Sentiment Analysis in Social Streams

Book Section

How to cite:

For guidance on citations see FAQs.

© 2016 Springer International Publishing

Version: Accepted Manuscript

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1007/978-3-319-31413-6;

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
Abstract In this chapter we review and discuss the state of the art on sentiment analysis in social streams –such as web forums, micro-blogging systems, and social networks–, aiming to clarify how user opinions, affective states, and intended emotional effects are extracted from user generated content, how they are modeled, and how they could be finally exploited. We explain why sentiment analysis tasks are more difficult for social streams than for other textual sources, and entail going beyond classic text-based opinion mining techniques. We show, for example, that social streams may use vocabularies and expressions that exist outside the mainstream of standard, formal languages, and may reflect complex dynamics in the opinions and sentiments expressed by individuals and communities.

1 Introduction

Sentiment Analysis is the field of study that analyzes the people’s attitudes towards entities –individuals, organizations, products, services, events, and topics–, and their attributes [36]: The attitudes may correspond to personal opinions and evaluations, affective states (sentiments and moods), or intended emotional effects. It represents a large problem space, covering different tasks, such as subjectivity identification, sentiment extraction and analysis, and opinion mining, to name a few.

Hassan Saif
Knowledge Media Institute, Milton Keynes, United Kingdom, e-mail: h.saif@open.ac.uk

F. Javier Ortega
Universidad de Sevilla, Seville, Spain, e-mail: javierortega@us.es

Miriam Fernández
Knowledge Media Institute, Milton Keynes, United Kingdom, e-mail: m.fernandez@open.ac.uk

Iván Cantador
Universidad Autónoma de Madrid, Madrid, Spain, e-mail: ivan.cantador@uam.es
Although some of the above tasks have been addressed on multi-modal data sources – e.g., sentiment extraction in audio and video –, from its origins, sentiment analysis has mainly focused on textual data sources [48]. Hence, it commonly refers to the use of natural language processing, text analysis, and computational linguistics to extract and exploit subjective information from text materials. In Chapters ?? and ?? the reader can find overviews of the state of the art in affective information representation and acquisition for various modalities.

With the advent of the Social Web, the amount of text material is huge and grows exponentially every day. The Web is a source of up-to-date, never-ending streams of user generated content; people communicate online with contacts in social networks, create or upload multimedia objects in online sharing sites, post comments, reviews and ratings in blogs and recommender systems, contribute to wiki-style repositories, and annotate resources in social tagging platforms.

The Web thus provides unstructured information about user opinions, moods and emotions, and tastes and interests, which may be of great utility to others, including consumers, companies, and governments. Hence, for instance, someone who wants to buy a camera may look in web forums for online opinions and reviews about different brands and models, while camera manufacturers implicitly/explicitly get feedback from customers to improve their products, and adapt their marketing strategies. Very interestingly, this information can go beyond reflecting the users’ subjective evaluations and sentiments about entities and their changes over time, by triggering chains of reactions and new events. For instance, identifying the overall concern, expressed in social media, on certain political decision may impact the modification or rejection of such decision.

The interest and potential exploitation of sentiment analysis in social streams – understood as social media in which user generated content emerges and changes rapidly and constantly –, are evident, and have been shown in numerous domains and applications, like politics and e-government [6][45][77], education and e-learning [76], business and e-commerce [85], and entertainment [23][72][80]. The reader is referred to several chapters of this book for detailed surveys of particular applications of affective information by personalized services, specifically by recommender systems (Chapters ??, ??, and ??), conversational systems (Chapter ??), multimedia retrieval systems (Chapters ?? and ??), and e-learning systems (Chapter ??).

The high availability of user generated content in social streams, nonetheless, comes with some challenges. The large volume of data makes difficult to get the relevant information in an efficient and effective way. Proposed techniques have to be simple enough to scale up, but have to deal with complex data. Some of these challenges are related to general natural language processing (NLP) approaches, such as opinion-feature association [31], opinion negation [32], irony and sarcasm [12][18], and opinion spam [33]. Others, in contrast, are related to issues characteristic of online user generated content, such as multiple languages, high level of ambiguity and polysemy, misspellings, and slang and swear words [70]. In this context, it is also important to mention the need of determining the users’ reputation and trust. For certain topics, the majority opinion (i.e., the wisdom of the crowd) may be the best solution [49], while for others, only the experts’ opinions should be the source
of information to consider [87]. Another relevant issue is the existence of particular pieces and forms of information existing in social streams: explicit citations to users, groups and organizations (e.g., @robinwilliams in Twitter), explicit forms for referring to concepts (e.g., Twitter hashtags #comedian and #funny), emoticons and slang terms noting emotions and moods (e.g., :D and lol), mechanisms to express interests and tastes (e.g., Facebook likes), and URLs to resources that complement posted information. There, the use of contextual metadata also plays a key role; extracting and mining time and geo-location metadata may be very valuable for sentiment analysis on dynamic and global social stream data.

