter* corpora, as updating the general models with specific *Twitter* data improved the results of only considering general purpose embeddings. According to the authors, considering the models defined in (Tang et al. 2014a) achieved worse results than the general purpose models. The authors hypothesised that such performance difference could be due to the diverse types of corpora used, and the size of the training dataset.

Unlike previous works, Jiang et al. (2011) decided not to consider the tweets syntax for extracting their feature claiming that the low parsing accuracy limits the performance of sentiment analysis tasks. Instead, they split tweets into a left and right context given a target term, and used distributed word representations and neural pooling functions to build the feature set. The method consists of five stages. First, tweets' terms were represented using the word embeddings in (Mikolov et al. 2013; Pennington et al. 2014). Second, the left and right contexts of a given term were extracted. The sentiment towards the given term resulted from the interaction of both contexts. Third, the contexts were extended using similarity methods by keeping or filtering words according to whether they appear on pre-defined sentiment lexicons. Fourth, pooling functions were used to automatically extract the features for representing the tweets. The pooling functions combined the features obtained for each context, extended the feature set with features corresponding to the full text, or replaced the original terms with statistics of their appearances. Finally, the resulting features were used as input for the sentiment classification task. Experimental evaluation was based on manually annotated tweets belonging to given topics (Dong et al. 2014), such as "bill gates", "google" or "xbox". Results showed that the approach obtained better results than state-of-the-art methods using syntax, including the one in (Tang et al. 2014a). According to the authors, their approach solves the potential limitation of syntax-based methods by avoiding the influence of noise. Moreover, classification performance was demonstrated to improve when considering multiple embeddings and pooling functions.

3.2 Based on groups of features

The techniques presented in the previous Section assume that data is independent and identically distributed. This assumption might be useful for long-texts, but short-texts present a different situation. Unlike in long-texts, most terms appearing in short-texts have low frequencies, thus, individual frequencies might not correctly assess term relevance. As a result, it is necessary to develop techniques that analyse groups of terms. The techniques reviewed in this Section are organised according to whether they leverage on information extracted from the linked nature of social media data or only include textual and metadata information.

3.2.1 Based on social network information

In real-world settings, data can be distributed in the form of networks or graphs. In the context of social media data, for each social post is available not only its content, but also the information regarding its authors and their social network. Thus, linked data differs from traditional attribute value data. Particularly, linked data is not independent and identically distributed, which is one of the most recurring assumptions of traditional text learning techniques (Alelyani et al. 2013), as exposed in the previous Section. As a result, linked data in social media presents new opportunities for developing novel FS techniques over linked data. In the last years, approaches that consider both the content of posts and the social information of their authors have been developed (Tang and Liu 2012, 2014b; Liu and Yu 2005). Interestingly, all techniques are based on computing arithmetic operations between high-dimensional matrices, which could increase the computational complexity of the approaches, hindering their applicability in real-time environments.

Tang and Liu (2012, 2014b) presented both supervised and unsupervised FS techniques based on links between the posts' authors. Tang and Liu (2012) suggested four types of relations between users based on social correlation theories such as homophily (McPherson et al. 2001) and social influence (Marsden and Friedkin 1993): *Co-Post* (posts by the same user comprise similar topics, i.e. posts of a user are more likely to have similar topics than randomly selected posts), *Co-Following* (if two users follow the same user, their posts are likely to have related topics), *Co-Followed* (if two users are followed by the same user, their posts are topically similar) and *Following* (a user follows another if they share interests, hence, their posts are more likely to have similar topics). Based on such relations, the authors defined *LinkedFS*, a supervised FS technique combining both content and social relations. Each social relation was formulated as an optimisation problem including spectral analysis and solving the minimisation problem derived from the $\ell_{2,1}$-norm. Interestingly, the technique showed to be more effective when small datasets were considered. This finding is important as it is difficult to obtain labelled social media data. Similarly, Tang and Liu (2014b) presented an unsupervised variation of the technique based on the definition of pseudo-class labels. Particularly, graph and social dimension regularisation were analysed for capturing the dependency among linked instances, and thus defining the pseudo-class label of each instance. Then, pseudo-class labels were used for finding the content information in a supervised manner by means of a spectral discriminative analysis. The approach has a quadratic computational complexity on the number of instances and features. Even though the authors defined the technique to be applied to unlabelled datasets by introducing the concept of pseudo-labels, experimental evaluation was performed only over labelled datasets. In both cases, the approaches rely on the definition of several parameters, which might be difficult to tune in dynamically changing environments. The authors did not provide means to automatically set parameters according to data characteristics.

In the same line of research, Gu and Han (2011) proposed a supervised FS technique based on Laplacian Regularised Least Squares (LapRLS) for networked data, which aimed at selecting the feature subset minimising the LapRLS error. The technique did not require to explicitly define the number of features to select, as such number was implicitly controlled by regularisation parameters. The approach used LapRLS to analyse the content information, and then adopted graph regularisation based on spectral graph theory to analyse link information. Alike the techniques presented in (Tang and Liu 2012, 2014b), graph regularisation was based on the basic assumption that if two nodes are linked in a network, they are likely to be topically related, and thus to have the same label. As the resulting optimisation

problem is a mixed integer programming problem that might be difficult to solve, the authors relaxed the problem into a $\ell_{2,1}$-norm constrained LapRLS problem, which was solved by an accelerated proximal gradient descent. Experimental evaluation compared the proposed approaches with four baselines: regularised least squares, LapRLS (a special case of the the approach considering all features), Fisher Score, and the probabilistic relation principle component analysis, for two datasets. Results showed that the proposed approach outperformed the baselines for most of the combinations tested. Particularly, it failed to outperform the regularised least squares method for one of the datasets. Alike the previous case, the technique relies on manually defined parameters.

The social relations considered by these techniques do not take into account the possibility that each type of link can lead to the formation of ties with different strength. Hence, further studies that consider the strength relevance of the different types of relations are needed in order to continue improving results. For example, studies could focus on the exploration of additional relevant information in social networks, or in measuring the strength of social relations by means of community detection techniques.

3.2.2 Based on textual and metadata information

Both Moradi and Rostami (2015) and Alexandrov et al. (2005) integrated the concept of graph clustering with node centrality and similarity metrics for performing FS. Moradi and Rostami (2015) presented an approach for unsupervised FS in three steps. First, a graph was built in which nodes represented features, and edges were weighted based on their similarity. As different similarity metrics might lead to different performances on graph-based FS, such metric has to be carefully selected. The authors chose the Pearson product-moment correlation coefficient instead of the traditional Euclidean distance. Second, features were clusterised aiming at grouping highly correlated features into the same cluster. Third, the most relevant and influential features of each cluster were selected by the Laplacian centrality. According to the authors, the technique is computationally efficient for high-dimensional datasets. However, it has three adjustable parameters that were tuned after performing preliminary runs on the training data, which might not be possible on OFS settings. Additionally, as the technique was evaluated for categorical data, its findings might not be applicable in the context of short-texts. Moreover, the authors evaluated the approaches considering two distinct training and test sets. Once all training instances were analysed, classifiers were never updated with the information conveyed by test instances, which could indicate that the authors assumed that both training and test instances are always relatively similar. Such assumption might not hold in social media due to the highly dynamic emergence of new topics, trends and posts.

Alexandrov et al. (2005) proposed a method to filter and group terms in clusters to compensate the effect of low frequencies by considering each new group as a new coordinate in the index space equal to the sum of the occurrences of all keywords in a given group. In addition, Ozdikis et al. (2012) considered the syntagmatic and paradigmatic relations between terms. The syntagmatic relation relies on the co-occurrence of words. Two words are syntagmaticaly related if they co-occur in more than a pre-defined number of tweets, and none of those words co-occur in a greater number of tweets with another word. On the other hand, the paradigmatic relation aims at finding pairs of words that can be used interchangeability. Terms were represented by a vector comprising term co-occurrences in tweets. Two words were considered to convey a paradigmatic relation if their similarity was higher than a pre-defined threshold. The similarity among term vectors define the degree of contextual commonality, which

allows to determine if two terms can be deemed as synonyms. The cosine similarity and the Manhattan distance assessed such similarity. Experimental evaluation performed on Turkish tweets showed that only the enrichment based on paradigmatic relations and Manhattan similarity was able to improve baseline results. Consequently, further experimental evaluations are required for effectively assessing the potential of the approach to improve topic detection accuracy. Additionally, the approach needs to be evaluated for tweets belonging to different languages in order to effectively test its language independence.

