Editor: Gustavo Rossi • gustavo@lifia.info.unlp.edu.ar

Harvesting Opinions and Emotions from Social Media Textual Resources

Despoina Chatzakou and Athena Vakali • Aristotle University of Thessaloniki

Harvesting sentiments from social media textual resources can reveal insightful information. The understanding and modeling of such resources are key requirements for accurately capturing the conveyed sentiments. Here, the authors consider multiple approaches, with an emphasis on detecting sentiments in Web 2.0 textual resources.

The amount and diversity of information shared in social media via activities such as posting and commenting sets a fertile ground for harvesting people's opinions and emotions (together known as sentiments). Up to now, many efforts in academia and industry have focused on apprehending people's sentiments expressed in social media. This is due to the pervasive interest of a wide range of stakeholders, such as companies, entrepreneurs, authorities, and the general public.

Two seminal approaches dominate the relevant research bibliography and market applications: *opinion mining* and *affective analysis* (interchangeably both referred to as sentiment analysis). Under opinion mining, texts are analyzed to capture people's opinions, typically falling into the dual polarities of positive or negative, with occasional consideration of a neutral standing. Affective analysis focuses more on people's sentiments, by tracking and revealing their emotions (such as anger, happiness, and disgust).

Revealing human sentiment is quite challenging, due to the non-standard or formal behavioral norm in people's expressions in social media. Because social, cultural backgrounds and demographics are the key factors that affect thoughts and perceptions, people of various origins may express the same sentiment differently. Additional challenges are posed by various linguistic phenomena, which are intense in social media. For example, grammar and syntactic flaws are due to informal and fast writing (for example, the use of several abbreviations) or texts' input limitations (such as limiting the number of characters in Twitter).

Because sentiment analysis methodologies require prior modeling of textual resources, here our work focuses on both key features tailored for social media textual resources modeling and the most popular approaches for sentiment detection in social media textual resources.

Web 2.0 Textual Content Modeling for Sentiment Analysis

Web 2.0 platforms are pervasive textual resource generators. Such textual resources are driven by social interactions and they're typically of a limited size (for example, posts). This is evident in microblogging threads, in comments, or even in metadata (such as tags, short summaries, and descriptive titles) in various social media platforms.

Web textual resources fluctuate from quite small snippets of text, such as a unique sentence (for example, a Twitter post), to a set of sentences that comprise a document (such as a product review). Having such resources at hand, we can view sentiment analysis as a text-processing approach, so a representation model (features' extraction) is required for textual resources' modeling. Complementary to the extensive earlier work in representation models for sentiment analysis,[1–3] we particularly focus on highlighting

 1089-7801/15/$31.00 © 2015 IEEE

popular processes followed in social media textual resources' modeling (see Figure 1). Initially, we outline the social media texts' modeling in a *sentence-based analysis*, and then we present the way that this modeling is used further in a *document-based analysis* (see Figure 1).