In this chapter, we review and discuss the state of the art on sentiment analysis in social streams, describing how opinion and affective information is extracted, processed, modeled, and exploited, in comparison to classic text-based opinion mining techniques.

The chapter is structured as follows. In Section 2 we overview the research literature in Sentiment Analysis, focusing on the main addressed tasks and applied techniques. In Section 3 we provide a description of social media, characterizing the user generated content and research challenges that arise from them. Next, in Section 4 we discuss Sentiment Analysis to social streams, and describe existing applications in such context. Finally, in Section 5 we discuss current and open research trends on Sentiment Analysis in social streams.

2 Sentiment Analysis

In the last fifteen years, Sentiment Analysis and Opinion Mining have been fed by a number of research problems and opportunities of increasing importance and interest [48]. In this section we review the main tasks addressed in the literature related to sentiment analysis, together with the different assumptions and approaches adopted. We then discuss some interesting proposals, resources and techniques intended to deal with those tasks.

2.1 Sentiment Analysis Tasks

The different sentiment analysis tasks can be categorized based on the granularity of their linguistic units they consider. In this sense, there are tasks where the document is assumed to be the main linguistic unit as a whole, while there are others where sentences or even words are considered as linguistic units. We can summarize these levels as follows:

- **Document-level**: At this level, it is assumed that each document expresses a particular sentiment, or at least it poses a predominant one. Many works have faced sentiment analysis tasks at the document level; see for example the survey presented in [78].
• **Sentence-level**: Some tasks could benefit from the determination of the sentiment in a text at a sentence level, as done in information extraction and question answering systems, where it is necessary to provide the user with particular information for a given topic.

• **Aspect-level**: In general, a sentence can contain more than one opinion about different aspects of an entity or topic. In an aspect-level approach, the context of the words are taken into account to determine the subjectivity of each expression in a sentence, and the specific aspect being opinionated [82][84]. This level can be useful, for example, in recommender systems [13], and in automatic processing of product reviews [16][44], where knowing individual opinions about each feature of a given product is crucial for the performance of the system.

• **Word-level** (also called as **entity level**): In this category we can find those tasks consisting of identifying the sentiment expressed by a given word regardless its context. Word-level analysis is useful in order to build resources like sentiment lexicons with the possible sentiment orientations of a word [29][64].

Another possible classification of sentiment analysis tasks can be made from the point of view of the dependency on the target domain. While some tasks are defined independently of the domain of application –like subjectivity detection–, some research works have shown the influence of domain-dependency on sentiment analysis problems –e.g., polarity detection [16][53][52][83].

In general, the following are the main goals of sentiment analysis:

• **Subjectivity detection**. Identifying subjective and objective statements.

• **Polarity opinion detection**. Identifying positive and negative opinions within subjective texts.

• **Emotion detection**. Identifying human emotions and moods.

Subjectivity detection can provide valuable knowledge to diverse NLP-based applications. In principle, any system intended to extract pieces of information from a large collection of texts could take advantage of subjectivity detection approaches as a tool for identifying and considering/discard non-factual information [57]. Such is the case of question answering [86] and information extraction systems.

Polarity detection aims to identify whether a text expresses a positive or a negative sentiment from the writer. Since it is very common to address this task only on subjective texts, usually a subjectivity detection stage is needed. Hence, in the literature we can find a number of works that tackle both problems –subjectivity and polarity detection– as a single one. Existing approaches commonly distinguish between three types of texts: positive, negative, and neutral or objective texts. Some works have shown this approach is much more challenging than the binary classification of subjective texts [57]. Applications of polarity classification are the identification of the writer’s political ideology –since it can be considered as a binary classification problem [21]–, and the analysis of product reviews –determining user positive or negative opinions about a given item (a product, a movie, a hotel, etc.) or even personal sentiments about specific features of such item.

In emotion detection, the main object of study is the user’s emotional attitude with respect to a text. In this context, we may aim to determine the writer’s mood
towards a text [82] or to identify the emotions “provoked” by the text to the reader [68].

### 2.2 Sentiment Analysis Approaches

In this section we discuss some interesting approaches intended to deal with the sentiment analysis tasks and goals previously described. For the sake of clarity, we classify them into two groups, according to the nature of the applied techniques:

- **Lexicon-based approaches** are those techniques that rely on a resource containing information about the affective terms that may occur in the texts, and usually additional information about such terms (e.g., polarity, intensity, etc.). These resources can be manually or automatically generated, domain independent or focused on a particular domain. Most of these approaches take advantage of the information available in a lexicon to compute subjective and affective estimations over the texts.

- **Machine Learning approaches** are those techniques that apply a machine learning method to address sentiment analysis tasks. In this case, a majority of techniques have been based on Support Vector Machines, which are usually fed with lexical and syntactic features, or even with lexicon-based features, to provide subjective and affective classifications.