Finally, besides social and content similarity relations, there could be several additional dimensions of information available for each data instance. For example, in the case of tweets, there is not only the text available, but also their associated hashtags. There are two straightforward strategies for applying techniques based on individual data to multi-dimensional feature spaces (Tang et al. 2013): concatenation (heterogeneous feature spaces are concatenated into one homogeneous feature space, i.e. all features are merged together), and separation (performing traditional FS on each feature space separately). The concatenation strategy ignores the differences among heterogeneous feature spaces, whilst the separation strategy considers each view independently. However, as the different feature spaces describe the same set of instances through different dimensions, the features spaces are inherently related. Consequently, techniques integrating the different data sources might achieve better performance than those considering each source independently. In this regard, (Tang et al. 2013; Fang et al. 2014) proposed techniques for leveraging different dimensions of information. Tang et al. (2013) proposed an unsupervised technique for simultaneously selecting features for all views by using spectral analysis to exploit the relations among them. The technique was not restricted to any particular kind of relation, and thus it could even consider the social ones. As in (Tang and Liu 2014b), the technique was based on defining pseudo-class labels to leverage the information from each view by means of spectral analysis. Then, relations among views were formulated as an optimisation problem. As it considered constraint vectors of zero norm mixed with integer programming, its solution was found by iteratively performing multiple arithmetic operations over high-dimensional matrices, which might negatively affect the performance of the approach.

Fang et al. (2014) proposed to combine three types of information dimensions among tweets: semantic, social tag and temporal. The semantic dimension was defined as the meaningful information provided by terms found across different tweets. A suffix tree was built to detect the common phrases between pairs of tweets. After building the trees, all the common phrases of each node were obtained by traversing from the root node to all the leaf nodes. The traditional vector space model was extended to assign extra weights to words in phrases detected by a suffix tree, i.e. the higher the number of words in the common phrase, the higher the weighting value that should be added to the word in it. When individually considered, this dimension might be inefficient due to the sparseness of the feature space. The social tag dimension was defined as the relation measured by hashtags, which could be regarded as a generalised description of the topics contained in tweets. The authors assumed that two tweets that do not share common meaningful words, but share hashtags have a high probability of belonging to the same topic. The temporal dimension was defined as the information provided by the posting time of tweets. Topics were assumed to be generated with a particular life cycle. For example, after some events such as natural disasters, related tweets are usually posted within a short time period. As tweets are not treated as a data stream, a Gaussian kernel function was applied to measure the temporal similarity of pairs of tweets. If several topics are generated in the same period, considering this dimension individually might not be useful. As each dimension has its own drawbacks when individually considered, the authors combined them by means of two techniques based on spectral clustering: Stage-based Mul-

tiview clustering (performs linear operations over the dimensions) and Co-training-based Multiview Clustering (projects back and forth from one relation to other iteratively). Experimental evaluation was based on a low-scale number of tweets containing at least one hashtag. Results showed that the Semantic relation was the most important individual relation. Finally, the authors stated that the results showed the superiority of the combination of relations over considering the individual relations, or any combination of two relations.

3.3 Summary

This Section reviewed several techniques for performing FS that assumed the existence of instances, and thus, a feature space fully known in advance. Techniques considered either the features individually or in groups. In turn, those techniques considering features individually were further classified according to whether they ranked, enriched or replaced features. Although ranking techniques have been widely used in the literature, the reviewed works exposed several limitations that can undermine their applicability in dynamic environments comprising short-texts. For example, the effect that low term frequency has on the ability of techniques to discover the truly important features. Also, most techniques require to predefine the number of features to select, which might be difficult to accurately estimate on dynamically changing environments such as social media data.

As regards techniques enriching the feature space, the reviewed techniques showed that enriching data by external sources may not be always feasible or suitable for real-time applications. For example, as the feature space is enriched, the "curse of dimensionality" is further aggravated. Additionally, the process of enrichment might introduce noise to the representation, causing the addition of semantically unrelated knowledge to the original text. Hence, a naïve combination of short-texts and semantically unrelated long-texts or topics might negatively affect learning performance. For unsupervised learning, the problem is even worse as there is no category information neither to guide the selection of auxiliary data nor to assess the quality of the enrichment. Finally, the computational complexity added by the semantic enrichment might not compensate the performance improvements. Replacement techniques obtained promising results whilst reducing the size of the feature space. Particularly, embedded techniques were shown to effectively capture implicit semantic and syntactic regularities in the textual data. Moreover, they showed potential for adapting to dynamically changing environments, at the expense of increasing the computational complexity of the task.

On the other hand, techniques considering groups of features were further classified according to whether they considered the social linked nature of social media data or only lexical and metadata relations. These techniques stated the benefits of considering groups of features to cope with the low frequency of terms, and to leverage on additional sources of information linking the different data instances. However, further studies are needed to accurately assess the importance of the different types of social relations for continuing to improve the performance of FS techniques. One possible limitation of the techniques involving social links is that they assume that link information is relatively stable. Thus, social links are never updated to reflect changes in the social network.

4 Online feature selection

The techniques introduced in the previous Section assume that all features and instances are known in advance. However, there are real-world applications in which either the full set of instances, features or both are unknown. In such applications, training examples could

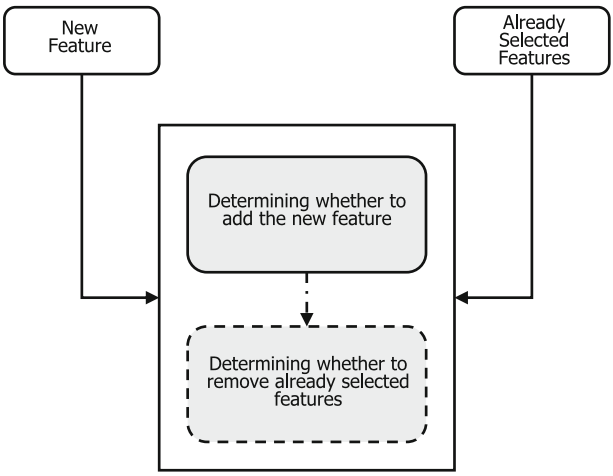

Fig. 2 A general framework for online feature selection

arrive sequentially, or it could be difficult to collect the full training set (Wang et al. 2014). For example, in the context of spam detection tasks, e-mails usually arrive sequentially, hindering the deployment of efficient and scalable batch FS techniques. Another example is the classification of newly arriving social posts, which could be used for event or trending topic detection, among other possibilities. In these situations, 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. At each moment, there is the possibility of not only selecting the most recently arrived features, but also of removing already selected features, or even including previously rejected ones (Perkins and Theiler 2003). Consequently, OFS techniques are particularly important in real-world systems in which traditional batch FS techniques cannot be directly applied (Wang et al. 2014). Although FS techniques have received considerable attention during the last decades, most studies are focused on developing batch techniques instead of facing the challenging problem of OFS. Efficient and scalable OFS becomes an important requirement in numerous large-scale social applications.

According to Perkins and Theiler (2003), OFS should be applied in two scenarios. In the first scenario, features are expensive to compute. Most learning approaches suppose that all features associated with the training data are known at the beginning of the learning process, ignoring the computational effort of computing those features. In the second scenario, the feature space is infinite. In both cases, a solution would include generating features, one at a time, and then selecting the best subset of generated features. As more features are known, the feature subset and its associated model will improve. When the model performance reaches a certain threshold, feature generation can be stopped. Interestingly, in both scenarios, the classification or clustering of instances only begins after the FS phase has ended, thus test instances are not considered for further updates of the model nor the selected features. This situation could result in low quality feature sets in highly dynamic environments in which training instances might not be representative of newly emerging ones.