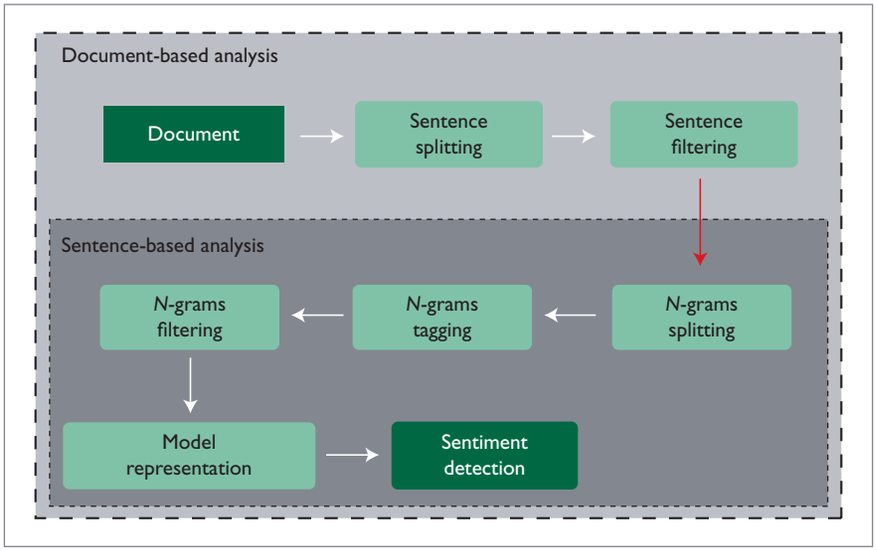

Figure 1. Outline of text modeling processes. Initially, we outline the social media texts' modeling in a sentence-based analysis, and then we present the way that this modeling is used further in a document-based analysis.

Sentence-Based Analysis

The popular bag-of-words (BoW) model is heavily utilized, since it adapts well to textual resources modeling. BoW is easily applied on sentences (being sets of words), while it disregards words' ordering. An alternative approach is the *n*-gram modeling (*N*-grams splitting), with an *n*-gram being a sequence of *n* items (words) of a sentence. Therefore, we model each sentence with a set of (overlapping) *n*-grams. Its simplest form is unigrams with words' unordered listing, which is the same as a BoW model. In general, *n*-grams (apart from unigrams) consider words' ordering in a sentence. Its word-ordering preservation enables understanding of inter-word sentimental influences and interactions.

To filter out words that probably aren't expressing sentiment, we use part-of-speech (for example, adverbs, adjectives, noun, and verbs) tagging (*N*-grams tagging), so that only words with specific tags are preserved for further processing (*N*-grams filtering). Farah Benamara and her colleagues[4] suggest that adjectives and adverbs are good indicators of sentiment.

We can utilize many features to represent the filtered sentences[1] (Model representation), such as the set of observed *n*-grams either unweighted or weighted by their frequency of appearance. As textual resources are characterized frequently by contradictions and different levels of sentiment expressions, we often use additional features such as intensifiers and negation words.[5] The intensifiers affect a sentiment, either by increasing (amplifiers such "very" or "much") or decreasing (downtoners such as "hardly" or "scarcely") its intensity. Negation words affect the polarity of an expression (word/phrase); for example, in the sentence "this is good," the polarity of the expressed sentiment is changed by adding the word *not*, so that it says "this is not good." The final issue to address is the approach to follow for extracting the concealed sentiments (Sentiment detection) with popular approaches being discussed later (see the "Sentiment Detection" section).

Document-Based Analysis

If the textual resource is a document, we can decompose it (Sentence splitting) into a set of sentences or paragraphs (in Web 2.0 textual resources, the *document* and *paragraph* concepts are often identical, so henceforth we'll use the term "document" to refer to a set of sentences for any kind of textual resource, such as a product review). Then we can use a filtering approach (Sentence filtering) to proceed with sentences that might better convey information about the document's sentiment. Such filtering could be based on different aspects of sentences, such as position, length, and subjectivity. For instance, the first and last sentences of a review are often quite indicative about its polarity.

Because a document is a set of sentences, its representation relies on the model derived for each individual sentence (sentence-based analysis). To capture the overall document's sentiment, the sentiments expressed in each individual sentence are typically aggregated using an appropriate measure (such as an averaging operator).

Topic-Based Sentiment Analysis

Often we need to identify sentiments expressed by a textual resource with respect to a specific topic, instead of the overall sentiment expressed in such a textual entity. The topic is the sentence's actual subject for which the sentiment is expressed. In this case, a sentiment detection approach should be able to initially identify the subject and then capture the sentiments expressed with respect to it. To proceed, we follow two steps[1]: first, we identify sentences relevant to the considered topic; and second, we apply the sentence-filtering process

