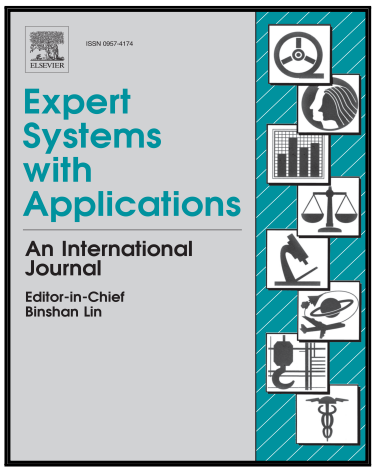


Detecting Variation of Emotions in Online Activities

Despoina Chatzakou, Athena Vakali, Konstantinos Kafetsios

PII: S0957-4174(17)30521-3
DOI: 10.1016/j.eswa.2017.07.044
Reference: ESWA 11460

To appear in: *Expert Systems With Applications*

Received date: 27 December 2016
Revised date: 25 July 2017
Accepted date: 26 July 2017

Please cite this article as: Despoina Chatzakou, Athena Vakali, Konstantinos Kafetsios, Detecting Variation of Emotions in Online Activities, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.07.044

Highlights

- Detect wide spectrum of emotions from online text sources.
- Reveal social emotions and affective states from online social networks content.
- A case study where explicit and implicit emotion experiences are monitored.
- Online text sources features for emotion detection.

Detecting Variation of Emotions in Online Activities

Despoina Chatzakou[a,*], Athena Vakali[a], Konstantinos Kafetsios[b]

[a]*Department of Informatics, Aristotle University, Thessaloniki GR 54124, Greece*
[b]*Department of Psychology, University of Crete, Rethymno GR 74100, Greece*

Abstract

Online text sources form evolving large scale data repositories out of which valuable knowledge about human emotions can be derived. Beyond the *primary emotions* which refer to the global emotional signals, deeper understanding of a wider spectrum of emotions is important to detect online public views and attitudes. The present work is motivated by the need to test and provide a system that categorizes emotion in online activities. Such a system can be beneficial for online services, companies recommendations, and social support communities. The main contributions of this work are to: (a) detect primary emotions, social ones, and those that characterize general affective states from online text sources, (b) compare and validate different emotional analysis processes to highlight the most efficient, and (c) provide a proof of concept case study to monitor and validate online activity, both explicitly and implicitly. The proposed approaches are tested on three datasets collected from different sources, i.e., news agencies, Twitter, and Facebook, and on different languages, i.e., English and Greek. Study results demonstrate that the methodologies at hand succeed to detect a wider spectrum of emotions out of text sources.

Keywords: Emotion detection, Machine learning, Lexicon-based approach, Hybrid process

*Corresponding author
Email addresses: deppych@csd.auth.gr (Despoina Chatzakou), avakali@csd.auth.gr (Athena Vakali), k.kafetsios@psy.soc.uoc.gr (Konstantinos Kafetsios)

Subject: Authors' Response to Reviewers Comments

The authors appreciate and thank the reviewers for their valuable comments, which have considerably contributed in improving the manuscript's content, and structure. Based on these comments the manuscript has been revised to meet reviewers' suggestions. In detail, in response to the reviewers' comments, the following revision actions have been followed.

Revisions in the text are shown using blue (added) for additions and red strike-through for deletions (~~deleted~~). Edited figures or tables are denoted by using blue on the caption title.

Responses to Reviewer #1 comments

Reviewer #1, Comment #1: The format of the reference in the text, I think that they are wrong.
The APA style format is now used for the references.

Reviewer #1, Comment #2: The authors abuse the use of bold text.
An important number of bold texts has been removed from the manuscript.

Reviewer #1, Comment #3: In the first sections, the authors claim the advantages of their methodology, but they don't provide results yet. These sections must be reformatted.
The Previous Work Section has been updated in order to provide better insights about the advantages of this work.
Also, the following paragraph was added in order to clearer demonstrate the advantage of the presented methodology: "In an effort to conclude to the best approach for detecting such extended set of emotions, as already has been stated, both machine learning and lexicon-based approaches are tested either separately

or under a hybrid scheme. Based on the analysis presented in Section 6, where a set of primary emotions is considered, we succeed to improve the overall performance by 13.34% after following a hybrid approach, while we also succeed to detect a wider spectrum of emotions (Section 7) quite satisfactory, considering the existing inherent difficulties when analyzing deeper humans' emotions.".

Reviewer #1, Comment #4: The figures are small and they can't be seen well.
The size of all figures within the manuscript has been increased.

Responses to Reviewer #2 comments

Reviewer #2, Comment #1: The research contribution is not strong enough, as lexicon-based and machine learning methods are widely used in emotions detection from text sources.
Similar to *Reviewer #1, Comment #3* the Related Work section has been importantly updated in order to provide a clearer indication of the paper's research contributions.
Also, the following paragraph was added within the manuscript: "**Summary.** Overall, in this paper, we tackle the problem of detecting humans' emotions by considering users textual online activity. We build on top of previous research works and already existing methodologies, i.e., machine learning and lexicon-based ones, either separately or in a hybrid mode (based on authors knowledge this is the first time that a hybrid process is used in order to detect a set of humans' emotions), geared to detect an extended set of emotions apart from only considering a set of primary ones. More specifically, while in previous works the emphasis has been placed on a set of primary emotions, with a limited consideration of those that characterize general affective states, here we build on top of them by also considering a set of social emotions.".

Reviewer #2, Comment #2: In this paper, original Greek texts were translated into the corresponding English texts using Google Translate tool, which is also commonly used in other research works to detect emotions/ sentiments from text sources.

The next sentence was added within the manuscript: "The selection of the Google Translate tool is made due to its increased use in already existing works, combined with the effectiveness that it has demonstrated in translating foreign texts to the English language.".

Reviewer #2, Comment #3: Translating foreign text into English for sentiment analysis may not work always.

The next paragraph was added within the manuscript: "Overall, the translation process is followed in this work since based on previous research efforts the translation systems succeed not to adversely affect the detection of the sentiments expressed in texts. However, as it is expected, since the detection of humans' emotions is a more complicated process than simply detecting the expressed sentiment, the translation of the foreign texts into the English ones may not always work perfectly. Up to now, although several well-structured lexicons are available for the English language, the same is not true for most of the other languages which makes it difficult to identify the expressed sentiment and emotions out of non English text sources by mainly building upon lexicon-based approaches. So, here, due to the lack of a well-structured and a comprehensive lexicon in Greek language we decided to proceed with the translation of the Greek sources to the corresponding English ones.".

Reviewer #2, Comment #4: Please discuss the Figure 1 in details. It is confusing.

The next paragraph was added within the manuscript: "Figure 1 overviews the features' modeling process followed for both machine learning and hybrid approaches. After the preprocessing of the available text sources, for the machine learning process a set of features is exploited, i.e., lexicon-based, emoticons, and

document feature vectors, which are combined (Hybrid vector generation - ML) in order to proceed with the emotion prediction process. During the hybrid approach apart from the already extracted features based on the machine learning approach, i.e., Hybrid vector generation - ML, both sentimental and emotional features are used (extracted based on a lexicon-based approach) for detecting the emotions expressed in the considered text sources under a hybrid scheme (Hybrid vector generation). Next sections present in detail the processes followed for the emotional detection analysis.". Also, Figure 1 was updated a little bit in order to distinguish better the hybrid vector generation processes among the machine learning and hybrid approaches.