Figure 2 depicts a general framework for OFS, which will typically perform the following steps:

1. A new feature arrives or it is generated.

2. It should be determined whether to add the new feature to the set of already selected features.
3. It should be determined whether to remove features from the set of already selected features. This step is optional. Some OFS algorithms choose to only implement Step 2.
4. Repeat Step 1 to Step 3.

One straightforward approach to OFS is to considerer the set of all discovered features at each time step, and then apply any traditional FS technique (Perkins and Theiler 2003). However, given that the size of the feature space can continuously increase, this approach might present scalability issues, thus resulting inefficient. For example, in online settings, filter approaches should recompute the statistics as new features are discovered, which might be time-consuming, even if new features are individually considered. The problem worsens if the quality of features is assessed by learning algorithms, as in wrapper methods. In an online setting, at each time-step, the feature set changes. In this context, evaluating the performance of the model considering each feature set would be inefficient and computationally complex. Ultimately, an OFS technique must allow performing efficient incremental updates. 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.

4.1 Based on individual features

Most OFS approaches, such as (Perkins and Theiler 2003; Wu et al. 2010; Wang et al. 2013), assume that features arrive sequentially and individually, whilst all training instances are known in advance, i.e. before the learning process starts. Perkins and Theiler (2003) proposed an approach based on gradient feature testing (i.e. grafting) for binary classification problems. Grafting involves defining weight vectors for each potential feature, ℓ_1-regularisation, additional un-regularised parameters, and an underlying model. The technique was based on the observation that adding newly discovered features into an existing model incurs in a regularised penalty. Features were only added if the reduction in the mean loss outweighed the penalty, i.e. features passed the gradient test. Otherwise, features were discarded. As the model was modified, a re-optimisation step was applied to all parameters by repeating the gradient tests. Weights adopting zero-values were removed. Model updates were efficiently performed as they only involved computing dot products, which imply a linear complexity on the number of elements over the computation is performed. However, the update time increased as the number of features to analyse grown. According to Perkins et al. (2003), the computational complexity of grafting is lower than the $\Theta(p^2)$ (where p represents the total number of weights in the weight vector) computational complexity of computing a full gradient descent, however it is still cubic on the number of weights to select for each feature. The authors concluded that grafting techniques provide an approach to OFS that combines the speed of filters with the accuracy of wrappers.

Besides grafting, α-investing is another technique traditionally used for stream FS (Zhou et al. 2006). The rationale behind α-investing is to adaptively control the threshold (α) for adding newly discovered features to the model. Such threshold corresponds to the probability of including a spurious feature, and it is adjusted using the acceptable number of insignificant features to select, i.e. false positives. Particularly, new features are added to the model when α is lower than the probability of the feature to be judged as insignificant when it is actually significant. Every time a new feature is not added to the model, α is decreased to avoid selecting more than a target fraction of spurious features. Similarly, every time a new feature

is added to the model, α is increased. Interestingly, adding new features does not trigger any further analysis of the already selected features. As a result, it cannot be guaranteed that no redundancy exists among features. The computational complexity of the approach is $\Theta\left(n^2 * f\right)$ where n represents the number of data instances and f the number of features. The quadratic cost corresponds to the statistics that the approach computes. However, such costs can be reduced if sampling strategies are applied. In such case, the cost would be $\Theta\left(n * s^2 * f\right)$, where s represents the size of the sample. Assuming that $s \ll n$, it follows that $n * s^2 \ll n^2$. Although this technique was reported to achieve better results than using a BOW representation in combination with SVM, neural networks and decision trees, it requires prior knowledge about the structure of the feature space to heuristically control the choice of candidate features (Wu et al. 2010). As it might be difficult or impossible to obtain such information from the feature stream, the technique might not be useful in truly online environments. More efforts would be needed to manage real-world streams in which the structure of features in unknown.

Wu et al. (2010) proposed a supervised two-step approach for selecting an optimal feature subset from an infinite feature stream, named Online Streaming Feature Selection (OSFS). The authors aimed at efficiently and effectively finding a feature subset based on feature relevance and redundancy, considering a setting in which candidate features are unknown as they are continuously generated. The first step analyses the relevance of new features regarding their class label. The second step analyses the redundancy of the selected features each time a relevant feature is found. Feature redundancy is assessed based on the conditional independence of the newly selected feature and each possible subset of already selected features of maximum size k. The steps are iteratively performed until a stopping criterion is satisfied. Both feature relevance and redundancy are analysed in terms of probabilistic conditional independence, although the analysis might fail in small datasets in which tests could be unreliable. To reduce the computational complexity of the redundancy analysis, the feature set was kept as minimal as possible. According to the authors, even though the technique was able to select all strongly relevant features, it could not select all non-redundant ones. The computational complexity of OSFS is $\Theta\left(|V|\,|BCF|\,k^{|BCF|}\right)$, where $|V|$ is the number of arrived features and $|BCF|$the number of selected features. Hence, the complexity depends on the number of conditional dependency tests to perform, which also represents the most time-consuming part of the technique. Aiming at reducing computational complexity, the authors presented a variation in which the redundancy analysis is divided into two phases. First, an inner-redundancy analysis that is performed every time a new relevant feature is found. This phase only re-examines the new relevant feature. Second, an outer-redundancy phase that is only performed when feature generation is stopped. This phase re-examines each relevant feature. The computational complexity of this variation is $\Theta\left((|V| + |BCF|)\,k^{|BCF|}\right)$.

Although the described approaches were reported to achieve promising results, their performance has only been assessed for long-text binary classification problems. Furthermore, all techniques still require to know all instances in advance, hindering its application in environments in which the set of instances is continually updated. Moreover, the grafting technique also requires all candidate features to be known in advance in order to set a parameter before starting the learning process. Thus, its application in online settings in which neither the full set of instances nor the set of features are known in advance is hindered. Finally, as features are assumed to be distinct from each other, and to arrive sequentially and individually, the techniques do not consider the possibility of having to analyse repeated features. In social media settings, features arrive grouped in instances, which can comprise

already known features or new unknown ones. In this setting, techniques might have to re-evaluate already known features, which could severely affect its computational time.

4.2 Based on groups of features

The approaches presented in the previous Section dynamically evaluate individual features as they arrive. However, all share the same limitation: they neglect the relationships between features, which could be highly important in certain environments. For example, when analysing short-texts it might be useful to consider both the text as a complete unit to compute statistics, and the social network of the authors of such short-texts. Some approaches designed for analysing newly arrived groups of features are described in this Section, divided according to whether they were designed for general purpose FS, i.e. applicable to multiple domains, or specifically designed for the short-text domain.

4.2.1 Multi domain techniques

Wang et al. (2013) proposed a variation of the approach in (Wu et al. 2010), named Online Group Feature Selection (OGFS), in which features are assumed to arrive in groups. The approach is divided into two steps: intra and inter-group FS. The intra-group step applies spectral analysis to dynamically select the most discriminative features of each newly arrived group. It requires performing arithmetic operations between matrices to compute a relevance score for each newly arrived feature and the subset of previously selected ones, resulting in a high computational complexity step. As traditional spectral FS approaches rely on global information, not available in OFS, the authors defined two criteria for selecting new features. The first criterion aims at only selecting features that maximise the inter-class distances by comparing the discriminative power of the already selected features with that of the already selected features plus the new feature. As it requires to compute at least eight matrix multiplications for each new feature, its computation might be excessively time-consuming. Furthermore, as the number of selected features increases, the criterion is supposed to be harder to satisfy. The second criterion aims at analysing the discriminative power of the individual features by performing a t-test to compute their statistical significance. A feature is selected if it satisfies any criterion. The inter-group step is performed to consider the group information disregarded by the intra-group step. It applies a sparse linear regression model of Least Absolute Shrinkage and Selection Operator (LASSO) to select a global optimal subset from the newly selected discriminative features and the previously selected feature set. LASSO involves solving an optimisation problem and the definition of several parameters for controlling the regularisation applied to the estimators. Both steps are iteratively performed until the stopping criteria is met. The authors defined three possible stopping criteria: 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 pre-defined threshold.