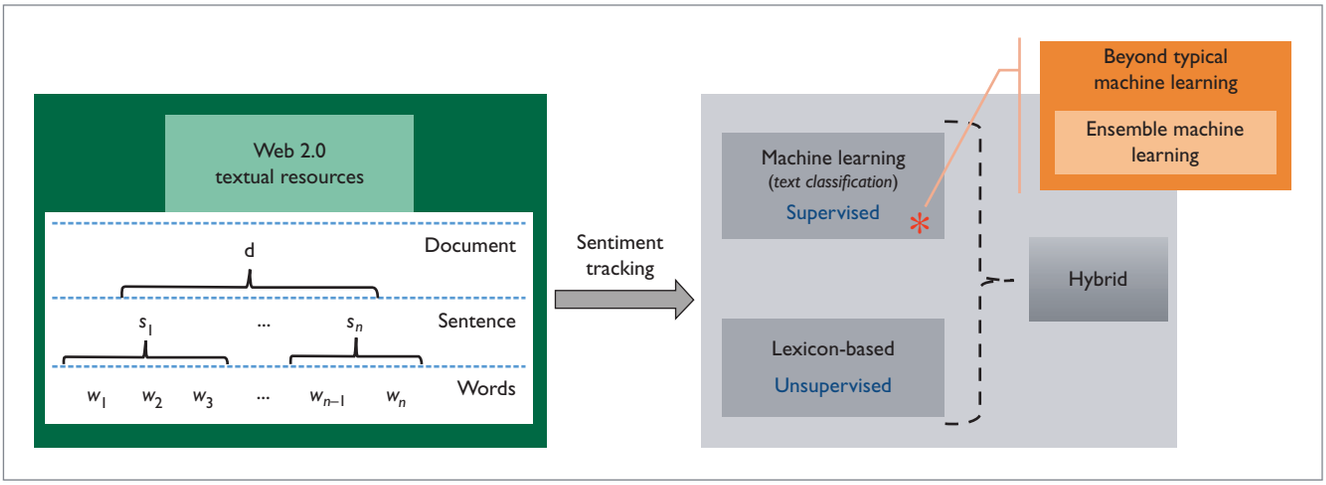

Figure 2. Sentiment detection approaches. We generally use one of the following three approaches: lexicon-based or machine learning approaches, with limited work on hybrid methodologies.

(see Figure 1). Finally, we use a model-representation approach (as previously described) to detect the sentiment of the on-topic sentences (Sentiment detection).

Sentiment Detection

We primarily capture the sentiment of a textual resource by lexicon-based or machine learning approaches, with limited work on hybrid methodologies (see Figure 2).

Lexicon-Based Approach

The lexicon-based approach is an unsupervised technique that identifies the sentiment of a text using lexicons that may be either domain-specific (such as movies, politics, music, or psychology) or domain-independent (see Table 1). Lexicons consist of words/phrases that are characterized by a sentiment value; usually a positive (for example, the score for the word "good" = 0.6) or negative (the score for the word "disgust" = –0.87) value is assigned to each word.

In the *opinion mining approach*, sentiment detection is realized by assigning scores to textual resources. These scores are assigned by utilizing lexicons with such word-scoring capabilities, and each word gets its respective value. *Affective analysis* deals with more fine-grained sentiment detection. Under this approach, a set of primary emotions is predefined (typical primary emotions include anger, disgust, fear, joy, sadness, and surprise[6]). Each of the primary emotions is associated with a further set of emotional words, extracted by lexicons (such as WordNet-Affect; http://wndomains.fbk.eu/wnaffect.html), the so-called secondary emotions (for example, for fear, *horror* and *nervousness* are among its secondary emotions). Then, given a textual resource, words which are in the secondary emotions list are captured and their association to their specific primary emotion is enabled. Finally, each secondary emotion's (lexicon-devised) score along with its other attributes (such as frequency of reference) are used for estimating the overall text's primary emotions' intensity.

Although several well-structured lexicons are available for the English language, the same isn't true for most of the other languages. However, some initial efforts are underway to develop methodologies for identifying sentiments in non-English texts.[7]

Machine Learning Approach

Text classification, a typical supervised machine learning approach, proceeds with learning from past information or experience (training data), to assign new data to a set of specific sentiment categories. In such an approach, the textual resource is split into the training and testing data. Training data (a set of texts with predefined sentiment) are used to identify the properties that are indicative for each sentiment and construct a model. Such a property could be, for instance, the frequent appearance of a word in texts that express a specific sentiment. So, based on the properties identified from the training data, a machine learning approach classifies the so-called testing data — that is, textual resources with unknown sentiment.