Reviewer #2, Comment #5: In sub-section 5.1, authors applied several classifiers to detect the emotion expressed in new texts using Weka. Authors experimented with Naive Bayes, BayesNet, (C4.5), Random Forest, NBTree, LADTree, SVM and Logistic Regression (LR). So, please include the results of BayesNet, Random Forest, NBTree, and LADTree in Table 4 and analyzed the results.
Table 4 was updated in order to include the results of all algorithms presented in Section 5.1. Also, the weighted average values are considered instead of simple average of precision, recall, and F1-score.

Reviewer #2, Comment #6: For ensemble classifier, please consider bagging and boosting (AdaBoost) algorithms. As you are considering Random Forest. It may increase the precision, recall and F1-score.
The following sentences were added within the manuscript: "Also, bootstrap aggregating ensemble based processes are followed, i.e., bagging with J48 and Random Forest as well, since they often lead to a good performance. Finally, a well known ensemble-based algorithm is the Adaptive Boosting (**AdaBoost**) which extends boosting to multiclass and regression problems. Here, we proceed with the AdaBoost M1 method with the J48, Random Forest, and Naive Bayes classifiers.". Also, the corresponding results were added in Table 4.

Reviewer #2, Comment #7: Please discuss in details, which ensemble classifiers you used for detecting emotion expressed in the text.
Reviewer #2, Comment #8: For machine learning based results over both basic & social emotions in Table 7, did you consider majority voting or weighted majority voting?
The next paragraph (which answers both questions) was added within the manuscript: "Here, different types of ensemble classifiers are tested: (i) J48, Naive Bayes, (ii) J48, NBTree, (iii) J48, LADTree, (iv) J48, LADTree, Random Forest, (v) J48, Random Forest, Naive Bayes, (vi) J48, Random Forest, and (vii) J48, LADTree, BayesNet. For all the above cases the decision is based on both the non-weighted majority vote (**MV**) and the maximum probability (**MP**) scheme. Also, bootstrap aggregating ensemble based processes are followed, i.e., bagging with J48 and Random Forest as well, since they often lead to a good performance. Finally, a well known ensemble-based algorithm is the Adaptive Boosting (**AdaBoost**) which extends boosting to multiclass and regression problems. Here, we proceed with the AdaBoost M1 method with the J48, Random Forest, and Naive Bayes classifiers.".

Reviewer #2, Comment #9: Weka is a collection of machine learning algorithms for data mining tasks, which is very easy to use. In general, it is very common to use Weka for implementing machine learning algorithms in research.
The sentence "For all the empirical parts presented in Sections 6, 7, and 8 we used WEKA data mining toolkit." was replaced with the following paragraph: "For all the empirical parts presented in Sections 6, 7, and 8 we use a widely known machine learning software, i.e., the WEKA data mining toolkit. WEKA was selected since on the one hand it is freely available under GNU General Public License, while on the other provides an easy access to a large collection of different data mining algorithms.".

Reviewer #2, Comment #10: In sub-section 5.3, the proposed hybrid approach extracted sentimental and emotional features by using lexicon-based approach

then machine learning methods used for detecting emotion in a text. But, in the literature both machine learning and lexicon-based approaches were used to detect Ekman's primary emotions in news headlines.

So far, in the emotional analysis related works (based on authors' knowledge) machine learning and lexicon-based approaches have not been used under a combined scheme. Only in sentiment analysis, which shows whether a text is positive or negative, the hybrid process has been used with promising results. In this sense, we decided to consider the hybrid scheme for the emotional analysis too in order to examine whether or not the hybrid process will have a positive impact on detecting humans' specific emotions. The corresponding information was also added within the manuscript: "Even though hybrid approaches have been used so far for detecting people sentiments out of text sources with promising results, e.g., (Prabowo and Thelwall, 2009; Khan et al., 2014; Asghar et al., 2017), this is the first time that a hybrid process is followed in an effort to detect humans' emotions by considering their online social activity.".

Reviewer #2, Comment #11: Table 5 presents the results with the hybrid approach, please consider weighted average instead of considering average of precision, recall, and F1-score.

Table 5 has been updated in order to contain the weighted average values of precision, recall, and F1-score. Also, Table 7 was updated to include the weighted average values.

Reviewer #2, Comment #12: Please follow the same format for all of the references.

The references have been updated in order to follow a uniform format.

1. Introduction

Web 2.0 technologies are increasingly dominating peoples' everyday life, such that constant and evolving digital social interactions are produced dynamically. Their impact in society is evident by the exponential rates of users and their interactions carried out in popular "mega" social networks platforms, e.g., the daily active users of Facebook overcome 1 billion (*Sept. 2016*).[1] Such intense and large scale online presence is characterized by many behavioral norms driven by people's emotions and views. The power of emotion is evident from recent work which documents that contagious effects in online social networks (OSNs) are due to users emotional states which are often transmitted from real to online life (Coviello et al., 2014).

Until recently, emphasis has been placed on capturing human sentiments by detecting positive and negative opinions on various text sources. However, since human emotions are more variable, not necessarily restricted to a dual emotional standing, a more challenging endeavor is to track and reveal wider peoples' emotions such as anger, joy, etc, since they are powerful elicitors and indicators of human motivational and perceptual states (Roseman et al., 1990). In the psychological science, ongoing debates classify emotions in terms of at least two categories:

- the *basic or primary*, i.e., a fixed number of emotions as the ones we experience instantly as a response to a pleasant (or unpleasant) stimulus. A widely recognized approach of Ekman and his colleagues (Ekman et al., 1982) identifies six primary emotions, i.e., 'anger, disgust, fear, joy, sadness, surprise'. Main characteristics of primary emotions are their automatic onset and pervasive impact on individuals' cognitive and behavioral outcomes.

- the *social emotions*, where a person's emotions are influenced by her fellows emotions and impact their emotions too (Parkinson et al., 2005). In-

[1] http://newsroom.fb.com/company-info, 2017

dicative social emotions are the 'rejection', and 'shame' which have been identified as quite important in social interactions (Kafetsios and Nezlek, 2012).

Up to now, efforts in emotion analysis on text sources have mostly focused on detecting emotions of individuals ignoring the social context's influence and impact. This work is motivated by the need to deepen emotional detection by exploiting not only primary emotions, but also other facets of emotions such as social emotions. A full list of emotions from a psychological perspective is not complete without reference to emotions that are neither basic or social but that characterize general individual affective states, such as feeling anxious, calm, and interest, i.e., emotional states that demonstrate a longer duration and cause less intensive experiences (Ekman, 1992). Existing work mainly targets in detecting primary emotions without considering the socials ones, while only limited work has targeted in detecting those that characterize general affective states.

This work proposes an extended emotions analysis approach which incorporates Ekman's primary emotions (enabling comparisons with existing work) together with a wider spectrum of emotions at which social and more general affective states are also considered. We leverage both machine learning and lexicon-based approaches which have dominated the literature on this area so far, under a separate or hybrid scheme. Motivation for building on a hybrid scheme originates from the fact that the existing lexicon-based approaches tend to achieve high precision and low recall (Nie et al., 2015), while machine learning approaches suffer in integrating syntactic with semantic information (El-Alfy et al., 2015). In summary, the main contributions of this work are as follows.