In the context of 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 OGFS 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 impacting on the performance of the technique. Although OGFS was reported to achieve promising results in UCI datasets [11] and images, it has not been applied to short-text classification. Finally, it is

[11] http://archive.ics.uci.edu/ml/.

interesting to highlight that even though the authors claimed that the intra-group step has a computational complexity of $\Theta(m)$ where m represents the number of selected features so far, the intra-group analysis includes computing several matrix multiplications, which results in a computational complexity of, at least, $\Omega\left(n^{2.807}\right)$ in the case of square matrices, depending on how multiplications are implemented (Strassen 1969). Furthermore, the approach requires continuous updates of several matrices, which could also affect the computational complexity.

Wang et al. (2014) proposed an Online Feature Selection with partial outputs technique (OFSp), which replicated the setting of real-world applications by considering a sequential arrival of instances. The authors proposed approaches for two OFS problems. In the first setting, all features belonging to each newly arrived training instance could be accessed by a linear learner. In the second setting, only a fixed number of features of each new training instance could be accessed by a linear learner, which selects the particular feature subset of each training instance. A straightforward approach for the first setting would be to modify the Perceptron algorithm by truncating the classifier to set everything but a pre-defined number of the highest absolute-valued features (B) to zero. However, that approach cannot guarantee that the values of the unselected features are sufficiently small, which could lead to classification mistakes. In this context, the authors proposed to maintain a linear classifier with at most B non-zero values. Each time a training instance is misclassified, the classifier is first updated by the online gradient descent and then projected to a L_2 ball to ensure that the classifier norm is bounded. Then, only the B non-zero elements with the highest absolute values are selected. Alike in the first setting, the selection of only the highest absolute-valued B features could lead to poor classification performance as the selected feature subset could never change. Instead, the authors proposed a greedy technique that randomly selects a feature subset, and then keeps only features with non-zero values in the resulting linear classifier. The algorithm first randomly chooses B attributes from the set of known attributes, and then chooses the B attributes for which the classifier obtained non-zero values.

As the technique was evaluated for long-texts in a binary-class setting, its results cannot be generalised for the case of short-text multi-class classification. Additionally, considering a multi-class setting increases the complexity of the binary approach from $\Theta(n * B * f)$ to $\Theta(n * m * B * f)$, where n corresponds to the number of instances to be classified, m the number of classes, B to the number of features to select, and f the number of features. The cost of selecting the B highest-valued features is $B * f$. In those cases in which $B \ll f$, the cost can be reduced to f. Also, in those cases in which $B \approx log(f)$, the cost would be $f * log(f)$, which is the cost of sorting an arbitrary structure. Finally, the evaluation considered two distinct training and test sets. Once all training instances were analysed, the linear classifier was never updated with the information belonging to test instances. This could indicate that the authors assumed that training examples are representative of testing examples. In social environments, this might not be the case due to the highly dynamic emergence of new topics, trends and posts.

4.2.2 Short-text oriented techniques

As regards OFS applied to short-text categorisation, most approaches available in the literature only rely 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. (2012) proposed a real-world event detection system (known as *Twevent*) using only non-overlapping segments of one or more consecutive words contained in tweets. Tweet segments contained in a large number of tweets were supposed to represent named entities or some semantically meaningful concept,

containing much more specific information than any individual uni-gram. Concepts reduce noise in the event detection process, facilitating the interpretation of the detected events. Given tweets published in a stream, *Twevent* first divides them into non-overlapping segments, and then bursty segments are identified within a fixed time window based on their frequency patterns, modelled as a Gaussian distribution. The tweet segmentation problem was formulated as an optimisation problem based on a stickiness function. The function was defined by considering the generalised Symmetric Conditional Probability for N-grams, which is defined to measure the "cohesiveness" of a segment by analysing all possible binary segmentations of the n-gram. A high stickiness score indicated that further splitting the segment would break the correct word collocation. A semantic component was added to the score to favour those segments frequently appearing as anchor texts in *Wikipedia*. Each bursty segment was described by the set of tweets containing such segment along with their timestamps, which were then used for computing the similarity among segments. Additionally, to detect events attracting a large number of users, the user frequency of each segment was computed. The usage of user frequency aimed at reducing the negative impact of Spam and Self-Promotion tweets, which do not convey real events. According to the authors, experimental evaluation showed that *Twevent* is an efficient, effective and scalable approach. The analysis of the computational complexity of the approach can be divided in three. First, tweet segmentation has a complexity linear to the number of words in tweets. Second, the event segment detection has a complexity linear to the number of segments on each window. Third, the tweet similarity computation has a computational complexity of $\Theta\left(n^2\right)$ where n represents the number of event segments. As the authors considered that the number of events to detect per time window is small, the time spent filtering the candidate events can be neglected. Finally, the authors proposed to study the effectiveness of adding more tweet features (e.g. retweet or hashtag rate), and to analyse the effectiveness of the approach when none of the segments is covered by the selected knowledge base.

Becker et al. (2011) aimed at differentiating between real-world events and non-event messages in a time-ordered stream of *Twitter* messages based on statistics of the expected volume of messages, user interactions, topical coherence of tweets' clusters and tag usage. The approach combined online clustering and filtering. First, tweets were clustered and then, the proposed features were computed for each cluster to reveal characteristics that might help to detect clusters that were effectively associated to events. Tweet clustering was the most computational complex part of the approach. The authors chose to use an incremental clustering algorithm, based on the cosine similarity. Thus, the complexity of the FS approach is $\Theta\left(n * m\right)$, where n represents the number of instances to be clustered and m the number of cluster centroids to which similarity needs to be computed. Features were broadly classified into four categories: temporal, social, topical and *Twitter*-centric features. Temporal features were based on the supposition that the volume of messages for a certain event increments during the event's associated time interval. Particularly, they considered the aggregated number of tweets containing each term per time interval (hours), and the total number of tweets posted per time interval. Other temporal features aimed at reflecting the degree to which the volume of messages containing a certain term exhibit an exponential growth in the hours leading up to the event. Social features aimed at capturing the interaction of users inside the tweet clusters. The authors assumed that the pattern of interactions (such as retweets, mentions and replies) between users might be different between events and other non-event messages. Topical features aimed at describing the topical coherence of a cluster based on the supposition that event clusters revolve around a central and unique topic, whereas non-event clusters do not. This type of features included the topical coherence of clusters (i.e. the average similarity of messages to the cluster centroid), the percentage of messages in the

cluster containing the N most frequent terms, and the percentage of the most frequent terms contained in at least the $K\%$ of cluster messages. *Twitter*-centric features aimed at helping to differentiate the *Twitter*-centric activities, which represent a particular class of non-event messages, from real-world events by computing statistics regarding tag usage. This type of features included whether tags comprised a concatenation of multiple words, and the percentage of cluster messages containing tags and the most frequently used tags. As clusters evolve constantly over time, features were updated accordingly. Although the approach was reported to improve baseline results, it suffered from misclassification. Particularly, clusters containing tweets belonging to broad topics posted following the same structure and using the same hashtags were classified as events, when they actually corresponded to spam. Additionally, the authors planned to extend the approach to improve prioritisation, ranking and filtering of the extracted content.