As text classification involves the assignment of texts to a number of predefined categories, in opinion mining the texts will be classified either as positive or negative, whereas in affective analysis the categories will be varied based on the primary emotions to be used. The introduction of the machine learning approach in sentiment analysis originates from Bo Pang and his colleagues,[8] where the most commonly used algorithms (see Table 1) are employed — such as Naïve Bayes, the Max Entropy classifier, and support vector machines. Utilizing an individual classifier might result in poor sentiment detection, since the performance of each classifier varies significantly, when for instance someone is using different features or weight measures.

Table 1. Overview of sentiment analysis approaches on social media.

Approach	Sentiment spectrum	Lexicons and classifiers	Social media sources
Lexicon-based	Positive, negative, (neutral)* Anger, disgust, fear, happiness, sadness, surprise, (love) Anger, sadness, joy, disgust Tension, anger, vigor, depression, fatigue, confusion Acceptance, surprise, anger, joy, sadness, anticipation, fear, disgust	SentiWordNet (http://sentiwordnet.isti.cnr.it) WordNet (http://wordnet.princeton.edu) SentiStrength (http://sentistrength.wlv.ac.uk) General Inquirer (www.wjh.harvard.edu/~inquirer) LIWC (www.liwc.net)	Social networks (Facebook, Digg, MySpace) Blogs and Microblogs (Twitter) Content communities (YouTube, Flickr, Vimeo)
Machine learning		Support Vector Machine Naïve Bayes Maximum entropy classifier K-nearest neighbor	

* Throughout the analysis of a textual resource, based on the expressed sentiment, the "sentimental word" presented in parentheses in this column might or might not be used.

Therefore, for overcoming the deficiencies of each classifier and to proceed with a more robust and successful sentiment detection process, we propose *ensemble classifiers* (that is, a combination of multiple classifiers). To optimally combine classifiers, first we estimate each classifier's error with an appropriate measure (with respect to text classification) and afterwards we merge the classifiers' results under a weighting scheme. Rui Xia and his colleagues[9] integrated the aforementioned classifiers[8] and proved the effectiveness of the ensemble technique in sentiment detection.

Hybrid Approach

Lexicon-based approaches suffer from their absolute dependence on lexicons, which are often characterized by words' shortage or inappropriate sentiment values' assignment. Even though machine learning approaches overcome lexicons' limitations, their need for a large volume of past information (training data) to accurately capture the concealed sentiments lessens their advantage.[10]

The hybrid approach targets at solving these limitations by combining lexicon-based and machine learning approaches. An exemplar use case scenario for a hybrid approach is to follow a two-step process to initially generate a set of training data by automatically identifying texts' sentiment score (using a lexicon-based approach), and then to proceed with classification (with a machine learning approach) that's independent from the lexicons' limitations. Hybrid approaches enhance the existing methodologies' stability and accuracy while exploring the strong characteristics of both the machine learning and lexicon-based approaches.

Sentiment analysis targets open issues in various fields (including politics, psychology, and society), because sentiments' understanding can largely impact interactions, policies, and decision making. An indicative example is to study the possibility of using social media textual resources as a substitute or even replacement of traditional polls (sentiments expressed in social media are in accordance with poll's results).[11] We can support arguments' extraction and policy making by capturing and considering citizens' opinions expressed in social media texts.[12] From the sociologists' perspective, it's interesting to see how people's sentiments are shaped in social media and to study their influence on a community's socio-economic well-being.[13]

As new phenomena emerge on Web 2.0, the need for sentiment harvesting and analytics spread also in problems such as the spotting of spammy sentiment information or the (almost) real-time sentiments' detection. People quite often produce and share fake content for the sake of publicity and profitability (for example, to increase a company's profits, widen a famous person's popularity, or generate public interest). Detecting fake sentiment is quite challenging and different from other forms of Web spam (link spam and content spam) or email spam. The latter cases are easier to recognize by visual inspection, whereas opinion spam is harder to detect (apart from the author, no one can infer with certainty whether the expressed sentiment is honest or fake). For example, in a product review, such as "It's an absolutely perfect monitor screen," by reading the content itself, it's almost impossible to conclude about its reliability. Typically, in sentiment spam-detection approaches, along with a machine learning or a lexicon-based approach, additional behavioral characteristics are examined, based on the hypothesis that sentiment spammers differ behaviorally from non-spammers (perhaps the textual resource

is published at an unusual time or a number of textual resources with almost similar content is repeatedly published from the same place, for example).[14]