1. We proceed with a hybrid approach which builds upon machine learning and lexicon-based approaches, to detect Ekman's primary emotions. To validate such approach and to compare it to existing work we utilize the

SemEval-2007 Affective text competition dataset.[2]

2. We examine distinct approaches to detect social emotions and those that characterize general affective states in addition to the primary ones. We experiment with a Twitter dataset annotated by a crowdsourcing process.

3. We implement a case study in which explicit (i.e., human reports) versus implicit (i.e., automatic detection of emotions) emotional experiences are monitored to cross validate results. The participants' native language ~~was~~is Greek, and so issues related to the Greek texting habits and the detection of emotions in non-English texts in general are also considered.

4. We share the annotated Twitter dataset at: http://bit.ly/2bLgVUP.

The remainder of the paper is organized as follows. Section 2 reviews literature on emotional analysis. Section 3 proceeds with the data preparation for analysis, while Section 4 presents the used datasets. Section 5 overviews the used methodologies. Sections 6 and 7 proceed~~s~~ with the emotions detection, i.e., the primary ones or the wider spectrum, accordingly. Section 8 presents the case study, while Section 9 concludes the paper.

2. Previous Work

Existing research on emotion detection out of English text sources has utilized various data sources which are summarized in Section 2.1 and heavily depended on machine learning and lexicon-based methodologies which are highlighted in Section 2.2. Also, sentiment detection out of non-English text sources is briefly outlined in Section 2.3. Table 1 shows a comparison of our work to others that are most relevant to our problem setting.

2.1. Text sources for Emotions analysis

Microblogging text sources offer a fertile ground for emotion analyses since they include interactions of emotional value and intensity. In practice, Twit-

[2] http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml

Table 1: Comparison of our work against alternatives

	textual sources		**emotions detected**			**methodology**		
	OSNs	no OSNs	primary	social	affective states	ML	LB	Hybrid
Our work	✓**(twitter, facebook)**	✓**(SemEval)**	✓	✓	✓	✓	✓	✓
(Wang et al., 2012)	✓(twitter)		✓			✓		
(Roberts et al., 2012)	✓(twitter)		✓			✓		
(Farnadi et al., 2014)	✓(facebook)		✓				✓	
(Mohammad et al., 2015)	✓(twitter)		✓			✓		
(Tumasjan et al., 2010)	✓(twitter)		✓		✓		✓	
(Kim et al., 2010)		✓(SemEval, Fairy tales, Isear)	✓			✓	✓	
(Smith and Lee, 2013)		✓(SemEval)	✓				✓	
(Inkpen et al., 2009)		✓(SemEval, Livejournal)	✓			✓		
(Strapparava and Mihalcea, 2008)		✓(SemEval)	✓			✓	✓	

ter text sources have been chosen widely in the emotions analysis literature, e.g., (Wang et al., 2012; Roberts et al., 2012; Mohammad et al., 2015), primarily due to the simplicity and openness of its Application Programming Interfaces (APIs). Other OSNs texts of emotional value such as Facebook have been used more rarely due to their data accessing limitations. As an example, in (Farnadi et al., 2014) authors use Facebook data to study the relations between human primary emotions and demographic characteristics. Also, an important number of studies has targeted in detecting emotions from non OSNs text sources. For instance, a popular competition was held at the SemEval-2007 (Strapparava and Mihalcea, 2007) addressing the problem of correlating Ekman's primary emotions and lexical semantics on news headlines. Thereafter, many other efforts have built upon these news headlines to conduct emotional analysis, e.g., (Inkpen et al., 2009; Smith and Lee, 2013)).

Since online text sources embed a lot of 'noise' (e.g., HTML tags, stop words, etc.), a preprocessing process is necessary. Data preprocessing is extremely important as it can significantly impact the performance (i.e., the accuracy of the extracted results) of an applied analysis (Kotsiantis et al., 2006). Popular preprocessing approaches are the removal of stop words, URLs, and punctuations, e.g., (Roberts et al., 2012; Mohammad et al., 2015), and the tokenization, e.g., (Roberts et al., 2012; Mohammad et al., 2015). Authors in (Wang et al., 2012)

also proceeded with the lowercasing of words and spelling correction.

In the present work, we leverage news headlines provided by the SemEval-2007 competition dataset to initially detect the expressed primary emotions. Concerning the detection of the wider spectrum of emotions, both Twitter and Facebook data sources were are used as they are among the most popular OSNs.[3] Finally, various preprocessing processes were are considered to ensure that meaningful information will be extracted from the texts under consideration.

2.2. Emotion detection methods

Lexicon-based (LB) and machine learning (ML) methods are quite common in emotion analysis. In (Kim et al., 2010) authors experiment with different lexicon-based approaches to detect 4four primary emotions, i.e., anger, fear, joy, and sadness, from various text sources ranging from fairy tales to news headlines (i.e., SemEval data), while in (Farnadi et al., 2014) and (Mostafa, 2013) authors detect emotions out of OSN sources. Finally, in SemEval-2007 competition (Strapparava and Mihalcea, 2007), various lexicon-based approaches were are tested to detect emotions in news headlines.

Machine learning in emotion analysis has been applied with different methodologies, such as Support Vector Machines (SVM), e.g., (Inkpen et al., 2009; Roberts et al., 2012; Mohammad et al., 2015), and Logistic regression (Mohammad et al., 2015). Authors in (Inkpen et al., 2009) consider various ML algorithms to detect Ekman's primary emotions in Livejournal's blog posts and SemEval-2007 data, while in (Roberts et al., 2012) and (Mohammad et al., 2015) SVM was is used to detect emotions on Twitter data. Finally, in (Strapparava and Mihalcea, 2008) both ML and LB approaches were are used (under a separate scheme) to detect Ekman's primary emotions in news headlines.

In text classification tasks, texts are modeled by many tokens (e.g., words, emoticons, punctuation marks) which make the classification process quite a hard process (Zareapoor and Seeja, 2015), e.g., increased time complexity due to

[3] goo.gl/qwoByg, 2017

large data volume. All such tokens can be used as features in a ML classification task. To simplify such a task, a features selection process takes place, i.e., the selection of the particular features that best describe the emotional valued information out of texts. Various features have been used in the emotional analysis tasks, as for instance N-grams, punctuation marks, WordNet synsets, emoticons, and emotional words, ~~(~~e.g., (Strapparava and Mihalcea, 2008; Inkpen et al., 2009; Roberts et al., 2012; Mohammad et al., 2015)~~)~~.

Here, we experiment with both machine learning and lexicon-based approaches, either separately or under a hybrid scheme to exploit their individual advantages towards an overall improved analysis. Even though hybrid approaches have been used so far for detecting people sentiments out of text sources with promising results, e.g., (Prabowo and Thelwall, 2009; Khan et al., 2014; Asghar et al., 2017), this is the first time that a hybrid process is followed in an effort to detect humans' emotions by considering their online social activity. The time complexity of the hybrid approach depends on both the machine learning and lexicon-based approaches. As LB approaches tend to be relatively fast (Augustyniak et al., 2014), the overall computational complexity is mainly in accordance to the ML process to be followed, e.g., Decision Tree classifiers tend to be faster than SVM. Hence, overall, the hybrid approach does not add important additional time complexity in the emotional analysis process. Finally, various feature selections ~~were~~are examined to conclude to those that best describe the available data sources.