Zubiaga et al. (2015) proposed a set of features for real-time classification of trending topics into four categories: news (trending topics that are produced by a newsworthy event that the major news outlets had already reported by the time the trend appeared or will report soon after its appearance in *Twitter*), ongoing events (trending topics generated by communities of users who tweeted about an ongoing event as it unfolds), memes (trending topics triggered by viral ideas initiated by individuals or organisations who are popular enough to widely spread tweets) and commemoratives (trending topics referring to birthdays, anniversaries or memorial days, among other celebrations) by considering a small set of language-independent features. The approach assumed that social behavioural patterns in the information diffusion process depend on the type of trigger that originated the trend. As a result, a set of 15 features divided into two groups (average values and feature diversity) was defined. The first group computed the average value of retweet ratio, depth of the retweet chain, hashtags, length of tweets, number of tweets with exclamation or question marks, number of tweets containing URLs, topic repetition, number of replies and spread velocity. The second group computed the Shannon's index of diversity, which analyses the variation of features across the trending topic, i.e. measures the information entropy of the distribution of values for a feature. The higher the diversity of features, the more different are the features from tweet to tweet within a trending topic. In particular, such variation was computed for the number of users posting about the trending topic, number of users who were re-tweeted, and hashtags, languages used and terms contained in the trend. Finally, the authors highlighted four characteristics that made their approach suitable for real-time tweet classification: it requires a small feature set that can be straightforwardly computed, it does not make use of any source of external data, it can outperform the predictive power of content, it has low computational cost (which is linear to the number of tweets to be analysed), and the number of features remains unchanged as the number of instances and their content increases. However, features were specifically designed for the four described classes, which might imply that cannot be generalised to other domains.

Finally, Li et al. (2015) proposed an unsupervised FS technique for social media. Unlike their previous works (Tang and Liu 2012, 2014b), this new technique is supposed to be applicable to OFS tasks. The authors aimed at investigating how to leverage on social link information for performing OFS. The technique considers both social information and feature information. First, the social latent factors for each instance are obtained based on the mixed membership stochastic block model (Airoldi et al. 2008). As the social factors for each linked instance are obtained, they are used as a constraint to perform FS through a regression model. Then, the importance of each feature is measured as its ability to differentiate diverse social latent factors. According to the authors, if two users post similar tweets, they are more likely to share similar social latent factors, like hobbies or education. In this regard, the social

latent factors of two instances are more likely to be consistent when their feature similarity is high. Feature information is modelled by a graph representing the feature similarity between different data instances. Then, the problem of deciding 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 if the feature has to be accepted. If the feature is accepted, the model is re-optimised. When new features are added, there is also the possibility of removing already selected features. Experimental evaluation was based on two real-world social media datasets. Four baselines were defined: laplacian score, spectral analysis, non-negative discriminative unsupervised feature selection and a previous approach of the authors (Tang and Liu 2012). Once all features were pre-processed they were used for clustering the instances by K-means. According to the authors, the presented approach outperformed all considered baselines for both datasets.

Although interesting, the paper has several limitations. First, it is not very clear on the exact definition of feature. Although the authors claim that the approach is supposed to be designed for streaming environments, it requires all data instances to be known in advance, as are features instead of instances which arrive one at the time. Second, the authors assumed that link information was relatively stable. As a result, social links between data instances were never updated to reflect changes in the social network. Third, the experimental evaluation was performed in a batch setting. Data instances were clustered only once all features were analysed. As a result, performance was not really assessed in a real online environment, where data instances continuously arrive in a stream and need to be clustered or classified in real-time. Fourth, even though the approach is supposed to have a computational complexity of $\Theta\left(n^2 s^2 t\right)$, where n represents the number of instances, s the number of selected features and t the total number of features, solving the optimisation problem requires the computation of arithmetic operations between high-dimensional matrices, which could further increase the computational complexity.

4.3 Summary

Interestingly, the numerous general purpose FS techniques that claim to be applicable on online or streaming settings suffer from different short-comings that might affect their applicability. First, they are intended to be applied on specific streaming environments in which all data instances are required to be known in advance, as are features instead of instances which arrive one at the time. Second, its evaluation is usually performed in a batch setting, i.e. all features are processed before instances are clustered or classified. As a result, performance is not assessed in a real online environment, where instances continuously arrive in a stream. Third, techniques might lack of scalability. Considering the high-dimensionality of feature spaces in streaming environments, techniques involving solving optimisation problems that require the computation of arithmetic operations between large-scale matrices might not applicable on online settings due to their extensive computational complexity. Moreover, considering such matrices could incur in substantial memory consumption. As regards techniques specifically designed for the short-text domain, two of them have low computational complexity. However, features seemed to be specifically designed for the described task to perform, which might imply that they cannot be generalised to other domains or tasks, or even they might not adequately cope with the dynamically changing environment of social media data. On the other hand, the last technique incurs in a quadratic cost, which might not result adequate for high-dimensional feature spaces. Finally, as for batch FS techniques, the

results obtained for the different techniques might not be comparable due to differences in the experimental settings and datasets.

5 Discussion

Tables 1, 2 and 3 summarise the main characteristics of the presented FS techniques. Each Table analyses the type of features used, the pre-processing technique, whether they update the set of selected features, the computational complexity, and whether the techniques are scalable, among other relevant characteristics. Although FS has received considerable attention during the last decades, most studies focus on developing batch techniques instead of facing the challenging problem of OFS. At the present day, efficient and scalable OFS is an important requirement in numerous large-scale social applications. However, despite presenting significant advantages in efficiency and scalability, the existing OFS techniques are not always accurate enough as they are not based on complete information, and still not sufficiently efficient when handling ultra-highly dimensional and massive-scale data (Wu et al. 2014). Table 4 analyses the suitability of the techniques described in Sect. 4 for performing OFS. As the Table shows, most techniques are not adequate for performing OFS.

Thus, in order to mine big data in real-world applications, new techniques to cope with the efficiency requirements of online learning processes have to be developed (Hoi et al. 2012).

The main shortcomings of existing techniques can be summarised as follows.

5.1 Individual versus Grouped features

Most techniques consider individual features assuming that they are independent and identically distributed. However, this might not hold in social media since, when measuring the relevance of features in isolation possible dependencies among them might be ignored. Linked data has become ubiquitous in social networks such as *Twitter* (in which not only tweets can be linked, but also their authors might be socially related) or *Facebook* (in which users can be connected by friendship relations). Interestingly, the problem of FS for link data is rarely addressed, and those approaches that do consider it might not be suitable for online environments. Note than only a few of the reviewed techniques analysed the relation between features, and only a small proportion assessed the linked social nature of social media data. The proportion increases when only OFS techniques are analysed, in which more than a half of the reviewed techniques considers groups of features. However, due to scalability issues, only a few of them might be suitable for real-time settings (Tables 2, 3, Column "*Scalable?*" and Table 4, Column "*Suitablity*"). These approaches aiming at considering the linked nature of data need to address two challenges (Tang et al. 2014c): how to exploit relations among data instances, and how to leverage those relations for FS.

5.2 Linked nature of social media data

Another challenge for FS techniques is posed by the linked nature of social media data (Tang and Liu 2014a), as most techniques cannot fully leverage the advantages of social data dimensions, which are different from the traditional feature value ones. The majority of FS techniques are designed for data containing uniform entities, i.e. feature-value data, which are typically assumed to be independent and identically distributed. However, social media data does not follow that assumption, as data instances are inherently linked through social

Table 1 Summary of batch feature selection techniques based on individual features

	Type of features (textual-social)	Data pre-processing?	Language dependence?	Updates selected features?	Computational complexity?	Explicit selection of feature number	Scalable?	Data source	Evaluation
Rosa and Ellen (2009)	Textual	Removal of rare terms	No	No	Low	Yes	Yes	Medium-scale number of short-text military chat	Classification
Saif et al. (2014)	Textual	No	No	No	Low to medium	Yes	Yes	Small-scale number of tweets	Classification
Wang et al. (2012)	Textual	Tokenisation. Nouns, verbs, and adjectives are kept	Yes	No	Medium	Yes	No	Chinese short-texts	Multi-class classification
Saif et al. (2012)	Textual	Remove of usernames, URLs, hashtags and non-alphabetic characters	Yes	No	Medium	Yes	No	Stanford *Twitter* sentiment dataset	Classification
Liu et al. (2010)	Textual	Tokenisation. Nouns, verbs, and adjectives are kept	Yes	No	Medium	Yes	No	*Sina* short-texts	Multi-class classification
Ferragina and Scaiella (2012)	Textual	No	Yes	No	Low to medium	Yes	No	Tweets	Comparison to other approaches
Tang et al. (2014b)	Textual	Tokenisation. Nouns, verbs, and adjectives are kept	Yes	No	Medium to high	Unknown	No	2450 tweets	Clustering
Perez-Tellez et al. (2010)	Textual	No	Yes	No	Medium to high	Unknown	No	English tweets of 20 companies	Clustering
Hu et al. (2009)	Textual	Stopword removal. Pre-process the world knowledge	Yes	No	Medium	Yes	No	Web snippets with less than 50 words	Clustering