It is worth to note that the creation, integration and use of lexicons are crucial in Sentiment Analysis, not only for lexicon-based techniques, but also for machine-learning techniques, which can be enhanced with the information available in such resources. In this context, General Inquirer [69] can be considered as one of the most relevant and widely used resources. It is a manually built lexicon formed by lemmas with associated syntactic, semantic and pragmatic information. It contains 4,206 lemmas manually tagged as positive or negative.

The MPQA (Multi-Perspective Question Answering) is a lexicon of news documents from the world press based on General Inquirer, including a set of words obtained from a dictionary and a thesaurus, and a set of automatically compiled subjective terms [57]. The MPQA lexicon is composed by 8,222 words with a set of syntactic and semantic features (type strength, length, part of speech, stem, and prior polarity).

Following the same schema, the Bing Liu’s English Lexicon (BLEL) [30] consists of an automatically generated list of words that have been classified into positive and negative. This classification is manually updated periodically. In total, BLEL contains 4,783 negative words and 2,006 positive words, including misspelled terms, morphological variants, and slang words, among others.

Maybe one of the most well-known and widely used lexical resources is WordNet [42], a thesaurus for English based on the definition of the so-called *synsets*, which are groups of words with the same meaning and a brief definition (*gloss*). To
relate synsets, WordNet provides a number of semantic relations, such as synonymy, hypernymy, and meronymy.

A very large number of works have used WordNet in a wide number of tasks and domains, and some of them have aimed to enrich or expand WordNet in different ways. In this context, it is worth to mention the Global WordNet Association\(^1\), a non-commercial organization devoted to provide a platform to ease the creation and connection of WordNet versions in different languages. Regarding the enrichment of WordNet, we can highlight WordNet Domains [5], a semi-supervised generated resource that augments WordNet with domain labels for all its synsets. Related to it, we find WordNet Affect [67], which assigns to each WordNet synset a set of affective labels encoding emotions, moods, attitudes, behaviors, etc. in order to build a resource suitable for emotion detection, in addition to subjectivity and polarity classification. Another affective extension of WordNet is SentiWordNet (SWN) [4], which attaches to each WordNet synset three sentiment scores in the range \([0, 1]\) summing up to 1, representing positivity, negativity and objectivity degrees of each synset. The polarities of words are assigned by means of a propagation of the polarity of some manually picked synsets through the relations in WordNet. SWN includes 117,000 synsets with sentiment scores.

The main advantage of WordNet-based resources and techniques over MPQA, BLEL or General Inquirer is the lack of semantic ambiguity between synsets, which unequivocally represent the term meaning. Word sense disambiguation constitutes a crucial problem in NLP, and most of the works using the above lexicons address such problem by computing the polarity at the level of words or lemmas by means of the polarity values from all the respective synsets [1][71]. In addition to this, the graph structure of WordNet-based resources allows for the application of graph-based techniques in order to better exploit the semantic information encoded within the relations.

Among the existing lexicon-based approaches, the technique presented in [78] has been a main reference work for many others. This technique is applied over manually selected sets of strongly positive words (such us excellent and good) and strongly negative words (such as poor and bad), which are considered as seed terms. The technique computes the Pointwise Mutual Information (PMI) between input words and the seeds in order to determine the polarity of the former. Since the polarity of a word depends on the relation between the word and the seed sets, the technique is usually called semantic orientation by association. A similar idea is proposed in [34], but replacing the PMI computation by building a graph with the adjectives in WordNet for computing the polarity of a word; specifically, by selecting the shortest graph path from the synset of the word to the synsets of the positive and negative seeds.

With respect to machine learning-based approaches, a considerable number of works has been done, applying well-known machine learning techniques, such as SVM and LSA, to deal with sentiment analysis tasks. These works usually include the exploitation of lexical, syntactic and semantic features suitable for the classifica-

\(^1\) http://globalwordnet.org/
tion problems that must be tackled in sentiment analysis for the subjectivity and polarity detection. Among these features, one may highlight n-grams, Part-Of-Speech (POS) tags, PMI and features extracted from lexicons [19][84]. In this context, it has to be noted that the joint use of lexicon- an machine learning-based approaches can be performed in the opposite direction, i.e., by using machine-learning techniques in order to improve lexicon-based approaches. For instance, in [51] LSA-based techniques are used to expand a given lexicon for different languages.

The work presented in [29] is another representative example of a machine learning-based sentiment analysis approach. It aims to predict the orientation of subjective adjectives by analyzing a large unlabeled document set, and looking for pairs of adjectives linked with conjunctions. It then builds a graph where the nodes correspond to terms connected by equal-orientation or opposite-orientation edges, according to the conjunctions that link the terms, and finally apply a clustering algorithm that partitions the graph into clusters of positive and negative terms.

A combination of ideas from Turney [78] and Hatzivassiloglou [29] is presented in [15], where a set of seed words is used to introduce a bias in a random-walk algorithm that computes a ranking of the terms in a graph of words linked according to the conjunctions that join them in the texts. In the generated rankings, positive and negative terms are respectively located into the highest and lowest positions. The word graph is also used as a mechanism to process the negations in the text by developing a PageRank-based algorithm that builds graphs with positive and negative weighted edges.