2.3. Multilingual textual online content analysis

Much research work in sentiment and emotional analysis builds upon English text content, since mature methodologies and tools have been developed and shared in this language. A practical approach to detect sentiments and emotions out of non English content is to translate foreign text into English. For instance, authors in (Martín-Valdivia et al., 2013) to detect sentiments on Spanish texts at first translate them in English by adopting a machine translation technique. Also, authors in (Balahur and Turchi, 2014) based on three machine

translation systems, i.e., Google Translate[4], Bing Translator[5] and Moses[6], proceed with a sentiment analysis process in 3 languages, i.e., French, German, and Spanish. Similarly, authors in (Mohammad et al., 2016) translate~~d~~ the Arabic texts to English prior to sentiment analysis. Surprisingly, in all previous cases the research outcome is that such automated machine translation systems do not negatively impact on the sentiment analysis process.

In the present research, similar to the previous approaches and for the needs of our case, original Greek texts ~~were~~are translated into the corresponding English ones by using the popular Google Translate tool (as in the (Balahur and Turchi, 2014) work). The selection of the Google Translate tool is made due to its increased use in already existing works, combined with the effectiveness that it has demonstrated in translating foreign texts to the English language. Overall, the translation process is followed in this work since based on previous research efforts the translation systems succeed not to adversely affect the detection of the sentiments expressed in texts.

However, as it is expected, since the detection of humans' emotions is a more complicated process than simply detecting the expressed sentiment, the translation of the foreign texts into the English ones may not always work perfectly. Up to now, although several well-structured lexicons are available for the English language, the same is not true for most of the other languages which makes it difficult to identify the expressed sentiments and emotions out of non English text sources by mainly building upon lexicon-based approaches. So, here, due to the lack of a well-structured and a comprehensive lexicon in Greek language we decided to proceed with the translation of the Greek sources to the corresponding English ones.

Summary. Overall, in this paper, we tackle the problem of detecting humans' emotions by considering users textual online activity. We build on top of pre-

[4] https://cloud.google.com/translate/docs
[5] http://www.bing.com/translator
[6] http://www.statmt.org/moses/

vious research works and already existing methodologies, i.e., machine learning and lexicon-based ones, either separately or in a hybrid mode (based on authors knowledge this is the first time that a hybrid process is used in order to detect a set of humans' emotions), geared to detect an extended set of emotions apart from only considering a set of primary ones. More specifically, while in previous works the emphasis has been placed on a set of primary emotions, with a limited consideration of those that characterize general affective states, here we build on top of them by also considering a set of social emotions.

In an effort to conclude to the best approach for detecting such extended set of emotions, as already has been stated, both machine learning and lexicon-based approaches are tested either separately or under a hybrid scheme. Based on the analysis presented in Section 6, where a set of primary emotions is considered, we succeed to improve the overall performance by 13.34% after following a hybrid approach, while we also succeed to detect a wider spectrum of emotions (Section 7) quite satisfactory, considering the existing inherent difficulties when analyzing deeper humans' emotions.

3. Background and Fundamentals

This section summarizes all fundamental concepts and processes required for some or all of the emotion analytics of this work, with an emphasis on the data preparation, the features modeling for the machine learning and hybrid processes ~~(see Figure 1)~~ and the emotions words specification, to predict then the texts' underlying emotions. Figure 1 overviews the features' modeling process followed for both machine learning and hybrid approaches. After the preprocessing of the available text sources, for the machine learning process a set of features is exploited, i.e., lexicon-based, emoticons, and document feature vectors, which are combined (Hybrid vector generation - ML) in order to proceed with the emotion prediction process. During the hybrid approach apart from the already extracted features based on the machine learning approach, i.e., Hybrid vector generation - ML, both sentimental and emotional features are used (extracted

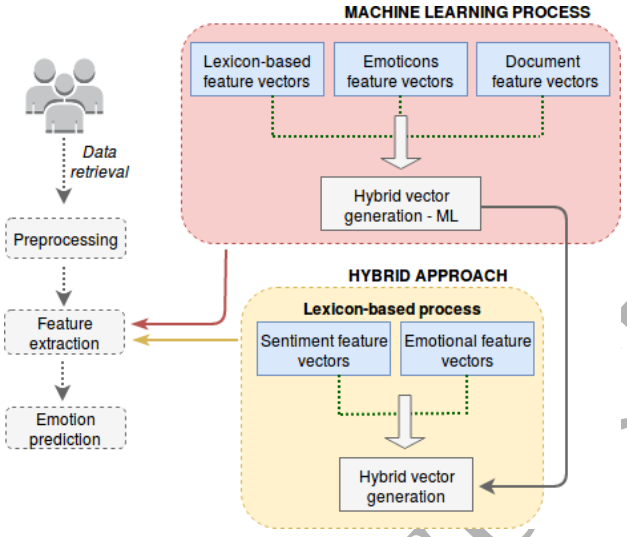


Figure 1: Features modeling for machine learning and hybrid approaches

based on a lexicon-based approach) for detecting the emotions expressed in the considered text sources under a hybrid scheme (Hybrid vector generation). Next sections present in detail the processes followed for the emotional detection analysis.

3.1. Preprocessing Process

Preprocessing tasks, such as cleaning and tokenization, are required to maintain a 'clean' dataset which is free of noisy data and which maintains textual emotional value. Next, we focus on the tasks which have been adopted in this work due to their suitability for emotion analysis overall.

~~**Cleaning.**~~*Cleaning.* The first step in the preprocessing process is to remove the noise from the data. As indicated at previous work (Section 2), initially we remove all stops words, i.e., words which carry no emotion information. These may be pronouns, prepositions, or conjunctions (e.g. 'a', 'is', 'are', 'by', 'for'). Moreover, we remove the URLs, the numbers, and the punctuation marks.

~~**Tokenization.**~~*Tokenization.* The cleaned texts are then tokenized on all

whitespace, i.e., we spit a text into pieces (tokens) using whitespace as separator. Also, we convert all characters of each word to lower case.

~~Stemming.~~*Stemming.* We keep the root (stem) of each word by removing any morphological affixes. For instance, the stemmed word 'love' remains for all of this word's derivatives – loving, loves, etc. Here, we use~~d~~ the Porter Stemmer (Porter, 1980) which has been used in sentiment/emotional analysis processes, e.g., (Khan et al., 2015).

3.2. Feature extraction for the Machine learning process

To proceed with an emotional analysis with a machine learning approach the feature extraction process is mandatory. Various features have been used in the literature (Section 2.2), so we also consider some of them to represent data in a suitable form to proceed then with emotion prediction / detection. Apart from the features presented next, we also experimented with some additional features. For instance, the punctuation marks and the uppercase texts (which can be indicative of intense emotional state, e.g., 'shouting') were also examined, but they were excluded from the analysis as they did not provide any additional value for distinguishing between different human emotions (based on the empirical study conducted in Section 6).

~~Lexicon-based features.~~*Lexicon-based features.* The lexicon-based features build upon emotional lexicons which contain words carrying an emotional connotation. A popular lexical database is the WordNet-Affect (Valitutti, 2004) which assigns a variety of affect labels to a subset of synsets in WordNet.[7]

WordNet is a lexical database in English, which groups words (i.e., nouns, verbs, adjectives, and adverbs) to sets of cognitive synonyms, the so-called synsets, and each synset expresses a distinct concept. For instance, the synonyms of the word 'small' are the words 'little, tiny, and mini', which belong to the same synset. Overall, WordNet facilitates the following processes: (i) indicates the semantic relations between words and synsets, and (ii) groups

[7] http://wndomains.fbk.eu/index.html

words together based on their meaning (synsets). The capturing of semantic information is important to understand the exact meaning conveyed in a text. E.g., 'explosion' is most probably negative when is associated with the concept of war, while will be positive when is associated with the expression of intensive positive emotions. So, ignoring the semantic information when detecting emotions may lead to inaccuracies.