Table 1 continued

	Type of features (textual-social)	Data pre-processing?	Language dependence?	Updates selected features?	Computational complexity?	Explicit selection of feature number	Scalable?	Data source	Evaluation
Tang et al. (2012)	Textual	No	No	No	Medium to high	No	No	*Facebook* and *Twitter* datasets	Clustering
Verma et al. (2011)	Textual and binary features.	Tokenisation. Stopword, URLS and special symbols removal	Yes	No	Low to medium	Unknown	No	Small-scale number of tweets	Classification
Ma et al. (2013)	Textual and contextual features	No	No	No	Medium to high	Unknown	No	31 million tweets	Classification
Amir et al. (2014)	Textual	Remove of usernames, URLs Normalised words including repeated characters	No	No	Medium to high	No	Needs GPU	*SemEval2014*	Classification
Tang et al. (2014a)	Textual	Tokenisation. Remove of usernames, URLs, tweets with less than 7 words	No	No	Medium to high	No	Needs GPU	*SemEval2013*	Classification
Severyn and Moschitti (2015)	Textual	Unknown	No	No	Medium to high	Yes	Needs GPU	*SemEval2015*	Classification
Jiang et al. (2011)	Textual	Remove tweets with less than 7 tokens	No	No	Medium to high	No	Needs GPU	5 million tweets	

Table 2 Summary of batch feature selection techniques based on group of features

	Type of features (textual-social)	Data pre-processing?	Language dependence?	Updates selected features?	Computational complexity?	Explicit selection of feature number	Scalable?	Data source	Evaluation
Alexandrov et al. (2005)	Textual	Removal of low-frequency words	No	No	Low to medium	No	No	CICLing-2002 abstracts	Clustering
Ozdikis et al. (2012)	Textual	Tokenisation. Stopword removal. Stemming	No	No	Medium	No	No	Over than 150,000 tweets	Clustering
Moradi and Rostami (2015)	Categorical, integer and real	No	No	Yes	Medium to high	Depends on a parameter	Yes	Datasets from the UCI repository	Binary and multi-class classification
Tang and Liu (2012)	Textual and social	Stopword removal. Stemming. TF-IDF irrelevant feature removal	No	No	High	Yes	No. Needs optimisation	*Digg* and *BlogCatalog* datasets	Multi-class classification
Tang et al. (2013)	Diverse textual features	Stopword removal. Stemming.	No	No	High	Yes	Yes	*Flickr* and *BlogCatalog* datasets	Clustering
Fang et al. (2014)	Textual, social and temporal	Stopword and special symbol removal	No	No	Medium	Yes	Yes	*Twitter*	Clustering
Tang and Liu (2014b)	Textual and social	No	No	No	High Cubic on the number of features	Yes	No	*BlogCatalog* and *Flickr* datasets	Clustering
Gu and Han (2011)	Textual and social	Unknown	No	No	High	No	No. Needs optimisation	*WebKB* and *Cora* datasets	Multi-class classification

Table 3 Summary of online feature selection techniques

	Type of features (textual-social)	Data pre-processing?	Language dependence?	Updates selected features?	Computational complexity?	Explicit selection of feature number	Scalable?	Data source	Evaluation
Perkins and Theiler (2003)	Categorical, integer, real	No	No	Yes	Medium	No	Yes	Datasets from the UCI repository	Binary classification
Zhou et al. (2006)	Numeric	No	No	No	Medium	Yes	Yes	Synthetic data with features created independently from a normal distribution	Binary classification
Wu et al. (2010)	Categorical, integer, real	No	No	Yes	High	No	No	Datasets from the UCI repository	Binary classification
Wang et al. (2013)	Categorical, integer, real	No	No	Yes	High. Cubic on the numner of features	No	Yes	Datasets from the UCI repository	Binary and multi-class classification
Wang et al. (2014)	Categorical, integer, real and textual	No	No	No	Medium to high	No	Yes	Datasets from the UCI repository, *Reuters corpus Volume 1* and *20Newsgroups*	Binary classification
Li et al. (2012)	Textual	Tweet segmentation	Yes	No	Medium to high	No	Yes	Tweets from Singaporean users	Clustering
Becker et al. (2011)	Textual, social and temporal	No	No	Yes	Low	Fixed number	Yes	Over than 2,600,000 tweets	Multi-class classification
Zubiaga et al. (2015)	Statistics from textual features	Stopword and special *Twitter* words removal	No	No	Low	Fixed number	Yes	Over than 500,000 tweets	Multi-class classification
Li et al. (2015)	Textual and social	No	No	Yes	High Cubic	Yes	No	*BlogCatalog* and *Flickr* datasets	Clustering

Table 4 Applicability of feature selection techniques for online feature selection

	Advantages	Disadvantages	Suitability
Perkins and Theiler (2003)	The time required to test previous features is small in comparison to the time required for adding a new feature. It combines the speed of filters with the accuracy of wrappers.	Considers features individually and independently. Requires all candidate features to be known in advance.	No
Zhou et al. (2006)	–	Considers features individually and independently. Does not fully exploit the combination of social and content information. It cannot be guaranteed that no redundancy exists among features.	No
Wu et al. (2010)	Analyses both feature redundancy and relevance.	Considers features individually and independently. High computational complexity.	No
Wang et al. (2013)	Features are considered in related groups. The number of features to select is dynamically chosen.	It requires computing several arithmetic operations between matrices. Does not provide any mechanism for dealing with repeated features.	No
Wang et al. (2014)	Considers a sequential arrival of instances. Does not need to know all instances in advance.	The authors assumed that both training and test instances are always similar.	No
Li et al. (2012)	The time complexity is linear on the number of detected segments in each time window.	Needs domain information. Do not fully exploits the combination of social and content information.	Yes
Becker et al. (2011)	Selected features are regularly updated.	Only considers statistical features. Does not fully exploit the combination of social and content information.	Yes
Zubiaga et al. (2015)	Requires a small set of features that can be straightforwardly computed. Does not make use of any source of external data. Requires low computational cost. The number of features remains unchanged as the number of instances and their content increases.	Only considers statistical features. Does not fully exploit the combination of social and content information.	Yes
Li et al. (2015)	Features are considered in related groups Considers both textual and social features. Selected features are updated.	Requires computing several arithmetic operations between matrices.	No

relationships. Nonetheless, social data provides extra information beyond the feature-value one, which can provide correlations between instances. For example, posts from the same user or two linked users are more likely to have similar topics. In this regard, the availability of link information enables performing advance research in FS techniques. Although social data seems to be important when performing OFS, only a third of the reviewed works considered social information (Table 3, Column "*Types of Features*"). When analysing techniques for batch FS, only a small proportion of them included some kind of social information (Tables 1 and 2, Column "*Types of Features*").