Effective sentiment discovery in real (or near-real) time has recently received significant attention. The astonishing amount of data flowing through social media (such as Twitter and Facebook) led to the *streaming sentiment classification* process, which extracts sentiments from content arriving in a stream (ordered and potentially unbounded texts' sequences). The formation of a mechanism adaptable to real-time sentimental changes is quite challenging. Albert Bifet and Eibe Frank[15] present a typical methodology that captures sentiments from social media streaming data.

As a promising future direction, we can improve the capturing of texts' sentiments by considering additional features (apart from features extracted from the text's content). Such features could be the social network's structure, temporal correlations, and users' demographics (adding value in the sentiment detection process). In this way, future work could reach a more accurate modeling of human behavior norms and attitudes and leverage it for sentiment detection approaches.

References

1. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1–2, 2008, pp. 1–135.
2. E. Cambria et al., "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, 2013, pp. 15–21.
3. Z. Nanli et al., "Sentiment Analysis: A Literature Review," *IEEE Proc. Int'l Symp. Management of Technology*, 2012, pp. 572–576.
4. F. Benamara et al., "Sentiment Analysis: Adjectives and Adverbs Are Better than Adjectives Alone," *Proc. Int'l Conf. Weblogs and Social Media*, 2007; www.icwsm.org/papers/3--Benamara-Cesarano-Picariello-Reforgiato-Subrahmanian.pdf.
5. D. Chatzakou et al., "Micro-blogging Content Analysis via Emotional-Driven Analysis," *Proc. Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 375–380.
6. P. Ekman, W.V. Friesen, and P. Ellsworth, "What Emotion Categories or Dimensions Can Observers Judge from Facial Behavior?" *Emotion in the Human Face*, P. Ekman, ed., Cambridge Univ. Press, 1982, pp. 39–55.
7. K. Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis," *Proc. Data Eng. Workshop*, 2008, pp. 507–512.
8. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.
9. R. Xia, C. Zong, and S. Li, "Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification," *Information Sciences*, vol. 181, no. 6, 2011, pp. 138–1152.
10. P. Goncalves et al., "Comparing and Combining Sentiment Analysis Methods," *Proc. 1st ACM Conf. Online Social Networks*, 2013, pp. 27–38.
11. B. O' Connor et al., "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *Proc. Int'l Conf. Weblogs and Social Media*, 2010; www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf.
12. C.I. Chesñevar et al., "Integrating Argumentation Technologies and Context-Based Search for Intelligent Processing of Citizens' Opinion in Social Media," *Proc. Int'l Conf. Theory and Practice of Electronic Governance*, 2012, pp. 166–170.
13. D. Quercia et al., "Tracking 'Gross Community Happiness' from Tweets," *Proc. Conf. Computer-Supported Cooperative Work*, 2012, pp. 965–968.
14. B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan and Claypool, 2012.
15. A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data," *Proc. Conf. Discovery Science*, 2010, pp. 1–15.

Despoina Chatzakou is a PhD student in the Department of Informatics at the Aristotle University of Thessaloniki, Greece. Her research interests include social network analysis, sentiment/affective analysis, and spam detection. Chatzakou has an MSc in informatics and management from Aristotle University. Contact her at deppych@csd.auth.gr.

Athena Vakali is a professor in the Department of Informatics at the Aristotle University of Thessaloniki, Greece. Her research interests include Web usage mining, content delivery networks on the Web, social networks, Web 2.0 data clustering, and Web data management on the cloud. Vakali has a PhD in computer science from Aristotle University. Contact her at avakali@csd.auth.gr.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.