WordNet-Affect is WordNet's extension and includes a subset of synsets which represents affective concepts (i.e., emotional states) correlated with affective words. E.g., the affective concept 'joy' is correlated with the affective words 'amusement, happiness, cheerfulness'. Similar to existing works, e.g., (Strapparava and Mihalcea, 2008; Wang et al., 2012), initially we extract all the WordNet-Affect affective concepts and the correlated affected words to be used then as features in our emotional analysis process.

~~**Emoticons as features.**~~*Emoticons as features.* Emoticons, i.e., pictorial representation of facial expressions, are considered as features since they are quite popular in OSNs and they do carry emotional value. This work exploits this emotion valued piece of information with the use of an emoticons list from the University of Maryland, at which each emoticon is associated with a score on a [-1,1] scale, based on how positive or negative is.[8] Then, all scores ~~were~~are rescaled in the interval [0,1] which shows the intensity of the expressed emotion in relation to a specific emoticon.

~~**Document feature vectors.**~~*Document feature vectors.* Vector space model (Salton et al., 1975) is widely used in information retrieval where each document is represented as a vector and each dimension of such vector corresponds to a separate word.[9] This vector size equals to the number of all unique words of all texts at hand. If a word occurs in the document then its value in the vector is non-zero.

[8] www.umaryland.edu

[9] In our case, the term document refers to the textual source under examination, i.e., tweet, news headline, message exchanged on facebook; so for the rest of the paper the terms *document* and *text* are used interchangeably.

For example, let's consider a dataset, $D = d_1, d_2, ..., d_n$, which contains N documents and a dictionary, $W = w_1, w_2, ..., w_m$, which contains all the unique words of D dataset (in our case, after the preprocessing task). Then a document $\vec{d_i} =< w_{1i}, w_{2i}, ..., w_{ni} >$, where w_{ki} represents the weight of k^{th} term in document i. There are different ways for estimating such weights, as for instance the binary weighting scheme (which considers only the appearance or absence of a word in a document), the term frequency, tf, weighting scheme, or the term frequency - inverse document frequency, $it - idf$, weighting scheme. The term frequency, $tf(w_j, d_i)$, equals to number of times a word, w_j, appears in the document, d_i, while the term frequency - inverse document frequency, $tf-idf(w_j, d_i)$, equals to $tf(w_j, d_i) * idf(w_j, d_i)$, where $idf(w_j, d_i) = \log N/|\{d_i \in D : w_j \in d_i\}|$. In our case, we use the term-frequency weighting scheme since it leads to a better performance based on the study conducted in Section 6.

Such document feature vectors, which permit the capturing of semantic information exists in the written language, have already been used in the emotional analysis process, e.g., (Sreeja and Mahalakshmi, 2016), showing promising results in the effort of detecting emotions from texts and so, ~~were~~are also used in our emotional analysis process.

~~**Hybrid features.**~~*Hybrid features.* Based on previous work (Giatsoglou et al., 2016), here we also proceed with a hybrid features process that considers lexicon-based features, with the emoticons features and the document features vectors. To be able to proceed with such an approach a *vectorization* technique should be applied to transform the lexicon-based and the emoticons features to vectors.

~~*Vectorization of lexicon-based features.* E~~Vectorization of lexicon-based features: each document, d_i, is represented as a vector of size M, where M equals to the number of the emotional words extracted from the WordNet-Affect lexical database. So, $\vec{d_i} =< ew_1, ew_2, ...ew_m >$, where each value of the vector can be one or zero if the emotional word, ew_j, exists or not in the document, accordingly.

~~*Vectorization of emoticons features.* S~~Vectorization of emoticons features: similarly as above, each document, d_i, is represented as a vector of size K, where K equals to the number of emoticons included in the emoticons list. So, $\vec{d_i} =< e_1, e_2, ...e_k >$ where each value of the vector can be (0,1] or zero if the emoticon, e_j, exists or not in the document, accordingly. If an emoticon exists in the document, then the corresponding value will be equal to the score of such emoticon in the emoticons list.

Therefore, after completing the above processes, all documents (i.e., texts) of our dataset are transformed to vectors which can then be combined (i.e., concatenation of the lexicon-based, the emoticons, and the document feature vectors) to conclude to a hybrid feature vector. Such hybrid approach permits the consideration of both emotional (i.e., first two types of features) and semantic (i.e., document feature vectors) information in the emotional analysis process.

3.3. The lexicons-based approach

Here, we focus on the lexicon-based (LB) approach and we initially present the building of a set of representative emotions, while then we describe attributes that will be considered in the emotional analysis via the lexicon-based process.

~~**Building the representative emotions.**~~ *Building the representative emotions.* To build upon a LB approach in an emotional analysis process a mandatory step is the extraction of representative emotions for each one of the primary, social, and the emotions which characterize affective states. The representative emotions further describe and triggered by the primary emotions, the social ones, and those that characterize general affective states (Becker-Asano and Wachsmuth, 2009). For instance, the primary emotion of 'joy' is an instant emotional state, while the representative emotion 'relief' is an emotional state that has been triggered from the 'joy' emotion. In summary, representative emotions are a set of words correlated with a primary, social, or an emotion that characterizes an affective state. So, to capture the emotions exist in a text initially we detect a text's representative words, while then we map the existing

representative words to the primary emotions, the social ones, or those that characterize affective states.

To create such a set of representative words we use the WordNet-Affect lexical database, which contains a list of words that further describe each one of the primary emotions, social ones, or those that characterize an affective state. But as such list is quite limited, we also use the WordNet lexical database to extract the synsets of each word included in the representative set of words.

In addition to WordNet-Affect and WordNet lexical databases, as far as the primary emotions, we also consider the NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2010) corpus for building a set of representative words. EmoLex is an emotional lexicon which contains a list of English words and their associations with eight primary emotions: 'anger, disgust, fear, joy, sadness, surprise, trust, and anticipation'. As it is a quite popular lexicon, especially in the sentiment analysis field, e.g., (Kiritchenko et al., 2014; Tang et al., 2016), and due to its included associations from words to emotions, it has been considered in the process of building a set of representative words for each primary emotion. The same process is not applied for the social emotions or those that characterize general affective states, as the EmoLex lexicon does not consider such emotions throughout its associations.

~~**Contextual valence shifters.**~~*Contextual valence shifters.* Even though some terms in a text tend to be clearly positive or negative, there are those that influence the emotions intensity of such positive or negative words. These words are known as contextual valence shifters and two popular types are the *intensifiers* and *negations* (Polanyi and Zaenen, 2006). By intensifiers we refer to words acting as either amplifiers (e.g., very, much) which increase the intensity of the associated emotion, or as the so called downtowners (e.g., hardly, scarcely) which decrease it. By negations we mean the action of negating an associated emotion. Here, we use a predetermined list of negations, such as "not", "never", "no", "none", "not", "cannot", etc. As contextual valence shifters affect or even alter the intensity of an associated emotion they

have been considered in the applied emotion detection process.