5.3 Stability

Stability can be defined as the sensitivity of the FS technique to data perturbation in the training set, usually caused by noise. Therefore, a good FS technique should be sufficiently robust to handle noise and return stable results that only contain relevant features across the different data samples. In real applications, domain experts would prefer stable FS techniques, as unstable FS ones would lower the confidence of results (Alelyani et al. 2011). However, domain experts would not be interested in a technique that yields very stable feature subsets but not effective predictive models (Han and Yu 2012). Consequently, stability can be considered a desired characteristic of FS techniques as it might affect their predictive performance (Han and Yu 2012). It was found that the stability of FS techniques can be greatly affected by the underlying data characteristics (Alelyani et al. 2011), such as the total number of features, the number of instances, data distribution, and the order in which data is processed. It was found that having a large number of instances has a positive impact on stability, whereas a large number of features has a negative impact. In addition, a large enough number of instances vanishes the impact of a large number of features. In this regard, it is important to consider the settings in which the experimental results were obtained. For example, numerous techniques were evaluated with small-size datasets including fewer than 10,000 instances (Tables 1, 2, 3, Column "*Data Source*"). Consequently, the obtained results might not be generalisable to a higher numbers of instances. Moreover, stability can be affected by the quality of data. Several OFS techniques assumed that test instances were similar to the training ones and did not update the selected features as more instances became available (Table 3, Column "*Updates Selected Features?*"). This situation could have a severe impact on the stability and generalisation of results when such assumption does not hold, as it might occur in dynamically changing environments in which new topics constantly emerge and old topics become obsolete. In this context, developing techniques with high predicting performance and stability remains an open challenge.

5.4 Number of data instances needed

FS techniques usually require a sufficient number of data instances to obtain statistically significant results. For example, it might be difficult to assess the relevance of features by only considering a unique data instance. In the case of supervised techniques, the problem aggravates as the number of distinct classes increases, and therefore the size of the needed training set also increases (Forman 2004). Hence, some classes will inevitably be more difficult than others to predict. Forman (2004) hypothesised that, in the case it is difficult to get good predictive features for some classes, existing techniques will focus on the features that are useful predictors for other easier classes, thus ignoring the difficult classes. This is even worse in social media data where the number of training instances for each

class might not be high (consider the techniques that were used for classifying posts—Tables 1, 2, 3, Column "*Evaluation*"—that considered a low-scale dataset - Column "*Data Source*"). FS techniques adequately dealing with an unbalanced set of training instances should be developed. Such techniques should be able to accurately assess the relevance of features, in presence of a reduced set of instances per class.

5.5 Number of features to select

Most techniques require pre-defining a fixed number of features to select, which might be difficult to choose without prior knowledge. This number does not only influence the degree of information lost, but also the efficiency of the learning task to be performed. As there is no standard or reliable method to choose the optimal number of features, it is generally chosen through experimental evaluation and never updated, which might be inadequate for dynamic environments. Interestingly, almost the half of the reviewed techniques manually defined the number of features to select (Tables 1, 2 and 3, Column "*Explicit Selection of Feature Number*"), further hindering their suitability in highly dynamic environments, in which new features constantly appear. As a result, techniques should be able to automatically choose the number of features according to some statistical threshold.

5.6 Scalability

The continuous grow of social media data puts in jeopardy the scalability of current techniques (Tang et al. 2014c), especially in real-time tasks. For example, most techniques require all data to be loaded into memory, which might not be possible when analysing social media data. Furthermore, other techniques might require iterative processes where each data instance is analysed more than once until convergence. Other aspect related to scalability is the computational complexity of techniques (Tables 1, 2, 3, Column "*Computational Complexity*"). Whereas for batch FS techniques computational complexity might not be highly relevant, in the case of OFS, computational complexity is critical. Techniques must be able to process data instances as they are generated in a reasonable amount of time to provide real-time answers. Consequently, techniques must be designed to achieve high performance. Most of the analysed techniques fail to achieve high performance as, for example, require repetitively computing arithmetic operations between high-dimensional matrices, or semantically enriching short-text by applying expensive language processing techniques. Moreover, several techniques were evaluated considering a limited number of instances (Tables 1, 2, 3, Column "*Data Source*"), which implies that the scalability of techniques has not been actually assessed. Note that only a a third of the reviewed techniques could be considered scalable (Tables 1, 2, 3, Column "*Scalable?*").

In summary, there is 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.

6 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. Dimensionality reduction helps to speed up data mining algorithms and thus, improve mining performance (Liu and Yu 2005). This survey discussed state-of-

the-art short-text FS approaches divided into two groups. The first group reviewed standard batch FS techniques, which assume the existence of a fixed set of instances, and therefore a feature space fully known in advance. Batch feature selection is effective as counts with complete information regarding the full set of instances and features, but it is not suitable for dynamic environments as training examples could arrive sequentially, features might appear incrementally or it could be difficult to collect a training set. Good examples of these dynamic environments are social network sites, such as *Facebook* or *Twitter*, where users are constantly creating new posts, tags or even words. The second group reviewed OFS techniques in which instances and their corresponding features arrive in a continuous stream. OFS techniques involve choosing a subset of features and its corresponding learning model at different time frames, i.e. they react to new features by adapting the subset of selected ones as new features appear. Both groups of techniques can be further categorised according to whether they process individual features or groups of them.

Although FS techniques have been the focus of extensive research for the last few decades, which implies the existence of a wide range of approaches, this work has identified several shortcomings shared by most approaches in the context of learning with social short-texts. First, most of the presented techniques do not leverage on the relationships between features, assuming that they are independent from each other. Although this assumption might be valid on traditional FS tasks, when considering short-texts in social media data, it does not further hold. In such scenario, data instances are linked through social links, which might provide valuable information on the correlation between instances. In this regard, the availability of link information enables the development of advanced FS techniques. Second, most FS approaches require to pre-define the number of features to select, in many cases with no prior information. The selection of this number directly affects the quality of the selected feature subset, and thus, that of the subsequent learning task to perform. Third, most techniques suffer from scalability limitations. For example, several techniques are based on performing arithmetic operations between high-dimensional matrices, which are inherently computationally complex. Most experimental evaluations were carried out with a limited number of instances, meaning that the scalability of techniques has not been actually assessed. Finally, in general terms, evaluations have been performed using different datasets, which are not publicly available on the majority of cases, hindering the generalisation and replication of results.

In summary, future research on FS should focus on facing the challenges posed by the new dynamics of social environments and leveraging all types of information provided by them. Approaches should be capable of effectively adapting to large-scale social applications by providing efficient and scalable solutions for handling massive-scale data with ultra-high dimensionality as produced by short-texts. Particularly, techniques should be capable of dealing with unbalanced datasets by accurately assessing the relevance of features. Additionally, the requirement of a pre-defined fixed number of features to select should be revisited in order to make FS techniques more adequate to dynamic environments in which new features are constantly emerging.

References

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. J Mach Learn Res 9:1981–2014

Alelyani S, Liu H, Wang L (2011) The effect of the characteristics of the dataset on the selection stability. In: Proceedings of the 23rd IEEE international conference on tools with artificial intelligence (ICTAI), IEEE Computer Society, pp 970–977

Alelyani S, Tang J, Liu H (2013) Feature selection for clustering: a review. In: Aggarwal CC, Reddy CK (eds) Data clustering: algorithms and applications - Chapman & Hall/CRC data mining and knowledge discovery series, Chapman and Hall/CRC, Boca Raton, pp 29–60

Alexandrov M, Gelbukh A, Rosso P (2005) An approach to clustering abstracts. In: Montoyo A, Muñoz R, Métais E (eds) Natural language processing and information systems, vol 3513, Lecture notes in computer science, Springer, Berlin, pp 275–285

Amir S, Almeida MB, Martins B, Ja Filgueiras, Silva MJ (2014) Tugas: exploiting unlabelled data for twitter sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). Association for computational linguistics and Dublin City University, Dublin, Ireland, pp 673–677

Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on twitter. In: Proceedings of the 5th international conference on weblogs and social media, The AAAI Press, Spain

Chen M, Jin X, Shen D (2011) Short text classification improved by learning multi-granularity topics. In: Walsh T (ed) Proceedings of the 22th international joint conference on artificial intelligence, The AAAI Press, IJCAI'11, pp 1776–1781

Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics. Association for computational linguistics, Baltimore, pp 49–54

Fang Y, Zhang H, Ye Y, Li X (2014) Detecting hot topics from twitter: a multiview approach. J Inf Sci 40(5):578–593

Ferragina P, Scaiella U (2012) Fast and accurate annotation of short texts with wikipedia pages. IEEE Softw 29(1):70–75