3 Sentiment Analysis on User Generated Content

Online social media platforms support social interactions by allowing users to create and maintain connections, share information, collaborate, discuss, and interact in a variety of ways. The proliferation and usage of these platforms have experienced an explosive growth in the last decade, expanding to all areas of society, such as entertainment, culture, science, business, politics, and public services. As a result, a large amount of user generated content is continuously being created, offering individuals and organizations a fast way to monitor people’s opinions and sentiments towards any form of entity, such as products, services and brands.

The nature and purpose of these platforms is manifold, and thus they differ in a variety of aspects, such as the way in which users establish connections, the main activities they conduct, and the type of content they share. These characteristics pose novel challenges and opportunities to sentiment analysis researchers. In the subsequent sections, we characterize the user generated content available in popular types of existing social media platforms, and present the major challenges to process such content in the context of sentiment analysis and opinion mining.
3.1 Characterizing User Generated Content

In the literature, social media platforms have been categorized in different ways [35]. Here we propose a categorization based on three dimensions: the type of user connections, the type of user activities, and the type of contents generated/shared within the platforms. We summarize such a categorization in Table 1.

- **Connections**: Users connections –e.g., friendship and following relations– in social media are based on three main models: explicit connections, which can be reciprocal –u follows v, and v follows u– and non-reciprocal –u follows v, but not necessary v follows u–, and implicit connections, where relations are extracted via interactions in the social platform –e.g., if user u posts a message and user v replies to that message, an implicit relation between v and u may be assumed. An example of a social platform that uses explicit reciprocal connections is Facebook via its friendship relations. Twitter, differently, uses explicit non-reciprocal connections via its follower-followee relation; if a user u follows a user v on Twitter, it does not necessarily imply that v follows u. Implicit connections, on the other hand, are more common in forums and blogs, where users post questions, evaluations or opinions, and other users react to the posted content.

- **Activities**: Users may perform different activities and have different goals when participating in a social media platform. In this chapter we mainly focus on five activities: nurturing social connections, discussing about particular issues and topics, asking for information, sharing content and, collaborating with others for certain tasks. Note that the majority of social media may allow performing various of these activities.

- **Types of contents**: The third dimension to categorize social platforms is the type of content that users share between them. Here we distinguish between six main types: text, micro-text, tags, URLs, videos and images. Text and micro-text contents differ on their number of characters. Micro-text is characteristic of micro-blogging platforms, such as Twitter, which allows a maximum of 140 characters in their text messages. Note that, as with activities, many of the existing platforms allow for multiple combinations of these content types, although their focus tends to be on few of them.

According to these three dimensions, social platforms can be described as follows:

- **Forums**: Forums and discussion boards are mainly focused on allowing users to hold conversations and to discuss about particular issues and topics. A user generally posts an comment, opinion or question, and other users reply, starting a conversation. All the posts related to a conversation are grouped into a structure

---

3 http://www.facebook.com
4 http://twitter.com
Table 1 Main characteristics of particular social media.

<table>
<thead>
<tr>
<th>User connections</th>
<th>Forous Q&amp;A systems</th>
<th>Wikis</th>
<th>Blogs</th>
<th>Microblogs</th>
<th>Social networks</th>
<th>Social tagging systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Non-reciprocal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implicit</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nurturing social connections</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussing issues and topics</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asking for information</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sharing content</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Collaborating in tasks</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Micro-text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tags</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URLs</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Videos</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

called thread. The predominant type of content in these platforms is the text generated with the evolution of the users’ discussions. User connections in forums usually are implicit. In general, users are not “friends” with each other explicitly, but connections between them can be extracted from the question-reply chains of their discussions. An example of this type of social platform is Boards.ie\(^5\), a popular Irish public forum board system, which is not restricted to certain topic, and where users discuss about any domain or topic, e.g., politics, sports, movies, TV programs, and music.

- **Q&A systems**: Question Answering (QA) platforms can be understood as a particular type of forums, where the main goal of their users is to ask for information, and therefore discussions are generated around the answers to formulated questions. A popular example of QA system is Stack Overflow\(^6\), where users ask a variety of questions about computer programming. A particular characteristic of Stack Overflow and other QA platforms, is that users can gain reputation points based on the quality of their contributions.

- **Wikis**: The key goal of wikis is to enable collaboration between users in order to create content (ideas, documents, reports, etc.). Users are therefore allowed to add, modify and delete content in collaboration with others. Connections in this type of platforms are generally implicit, and are derived from common editing of

---

\(^5\) [http://www.boards.ie](http://www.boards.ie)
\(^6\) [http://stackoverflow.com](http://stackoverflow.com)
a particular resource: a wiki page. The main type of content generated in wikis is text, but other content types, such as images and URLs, are also quite common. One of the most popular examples of this type of platforms is Wikipedia\(^7\), a wiki with more than 73,000 editors around the world, who have contributed to the creation of a very large open online encyclopedia.