~~**Emoticons.**~~*Emoticons.* As the use of emoticons in computer-mediated communication is a way of expressing emotions (Derks et al., 2008), we associated each primary, social, and those that characterize affective states emotions with the list of emoticons presented in Section 3.2. To map such emoticons to the emotions under examination a manual process was followed.

3.4. Feature extraction for the Hybrid approach

To proceed with the hybrid approach, apart from the features presented in Section 3.2 we also consider the *sentimental* features, which indicate how positive, negative, or neutral a text is, and the *emotional* features, where we consider the intensity of each primary, social, and emotions which characterize general affective states (see Section 5.2 for the features extraction process).

As both the sentimental and emotional features will be combined with the features presented in Section 3.2, similar to the *hybrid features process*, a vectorization technique should be applied.

~~*Vectorization of sentiment features (SF).* E~~Vectorization of sentiment features (SF): each document, d_i, is represented as a vector of size S, where S equals to 2, i.e., to positive and negative scores. So, $\vec{d_i} =< pos, neg >$, where each value of the vector can be [0,1] depending on how positive or negative is the expressed sentiment in the document.

~~*Vectorization of emotional features (EF).* E~~Vectorization of emotional features (EF): each document, d_i, is represented as a vector of size E, where E equals to the number of the emotions under consideration. For instance, considering only the primary emotions, $E = 6$, while if the social and the emotions that characterize general affective states are also considered then $E = 12$. So, $\vec{d_i} =< em_1, em_2, ..., em_e >$ where each value of the vector can be on a [0,1] scale, depending on the intensity of the corresponding emotion in a document.

4. Dataset and Study setups

This work exploits various datasets which are used at its different experimentation phases and setups. Next, we summarize these datasets and study setups to increase comprehension of the next Section's methodologies.

~~**Detecting primary emotions.**~~ *Detecting primary emotions.* The initial study setup serve~~d~~s as a baseline for comparisons with the existing state-of-the-art approaches and prior to proceeding with the wider spectrum of emotions detection. Here, we experiment~~ed~~ with the SemEval dataset due to its popularity in already existing emotional analysis work, e.g., (Strapparava and Mihalcea, 2007; Inkpen et al., 2009; Smith and Lee, 2013). It is available under two separate datasets, i.e., the training set which consists of 250 annotated headlines, and the test set which is comprised of $1,000$ annotated data.[10] The included headlines are annotated in a six scale based on Ekman's six primary emotions.

~~**Detecting wider spectrum of emotions.**~~ *Detecting wider spectrum of emotions.* In this study setup we proceed with the detection of social emotions and those that characterize general affective states in addition to the primary ones. As discussed in the Introduction, the consideration of all such emotions is especially valuable as throughout social communications the emotions of one are importantly affected from those of others. Here, we proceed with a Twitter dataset, where based on Twitter streaming API[11] we randomly collected a set of $3,000$ tweets in English. The random selection of tweets was made to ensure that our method does not only apply to a specific domain or topic of interest.

The $3,000$ tweets were then annotated based on the expressed emotions by following a crowdsourcing process. To perform the tweets annotation we developed a web application at which the users were able to characterize a tweet by selecting only one specific emotion among a list of twelve emotions: i) primary - 'anger, disgust, fear, joy, sadness, surprise', ii) social - 'enthusiasm, rejection,

[10] http://nlp.cs.swarthmore.edu/semeval/tasks/task14/data.shtml
[11] https://dev.twitter.com/streaming/

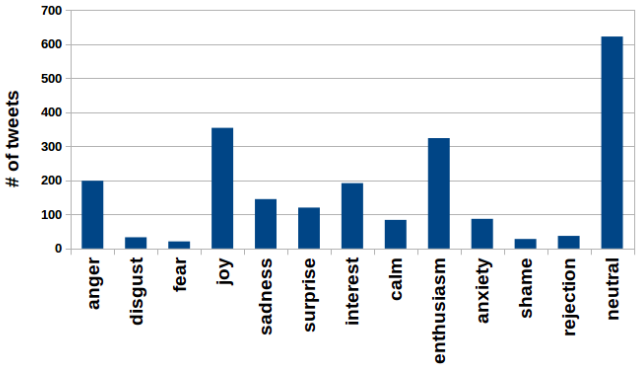


Figure 2: Distribution of emotions

shame', and iii) emotions that express affective states - 'anxiety, calm, interest'. Tweets that did not bear an emotion for the user could be characterized as neutral. Figure 2 shows the distribution of emotion categories as reported by the annotators. We observe that most texts have no emotion (28%) which is quite reasonable as in many cases Twitter serves as a mean of broadcasting news (Bhattacharya and Ram, 2015). Overall, the most common emotions are the joy with 16%, the enthusiasm with 14%, and the emotion of anger with 9%.

Each dataset item (i.e., Twitter post) was annotated by overall 4 annotators, not necessarily by the same ones, to reduce chance of biased annotators in their declaration of emotions. Annotators were instructed to characterize the text sources based on the instant emotion they experienced once they read the text for the first time. The majority vote of annotations was used for creating the final annotation labels. If no majority vote could be determined (i.e., if all annotators assigned a different category), the corresponding tweet was excluded from the dataset, concluding in the end to $2,246$ tweets out of the initial $3,000$ ones. To evaluate the overall degree of agreement among the annotators we calculated the inter-rater agreement using the Fleiss' kappa[12] measure (Fleiss et al., 1971). The overall *kappa* value equals to 26.24%, while the inter-rater

[12] Statistical measure to assess the reliability of agreement between a fixed number of raters.

reliability measure equals to 0.66, which based on (Landis and Koch, 1977) can be characterized as a fair agreement between the annotators. Even though 26.24% it is not a very high agreement, it can be considered quite satisfactory as without the existence of vocal inflections or physical gestures it can be tough to understand the emotions expressed in texts (e.g., sarcasm, cold or serious tone) (Fagerberg et al., 2004). Also, this fair agreement among annotators further indicates the existing difficulties of understanding the underlying emotions out of texts.

5. Emotion detection methodologies

As indicated at the Introduction, lexicon-based approaches tend to result in high precision and low recall, while machine learning approaches do not consider the syntactic and semantic~~s~~ attributes, so both approaches embed emotions misinterpretation risks. Thus, a hybrid process which builds on the advantages of both approaches is proposed. Next, we outline both the machine learning and the lexicon-based approaches, concluded in the end to the hybrid one.

5.1. Emotion Classification with Machine Learning

Here, we outline the machine learning approach, and overview its main processes/phases (Figure 3), which will be followed in both study setups.