Forman G (2004) A pitfall and solution in multi-class feature selection for text classification. In: Proceedings of the 21st international conference on machine learning, ACM, New York, NY, USA, ICML'04, p 38

Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305

Gabrilovich E, Markovitch S (2006) Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st national conference on artificial intelligence. MA, USA, Boston, pp 1301–1306

Gu Q, Han J (2011) Towards feature selection in network. In: Proceedings of the 20th ACM international conference on information and knowledge management, ACM, New York, NY, USA, CIKM'11, pp 1175–1184

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Han Y, Yu L (2012) A variance reduction framework for stable feature selection. Stat Anal Data Min 5(5):428–445

Hoi SCH, Wang J, Zhao P, Jin R (2012) 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, BigMine'12, pp 93–100

Hu X, Sun N, Zhang C, Chua TS (2009) Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM'09, pp 919–928

Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies—vol 1, Association for computational linguistics, Stroudsburg, PA, USA, HLT'11, pp 151–160

Jin O, Liu NN, Zhao K, Yu Y, Yang Q (2011) Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, New York, NY, USA, CIKM'11, pp 775–784

John GH, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: Proceedings of the 11th international conference of machine learning, Morgan Kaufmann, ICML'94, pp 121–129

Li J, Hu X, Tang J, Liu H (2015) Unsupervised streaming feature selection in social media. In: Proceedings of the 24th ACM international on conference on information and knowledge management, ACM, New York, NY, USA, CIKM'15, pp 1041–1050

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management, ACM, New York, NY, USA, CIKM'09, pp 375–384

Li C, Sun A, Datta A (2012) 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, CIKM'12, pp 155–164

Liu ZLZ, Yu WYW, Chen WCW, Wang SWS, Wu FWF (2010) Short text feature selection for micro-blog mining. In: Proceedings of the 2nd international conference on computational intelligence and software engineering, IEEE, CISE'10, pp 4–7

Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502

Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in twitter. J Am Soc Inf Sci Technol 64(7):1399–1410

Marsden PV, Friedkin NE (1993) Network studies of social influence. Sociol Methods Res 22(1):127–151

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

Medelyan O, Milne D, Legg C, Witten IH (2009) Mining meaning from wikipedia. Int J Hum Comput Stud 67(9):716–754

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th annual conference on neural information processing systems. Lake Tahoe, Nevada, USA, pp 3111–3119

Moradi P, Rostami M (2015) A graph theoretic approach for unsupervised feature selection. Eng Appl Artif Intell 44(C):33–45

Ozdikis O, Senkul P, Oguztuzun H (2012) Semantic expansion of tweet contents for enhanced event detection in twitter. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining, IEEE Computer Society, Istanbul, Turkey, ASONAM'12, pp 20–24

Peng Y, Xuefeng Z, Jianyong Z, Yumhong X (2009) Lazy learner text categorization algorithm based on embedded feature selection. J Syst Eng Electron 20(3):651–659

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. Doha, Qatar, pp 1532–1543

Perez-Tellez F, Pinto D, Cardiff J, Rosso P (2010) On the difficulty of clustering company tweets. In: Proceedings of the 2nd international workshop on search and mining user-generated contents, ACM, New York, NY, USA, SMUC'10, pp 95–102

Perkins S, Lacker K, Theiler J (2003) Grafting: fast, incremental feature selection by gradient descent in function space. J Mach Learn Res 3:1333–1356

Perkins S, Theiler J (2003) Online feature selection using grafting. In: Fawcett T, Mishra N (eds) Proceedings of the 21st international conference on machine learning, AAAI Press, ICML'03, pp 592–599

Rafeeque P, Sendhilkumar S (2011) A survey on short text analysis in web. In: Proceedings of the 3rd international conference on advanced computing, IEEE, Chennai, India, ICoAC'11, pp 365–371

Rosa KD, Ellen J (2009) Text classification methodologies applied to micro-text in military chat. In: Proceedings of the 2009 international conference on machine learning and applications, IEEE Computer Society, Washington, DC, USA, ICMLA'09, pp 710–714

Saeys Y, In Inza, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Saif H, Fernández M, He Y, Alani H (2014) On stopwords, filtering and data sparsity for sentiment analysis of twitter. In: Proceedings of the 9th international conference on language resources and evaluation, European Language Resources Association (ELRA), Reykjavik, Iceland, LREC'14, pp 810–817

Saif H, He Y, Alani H (2012) Alleviating data sparsity for twitter sentiment analysis. In: Proceedings of the 2nd workshop on making sense of microposts: big things come in small packages at the 21st international conference on the World Wide Web, CEUR Workshop Proceedings, MSM'12, pp 2–9

Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the 6th international conference on new methods in language processing, Manchester, UK, NeMLaP'94

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47

Severyn A, Moschitti A (2015) Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on research and development in information retrieval, ACM, New York, NY, USA, SIGIR'15, pp 959–962. doi:10.1145/2766462.2767830

Strassen V (1969) Gaussian elimination is not optimal. Numer Math 13(4):354–356

Tang J, Wang X, Gao H, Hu X, Liu H (2012) Enriching short text representation in microblog for clustering. J Front Comput Sci China 6(1):88–101

Tang J, Alelyani S, Liu H (2014c) Feature selection for classification: A review. In: Aggarwal CC, Reddy CK (eds) Data classification: algorithms and applications - Chapman & Hall/CRC data mining and knowledge discovery series, Chapman and Hall/CRC, Boca Raton, pp 37–64

Tang J, Hu X, Gao H, Liu H (2013) Unsupervised feature selection for multi-view data in social media. In: Proceedings of the SIAM international conference on data mining, SIAM, SDM'13, pp 270–278
Tang J, Liu H (2012) Feature selection with linked data in social media. In: Proceedings of the 12th SIAM International conference on data mining, SIAM / Omnipress, pp 118–128
Tang J, Liu H (2014a) Feature selection for social media data. ACM Trans Knowl Discov Data 8(4):19:1–19:27
Tang J, Liu H (2014b) An unsupervised feature selection framework for social media data. IEEE Trans Knowl Data Eng 26(12):2914–2927
Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014a) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, The Association for computer linguistics, Baltimore, MD, USA, pp 1555–1565
Tang G, Xia Y, Wang W, Lau R, Zheng F (2014b) Clustering tweets using wikipedia concepts. In: Proceedings of the 9th international conference on language resources and evaluation, European Language Resources Association (ELRA), Reykjavik, Iceland, LREC'14
Verma S, Vieweg S, Corvey W, Palen L, Martin JH, Palmer M, Schram A, Anderson KM (2011) Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In: Proceedings of the 5th International AAAI conference on web and social media, The AAAI Press, ICWSM'11
Wang Bk, Huang YF, Yang Wx, Li X (2012) Short text classification based on strong feature thesaurus. J Zhejiang Univ Sci C 13(9):649–659
Wang J, Zhao P, Hoi S, Jin R (2014) Online feature selection and its applications. IEEE Trans Knowl Data Eng 26(3):698–710
Wang J, Zhao ZQ, Hu X, Cheung YM, Wang M, Wu X (2013) Online group feature selection. In: Proceedings of the 23rd international joint conference on artificial intelligence, AAAI Press, IJCAI'13, pp 1757–1763
Wu Y, Hoi SCH, Mei T (2014) Massive-scale online feature selection for sparse ultra-high dimensional data. Computing Research Repository abs/1409.7794. https://arxiv.org/abs/1409.7794
Wu X, Yu K, Wang H, Wei D (2010) Online streaming feature selection. In: Proceedings of the 27th international conference on machine learning (ICML-10), Omnipress, ICML'10, pp 1159–1166
Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proceedings of the 14th international conference on machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'97, pp 412–420
Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
Zhou J, Foster DP, Stine RA, Ungar LH (2006) Streamwise feature selection. J Mach Learn Res 7:1861–1885
Zubiaga A, Spina D, Martínez R, Fresno V (2015) Real-time classification of twitter trends. J Assoc Inf Sci Technol 66(3):462–473