- **Blogs**: Blogs represent a more “personal” type of platform with respect to forums and wikis. When using these platforms, the main goal is to share information, although this often generates discussions. A user does not participate in a blog, but owns it, and uses it to share explanations, opinions or reviews about a variety of issues. Other users can comment about particular blog posts, sometimes generating large discussions. Differently to forums, these discussions are not grouped into threads, but are located under a particular blog post. Multimedia content (photos, videos) are also frequent within this type of platforms. Popular examples of blogging platforms are Blogger\(^8\) and WordPress\(^9\).

- **Microblogs**: Microblogs can be considered as a particular type of blog, where the posted content typically is much smaller. Microblogs are also focused on sharing information, but in this case, information is exchanged in small elements, such as short sentences, individual images, videos, and URLs. As opposed to blogs, microblogs generally allow for explicit user connections, both reciprocal and non-reciprocal. One of the most popular micro-blogging platforms is Twitter, which allows a maximum message length of 140 characters. This limitation forces users to use abbreviations and ill-formed words, which represent important challenges when analyzing sentiments and opinions.

- **Social networks**: The main goal of social networks is to maintain and nurture social connections. With this purpose they enable the creation of explicit, reciprocal relations between users. Most of these platforms also support other types of activities, such as sharing content and enabling discussions. In this sense, users share text, URLs and multimedia content within a platform. Popular examples of social networks are LinkedIn\(^10\), which is focused on professional connections, and Facebook, which tends to be more focused on personal relations.

- **Social tagging systems**: In these platforms, users create or upload content (e.g., images, audios, videos), annotate it with freely chosen words (called *tags*), and share it with others. The whole set of tags constitutes an unstructured collaborative categorization scheme, which is commonly known as *folksonomy*. This implicit categorization is then used to search for and discover resources of interest. In principle, social tagging systems are not conceived for connecting users. Nonetheless, the shared tags and annotated items are usually used to find implicit relations between users based on common interests and tastes. Moreover, tags do not always describe the annotated items, but reflect personal opinions and emo-

---

\(^7\) [http://www.wikipedia.org](http://www.wikipedia.org)

\(^8\) [http://www.blogger.com](http://www.blogger.com)

\(^9\) [http://www.wordpress.com](http://www.wordpress.com)

\(^10\) [http://www.linkedin.com](http://www.linkedin.com)
tions concerning such items [10]. Popular sites with social tagging services are Flickr\(^{11}\), YouTube\(^{12}\) and Delicious\(^{13}\).

Note that our purpose is not to provide an exhaustive categorization of social media, but an overview of the main types of platforms used in the literature to extract and capture opinion and affective information. Other categorizations and platforms exist, such as social bookmarking systems and multimedia sharing sites. In the following subsection, we explain the challenges and opportunities that social media content poses to the extraction and analysis of the above information.

### 3.2 Challenges of Sentiment Analysis in Social Media

Content generated by users via social media in general, and micro-blogging platforms in particular, poses multiple challenges to sentiment analysis [38][60]. In this section we aim to overview and summarize some of these challenges.

- **Colloquial language**: Social platforms, except those targeting professional circles, are commonly used for informal communication. Colloquial written language generally contains spelling, syntactical and grammatical mistakes [73]. In addition, users tend to express their emotions and opinions using slang terms, emoticons, exclamation marks, irony and sarcasm [38]. Processing ill-formed text, understanding the semantics of slang language, emphasizing the detected emotion/opinion level according to exclamation marks, and detecting that the emotion expressed by a user is the opposite than the emotion reflected within the text due to sarcasm, represent difficult challenges for current NLP and sentiment analysis tools.

- **Short texts**: Small pieces of text are typical in micro-blogging platforms, such as Twitter, where a maximum of 140 characters per message is allowed. To condense their messages, users make use of abbreviations (e.g., lol for laugh out loud), ill-formed words (e.g., 2morrow for tomorrow), and sentences lacking syntactical structure (e.g., TB Pilot Measuring up (Time):<1 week from data sharing). The lack of syntactical structure, as well as the appearance of abbreviations and contemporaneous terms not recorded in dictionaries, represent important challenges when attempting to understand the affective information expressed within the texts [60].

- **Platform-specific elements**: Some social platforms have their own symbols and textual conventions to express opinions (e.g., Facebook “likes”, Google+ “+1”, and StackOverflow points to reward high quality answers), topics (e.g., Twitter hashtags), and references to other users (e.g., Twitter @ symbol). To exploit these conventions, sentiment analysis methods and tools have to be adapted [75].