In each machine learning approach a training set is required (**PHASE M1**), namely a set of text entities for which the underlying emotion is already known. A machine learning algorithm uses the training data to identify the properties that are indicative for each emotion to predict then the emotions of new texts (Chatzakou and Vakali, 2015). Having preprocessed the available text sources, as presented in Section 3.1, we proceed with the *model representation* phase (**PHASE M2**)~~.~~, which ~~**Model representation.** It~~ involves the features selection for representing the preprocessed texts. Here, we build upon features presented in Section 3.2. ~~**Machine learning algorithm. This**~~ Finally, the *machine learning algorithm* phase includes the selection of the machine learning

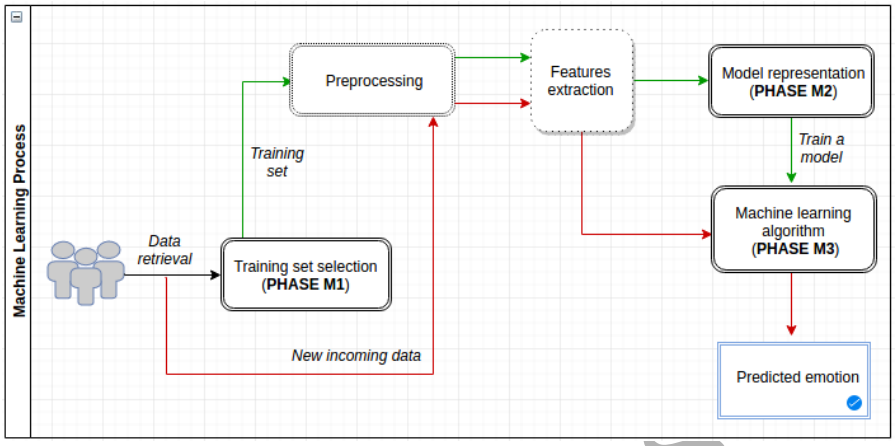


Figure 3: Overall Machine Learning process

algorithm / classifier (**PHASE M3**) which best detects the emotion expressed in new texts. We considered various classifiers, either probabilistic, tree-based, statistical-based, or ensemble ones. For all the empirical parts presented in Sections 6, 7, and 8 we used a widely used machine learning software, i.e., the WEKA data mining toolkit.[13] WEKA was selected since on the one hand it is freely available under GNU General Public License, while on the other provides an easy access to a large collection of different data mining algorithms.

~~**Probabilistic classifiers.**~~ *Probabilistic classifiers.* Probabilistic classifiers (Friedman et al., 1997) are among the most popular classification models. They are based on the Bayes rule to estimate the condition probability of a class label y, based on the assumption that such probability can be decomposed into a product of conditional probabilities:

$$P_r(y|x) = P_r(y|x^1, x^2, ..., x^n),$$

where $x = (x^1, x^2, ..., x^n)$ is the $n-$dimension feature vector.

We experimented with various probabilistic classifiers, such as Naive Bayes (**NB**) and BayesNet. Our results suggest that the former demonstrates the

[13] http://www.cs.waikato.ac.nz/ml/weka/

best performance.

~~Tree-based classifiers.~~ *Tree-based classifiers.* Tree-based classifiers are considered relative fast compared to other classification models (Quinlan, 1986). They consist from three types of nodes, i.e., the *root node*, the *internal nodes*, and the *leaf node*. The *root node* has no incoming edge, and zero or more outgoing edges. Each *internal node* has one incoming edge and two or more outgoing edges. Finally, the *leaf node* has one incoming edge and none outgoing edges. Both the root and each internal node are considered as the feature test conditions, where in the simplest form each test corresponds to a single feature, for distinguishing data based on their characteristics. On the other hand, the leaf nodes correspond to the available classes, i.e., in our case to the emotions under examination.

Here, we experiment~~ed~~ with various tree-based classifiers, e.g., J48, Random Forest, NBTree, and LADTree.

~~Statistical-based classifiers.~~ *Statistical-based classifiers.* This family of classifiers constitutes of a set of widely known classifiers, such as Support Vector Machine (**SVM**) and Logistic Regression (**LR**). Here, we experimented with both such classifiers. SVM separate the different classes of data by a hyperplane, while the objective is to maximize the margin between classes (Burges, 1998). As far as the LR (Kleinbaum and Klein, 2010), is a regression model which learns the conditional distribution $P(y|x)$ by extracting a set of features from the input x, combining them linearly and then applying a function to this combination for predicting the class y.

~~Ensemble classifiers.~~ *Ensemble classifiers.* To overcome the deficiencies of each classifier we proceed with ensemble classifiers, i.e., combination of multiple classifiers (Chatzakou and Vakali, 2015). The decisions of each individual classifier are combined typically by a weighted or unweighted voting. Under the ensemble classifiers the classification results are less dependent on peculiarities of a single classifier as the uncorrelated errors of individual classifiers can be eliminated by averaging (Dietterich, 2000). Here, different types of ensemble

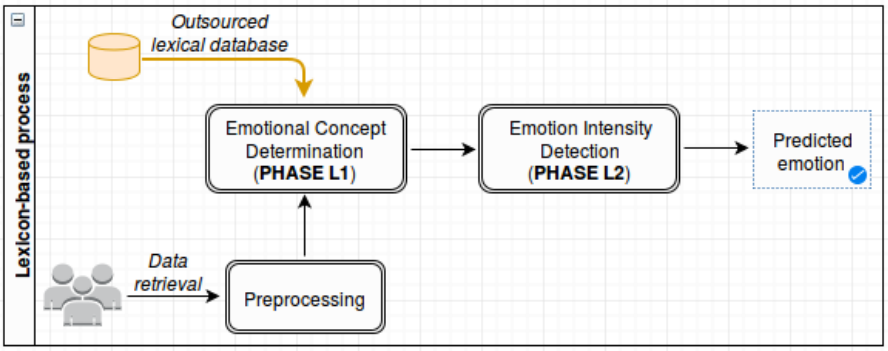


Figure 4: Overall Lexicon-based process

classifiers are tested: (i) J48, Naive Bayes, (ii) J48, NBTree, (iii) J48, LADTree, (iv) J48, LADTree, Random Forest, (v) J48, Random Forest, Naive Bayes, (vi) J48, Random Forest, and (vii) J48, LADTree, BayesNet. For all the above cases the decision is based on both the non-weighted majority vote (**MV**) and the maximum probability (**MP**) scheme. Also, bootstrap aggregating ensemble based processes are followed, i.e., bagging with J48 and Random Forest as well, since they often lead to a good performance. Finally, a well known ensemble-based algorithm is the Adaptive Boosting (**AdaBoost**) which extends boosting to multiclass and regression problems. Here, we proceed with the AdaBoost M1 method with the J48, Random Forest, and Naive Bayes classifiers.

~~**Evaluation.**~~ *Evaluation.* We assess our results based on the *precision* (Prec.),[14] *recall* (Rec.),[15] and *F1-score* (F1). The F1-score, also known as F1-measure, considers both the precision and the recall scores to compute its score[16] and so it shows the balance between them. Thus, the F1-score is used as indicator for evaluating the overall performance of the examined approaches.

5.2. Emotion Extraction based on a lexicon-based approach

Lexicon-based approaches adopt lexicons to proceed with emotion detection by counting and weighting meaningful words, i.e., words that carry emotional information. Contrary to the machine learning approach that just detects the emotion expressed in a text, lexicon-based approaches can also capture the intensity of the underlying emotions. Since human emotions vary greatly in intensity, the exclusion of such aspect from the analysis may lead to misinterpretations of empirical finding (Reisenzein, 1994). To cover such emotions intensity detection, we ~~have~~ also implement~~ed~~ a lexicon-based approach. Based on authors' previous work (Chatzakou et al., 2013), after texts' preprocessing (Section 3.1), next the *emotional concepts determination* (**PHASE L1**) and the *emotions intensity spotting* (**PHASE L2**) phases take place (Figure 4). ~~**Emotional Concepts Determination.** Here,~~In the *emotional concepts determination* we extract the set of the representative emotions (see Section 3.3) to further describe the primary emotions, the social ones, and those that characterize general affective states. ~~**Emotion Intensity Detection.** This~~As far as the *emotion intensity detection* phase, it involves two processes, the estimation of a word's sentiment score and the overall text's emotion. With the word's sentiment score estimation process we detect how positive or negative a word is (for all words of a text), i.e., we capture its intensity, to proceed then with the emotions detection process.