\(^{11}\) http://www.flickr.com

\(^{12}\) http://www.youtube.com

\(^{13}\) http://delicious.com
• **Real-time Big Data**: User generated content coming from popular social media, such as Facebook and Twitter, is characterized by the Big Data challenges, including: *volume* – data size, *velocity* – the speed of change, *variety* – different types of data, and *veracity* – the trustworthiness of the data. Hence, sentiment analysis techniques applied to social media platforms have to deal with: processing massive amounts of data in short periods of time, dealing with the constant emergence of new words and topics, managing data in different formats (text, image, video), and assessing the veracity of data sources [7][8][65]. From these aspects, we highlight the *velocity* aspect, which implies not only to capture and process the user generated content in real time, but also to perform a response (e.g., recommendation, news provision, trending topic detection) as fast as possible, since it is a common demanding functionality from social media users.

4 Sentiment Analysis in Social Streams

Once presented the main Sentiment Analysis tasks and techniques (Section 2), and described the characteristics of user generated content with regards to the expression of personal opinions and sentiments (Section 3), in the subsequent sections we focus on particular problems and applications of Sentiment Analysis in social streams.

4.1 Sentiment Analysis Problems Addressed in Social Streams

Sentiment Analysis is an essential processing task for personalized services that aim to exploit textual content –such as microblog messages and social tags– generated in social streams, since they usually reflect the users’ subjectivity, in terms of opinions and sentiments for certain issues and topics. For such purpose, in addition to the fundamental sentiment analysis problems –such as entity and opinion recognition, and sentiment polarity estimation–, there are aspects that have to be taken into account. User-generated content in social streams presents a number of interesting phenomena, namely opinion spam, user reputation, irony, sarcasm, and emotion dynamics. If we intend to address these issues, we have to go beyond classic text-based opinion mining techniques.

Opinion spam [33] is aimed to disturb the normal behavior in social media services, especially those integrated in recommendation and e-commerce systems, by introducing a bias towards a specific opinion tendency that promotes or demotes an entity (e.g., a product, a service, a brand), or makes users express reviews and opinions in a certain direction. The identification of opinion spam represents a crucial problem for opinion mining and sentiment analysis approaches, which should be able to detect deceptive opinions that try to simulate real user reviews that increase or harm an entity’s reputation [28][47]. In certain media, such as social networks and microblogging platforms, the users’ responses (e.g. by unfollowing contacts,
and posting complaint comments) may represent a valuable source of information to detect spam content.

The writers’ reputation is another important aspect of sentiment analysis of user-generated content. From the point of view of a review site, the higher the reputation of a review author, the more reliable the review can be to other customers, and sometimes vice versa: A review that is seen as reliable by the users can provide high reputation to its author. In this sense, determining the reputation of the authors of a content can be helpful for opinion spam detection. Moreover, some sites have adopted reputation systems as a tool for avoiding or at least discouraging the production of opinion spam. The reputation of users also plays an important role in social networks, since automatically determining the most influential users in a given network can be really valuable [20].

Given the subjective nature of user-generated content, another relevant phenomenon is the existence of irony or sarcasm in the texts. This can constitute a serious problem for many tasks in Sentiment Analysis, like detection of subjectivity and the classification of the polarity of a given opinion, since the explicit text content reflects the opposite of the sentiment really expressed by the writers. Most of published works has focused on the identification of one-liners (jokes or humorous contents in short texts), but there are some researches aimed to extract humorous patterns from longer texts. In a different way, there are also approaches that use results of sentiment analysis in order to detect humor in texts, for example the (negative) polarity of a text has been taken as a feature to retrieve patterns of humorous contents [41], and syntactic and semantic features have been used as indicatives of humor [56], e.g., semantic ambiguity, the appearance of emoticons, idioms and slang language, and the abundance/absence of punctuation marks, to name a few. In the case of social media, certain user responses, such as expressions of amusement and laughing emoticons, may be used as a source for identifying contents with irony and sarcasm, which may be difficult to detect if no additional information apart from the contents themselves exist.

4.2 Applications of Sentiment Analysis on Social Streams

Sentiment analysis over social platforms offers a fast and effective way to monitor the public’s opinions and feelings towards products, brands, governments, events, etc. Such insights can be used to support decision making in a variety of scenarios. In this section we present three scenarios where the extraction and exploitation of affective information from social streams have become key for certain decision making tasks.

- **Sentiment Analysis in politics and e-government**: Designing and implementing a policy at any level of government is a complex process. One of the key difficulties is finding and summarizing public opinions and sentiments. Citizens do not actively participate in e-government portals [43], and policy specialists
lack of appropriate tools to take into account the citizens’ views on policy issues expressed in real time through social network discussions. Governments are currently investing in research and development\textsuperscript{14} to learn about the citizens, by summarizing public opinion via popular social platforms, and to engage them more effectively. One of the key challenges that arises from this scenario is the lack of awareness of the characteristics of those users that discuss political issues in social media, and whether those users really represent the public opinion [24]. Sentiment analysis tools are therefore challenged in this scenario to complement affective information with details about the citizens and organizations behind the gathered opinions. Another common task in which social streams are used as a source of affective and opinion information about politics is the prediction of the outcome and evolution of events, such as elections [45][77], and crises and revolutions [6], as in the case of the Westgate Mall Terror Attack in Kenya [66].

- **Sentiment Analysis in education and e-learning**: Schools and universities strive to collect feedback from students to improve their courses and tutoring programs. Such feedback is often collected at the end of a course via survey forms. However, such methods are too controlled, slow, and passive. With the rise of social streams, many students are finding online social streams as perfect venues in which sharing their experiences, and seeking for peer help and support. To address this issue, educational institutions –such as the Open University\textsuperscript{15}– are working towards the development of platforms that allow capturing and monitoring the students’ sentiment and opinion in open social media groups [76]. The aim is to speed up the reaction to the concerns and challenges raised by students. In this scenario, one of the key challenges that arises is the need for adapting the sentiment and opinion extraction processes to the particularities of the domain. For example, discussions around a World War lecture will generally have a negative connotation. Sentiment analysis tools need to isolate the opinions targeting the logistics of a course, with respect to the opinions targeting themes inherent to the course.

- **Sentiment Analysis in business and e-commerce**: Public as well as corporate social platforms generate major economic value to business, and can form pivotal parts of corporate expertise management, corporate marketing, product support, customer relationship management, product innovation, and targeted advertising. Public social platforms are generally used to monitor public opinion and reputation about brands and products [85]\textsuperscript{16}. Corporate social platforms, on the other hand, are more focused on providing product support and knowledge interchange within a company. One of the new challenges associated with managing these online communities is the ability to predict change in their “health”. Providing owners and managers of the social platforms with early warnings (by monitoring the members’ contributions, opinions, levels of satisfaction, etc.) may facilitate their decisions to safeguard the communities, e.g. by maintaining engagement,

\textsuperscript{14} http://www.wegov-project.eu, http://www.sense4us.eu

\textsuperscript{15} http://data.open.ac.uk

\textsuperscript{16} http://www.brandwatch.com/, http://www.lithium.com/
reducing community churn, and providing adequate support. The identification of sentiments is key as an initial indicator, but it does not necessarily represent the overall health of the community. The challenge of sentiment analysis tools in this scenario is to complement sentiment extraction with techniques for risk detection in the context of business domains, helping owners and hosts to ensure a sustained stability of their communities.

- **Sentiment Analysis in entertainment**: In an online entertainment scenario, it is well accepted that (i) the user’s current mood may affect the type of resource (e.g., a song, a tv series episode, a video game) she prefers to consume at a particular time—partial or completely regardless her personal tastes— and, in the opposite direction, that (ii) emotions evoked by consumed resources may affect the user’s current mood. These facts are the basis for the investigation and development of sentiment-aware engaging services in social media. How user moods and item-provoked emotions can be determined [25], how they can be related each other [88], and how they and their relations can be exploited for user entertainment applications are indeed emerging research topics, such as those addressed in recommender systems [72], and multimedia retrieval and entertainment [23][80].

All these application scenarios come with an additional common challenge: scalability. Social platforms can easily exceed a million users with hundreds of thousands online each day. Content generation may be of Gigabytes per day, and orders of magnitude more data is derived from observing interaction of the users within a system. Existing data analysis approaches, and in particular sentiment analysis tools, currently struggle with these scalability challenges.

5 Discussion

In the previous sections, we have reviewed and discussed the state of the art on sentiment analysis in social streams. We have explained the different types of user generated content existing in social media platforms, as well as some of the most common challenges that this type of content poses when analyzing affective and opinion information. We have described the different problems and tasks addressed in the sentiment analysis research area, as well as the variety of techniques that have been developed to approach them. We have shown examples of applications that use sentiment analysis on social streams to support decision making process in a variety of domains. In this section, we provide an overview of directions that sentiment analysis area is currently following, and what are the main factors driving the research into these directions.

- **Sentiments are dynamic**: Social streams, such as Twitter, may exhibit very strong temporal dynamics with opinions about the same entity or event changing rapidly over time. Since sentiment analysis approaches generally work by aggregating information, a key challenge faced by current sentiment analysis approaches is to detect when new opinions are emerging, so that the new information is not aggre-
gated to an existing opinion for the given entity. For example, the opinion about the Nexus4 smartphone is generally determined based on a set of posts expressing sentiment about this particular device. Opinions about it may change over time, e.g., as new technical problems or bugs are discovered. Sentiment analysis approaches should therefore be able to identify opinion changes for entities and/or events as long as new issues regarding them emerge. An option adopted by several approaches is to define a time-window (minute, hour, day) in which sentiment is aggregated for the particular entity that is being monitored. However, discussions in social media may emerge and spread really fast, or cool down for long time periods. Therefore, assessing the right granularity level is key to not lose relevant information when discussions spike, and not waste resources when discussions about target entities or events are not present [7][39].

• **Sentiments are entity-focused**: As discussed in Section 2, sentiment is generally computed at document and/or sentence level. Multiple sentiments, nonetheless, can be expressed within the same document or the same sentence towards different targets. For example, the post “I love Nexus4 but I don’t like Nexus5 at all!” expresses two different sentiments towards two different targets, the Nexus4 and Nexus5 devices. Additionally, when monitoring the sentiment or particular brands, events or individuals in social media, sentiment analysis approaches should consider if the sentiment of the posts referencing the brand, event or individual do indeed express sentiment towards those entities. For instance, a significant number of negative posts do exist in social streams mentioning the WWF (the World Wildlife Fund) organization, which do not criticize it, but the negative impact of climate change, the danger of extinction suffered by a number of species, and other sustainability issues. Furthermore, approaches in the literature of sentiment analysis have emerged in the last few years that aim to identify sentiment targets within a given text, focusing on entity-level and aspect-level sentiment analysis detection [37][40][54][85], i.e., they first identify the entities and events appearing in the text, and then check the sentiment expressed towards them.

• **Sentiments are semantics-dependent**: Most of existing approaches to sentiment analysis in social streams have shown effective when sentiment is explicitly and unambiguously reflected in text fragments through affective (opinionated) words, such as “great” as in “I got my new Android phone, what a great device!” or “sad” as in “so sad, now four Sierra Leonean doctors lost to Ebola.” However, merely relying on affective words is often insufficient, and in many cases does not lead to satisfactory sentiment detection results [8][27][61]. Examples of such cases arise when the sentiment of words differs according to (i) the context in which those words occur (e.g., “great” conveys a negative connotation in the context “pain” and positive in the context “smile”), or (ii) the conceptual meaning associated with the words (e.g., “Ebola” is likely to be negative when its associated concept is “Virus” and likely to be neutral when its associated concept is “River”). Therefore, ignoring the semantics of words when calculating their sentiment, in either case, may lead to inaccuracies. Recent research in sentiment analysis is therefore focusing on investigating the identification and use of contextual and
Sentiment Analysis in Social Streams

Semantic information to enhance the accuracy of traditional machine learning [59] and lexicon-based approaches [61].

- **Sentiments are domain-dependent**: Sentiment is expressed in social streams within multiple domains. For example, the domain of death is generally more negative than the domain of birth, although both use common terminology, such as hospital, family, etc. Sentiment analysis approaches need to establish the sentiment of the targeted domain to be able to establish the positivity/negativity of the posts. It has been observed that current sentiment classifiers trained with data from one specific domain do indeed fail when applied to a different domain [3]. Similarly, while lexicon-based approaches have a higher tolerance to domain changes, these approaches do suffer when the vocabulary of the domain under analysis is not well covered by the available sentiment lexicon. Given the great variety of topics and domains that constantly emerge in social streams, domain constraints currently affect the applicability of sentiment analysis approaches. Research is currently being conducted to adapt to new domains, by automatically assigning sentiment to terms not previously covered by the lexicons, and by providing dynamic re-training of existing classifiers [9][50][61][65]. There are also recent approaches aimed to generate domain-dependent lexicons, such as that presented in [25]. In that work, automatic lexicons\(^{17}\) with emotional categories for the movie, music and book domains –e.g., gloomy movies, nostalgic music compositions, and suspenseful books– are automatically generated and modeled by exploiting information available in social tagging systems and online thesauri. The terms of these lexicons are also linked to a core lexicon which is composed of weighted terms associated to 16 general emotions –e.g., happiness, calmness, and tension– of the well-known Russell’s circumplex model of affect [58].

- **Sentiments are language- and culture-dependent**: An important problem when analyzing sentiment in social media streams is that posts are written in different languages. Even individual posts may include terminology from a variety of languages within them. Language identification tools are therefore needed to detect the language in which posts are written [11]. An even more complex problem is that sentiment is culturally dependent. The way in which we express positivity or negativity, humor, irony or sarcasm varies depending on our cultural background [22]. Sentiment analysis tools therefore need to account for language and culture variances to provide accurate sentiment identification. Few research works have been recently conducted in this vein, focusing mainly on demographic language variations (e.g., age, gender) of users to improve sentiment analysis performance [79, 74].

- **Sentiments are personality-dependent**: The relationships between emotional states and personality have been a topic of study in psychology in the last twenty years (see e.g. seminal works as [55]). The reader can find more details on this in Chapter ???. In particular, several studies have revealed associations between extraversion and neuroticism (sometimes referred as emotional stability) personality traits with individual differences in affective level and environmental response.

\(^{17}\)http://ir.ii.uam.es/emotions/
This, together with the facts that (i) it has been shown that there exist correlations between user personality traits and user preferences in several domains [26], and that (ii) approaches have been proposed to infer user personality from data about user activity and behavior in social streams [2] (see Chapter ??), raise new research opportunities and applications –such as customer characterization, market segmentation, and personalized recommendation– for sentiment analysis in the context of the Social Web.

References