~~**Word's sentiment score.**~~ *Word's sentiment score.* There are various sentiment lexicons, i.e., lists of words and/or phrases with associations to positive and negative scores which indicate the intensity of the expressed sentiments (Kiritchenko et al., 2014). The EmoLex, the MPQA Subjectivity Lexicon (Wilson et al., 2005), and the SentiWordNet (Baccianella et al., 2010) are some indicative examples. Here, we use the SentiWordNet lexicon which is already used in

[14] Precision = #true_positive / (#true_positive + #true_negative)
[15] Recall = #true_positive / (#true_positive + #false_negative)
[16] F1-score = 2*((precision*recall) / (precision+recall))

the literature, e.g., (Rao et al., 2014), which assigns scores to its words based on how positive or negative they are in a [0, 1] scale.

Within a text, apart from the nouns, there are the adverbs, or adverbial phrases (i.e., intensifiers) that strengthen / weaken the meaning of the word in which they refer to and show emphasis, and so they can importantly affect the intensity of the expressed emotion. To include such intensifiers in the estimation of a word's sentiment score, based on the set of intensifiers presented in Section 3.3, the overall sentiment score of a word, et_{ir}, associated with an intensifier, $intens_j$, is as follows:

Definition 1. *The intensifier aware word score.*

$$WScore(et_{ir}) = \|(1 + score(intens_j)) * score(et_{ir})\|$$

Similarly, negations (see Section 3.3) which tend to negate another word are considered in the estimation of a word's sentiment score, as follows:

Definition 2. *The negation inclusion word score.*

$$WScore(et_{ir}) = \|1 - score(et_{ir})\|$$

~~**Overall text's emotion.**~~ *Overall text's emotion.* Following the previous process, we derive the sentiment score of all words within a text, and so, we can proceed with the detection of a text's overall emotions. The followed approach is based on the assumption that the representative emotions can determine the emotions expressed in a text by estimating the similarity between emotional words within a text and the representative words of each emotion.

Based on authors previous work (Chatzakou et al., 2013), the overall score between a text t_i and a emotion e_j is estimated based on the term-frequency (tf) similarity measure for each emotional word et_{ir} of the t_i text. The tf(et_{ir}, e_j) is defined as the number of the representative words of emotion e_j that matches the emotional word et_{ir}. So, the overall score of a text source for each emotion is defined as:

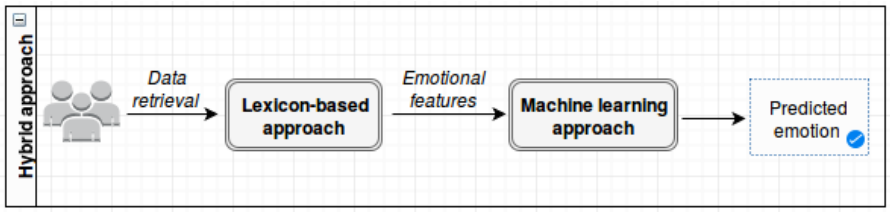


Figure 5: Hybrid process

Definition 3. *Overall score for each emotion.*

$$Score(E_i) = \frac{\sum_{\forall et_{ir} \in ET_i} tf(et_{ir}, e_j) * WScore(et_{ir})}{\sqrt{\sum_{\forall et_{ir} \in ET_i} tf(et_{ir}, e_j)^2}}$$

As in a text multiple emotions can be expressed, we have to conclude to the one that can best describe the text under examination, i.e., predicted emotion. So, the emotion with the highest overall score is considered as the 'predicted' one.

~~**Evaluation.**~~ *Evaluation.* Similar to the machine learning process, here, the precision, recall, and F1-score are used for the evaluation process.

5.3. A hybrid approach for emotion detection

To overcome deficiencies of the lexicon-based and machine learning approaches, we proceed with a hybrid process (Figure 5). Our intuition is that the intrusion of sentimental and emotional features extracted with a lexicon-based approach into the machine learning process will permit the more accurate detection of the expressed emotion in a text (**Testing hypothesis**).

Initially we proceed with a lexicon-based approach to extract two types of features, i.e., the sentimental and the emotional ones (Section 3.4). So, in this case, the lexicon-based approach is used for the feature extraction process and not for detecting a text's underlying emotions. To extract the sentimental features we follow a lexicon-based approach similar to the one described in Section 5.2. Instead of using SentiWordNet as a lexical source, we also experimented with additional lexicons, i.e., the Opinion Lexicon (Hu and Liu,

2004), the Multi-Perspective Question Answering (MPQA) subjective lexicon, the Harvard General Inquirer (Stone, 1966), and the Sentiment140 lexicon (Mohammad et al., 2013). We concluded to Sentiment140, which includes 62,468 unigrams, 677,698 bigrams, and 480,010 non-contiguous pairs, as by using it we succeeded the best F1-score, based on the *primary emotions detection* study setup (Section 6).

Both sentimental and emotional features (after the vectorization process) combined with the features presented in Section 3.2 are used in the machine learning process to finally detect the emotion of a new text.

6. Study I: primary emotions detection

The first study setup serves as a baseline as we compare our methodologies with the already existing ones, by using Ekman's primary emotions, i.e., anger, disgust, fear, joy, sadness, and surprise. Initially, we experiment with the lexicon-based approach, followed by the machine learning process, to finally conclude to the hybrid one. Similar to the SemEval-2007 Affective text task, our objective is to classify news headlines extracted from news websites.

6.1. Results based on a lexicon-based approach

We build upon two processes to extract the set of representative words for each primary emotion (Section 5.2), i.e., the WordNet-Affect, either combined (WN-EL) with the EmoLex corpus or not (WN). Table 2 shows that the WF-Affect lexicon itself achieves better recall in most cases, while when it is combined with the EmoLex corpus there are cases with a higher value in precision. Overall the F1-score in almost every emotion under examination is higher when following the WN process, so in the next experiments we proceed with the representative words extracted with the WordNet-Affect lexical database.

To evaluate the performance of the followed lexicon-based approach we compare our results with previous works: (i) SemEval-2007 Affective-task (Strapparava and Mihalcea, 2007) competition, and (ii) Strapparava, et. al, work (Strapparava and Mihalcea, 2008) where various approaches are examined. All the

Table 2: Results based on a different list of representative words

	anger		disgust		fear		joy		sadness		surprise	
	WN	WN-EL	WN	WN-EL	WN	WN-EL	WN	WN-EL	WN	WN-EL	WN	WN-EL
Prec.	23.26	**25.00**	**62.50**	5.97	**89.66**	81.25	**73.77**	72.34	60.42	**69.49**	46.67	**50.10**
Rec.	**15.87**	12.70	19.23	**46.15**	**16.77**	8.39	**12.47**	9.42	**28.43**	20.10	3.63	3.63
F1	**18.87**	16.84	**29.41**	10.57	**28.26**	15.20	**21.33**	16.67	**38.67**	31.18	6.73	**6.76**

Table 3: Overall average results for all the systems under examination

	Prec.	**Rec.**	**F1**
SWAT	19.46	8.61	11.57
UA	17.94	11.26	9.51
UPAR7	27.60	5.68	8.71
WN-AFFECT PRESENCE	38.25	1.54	4.00
LSA SINGLE WORD	9.88	66.72	16.37
LSA EMOTION SYNSET	9.20	77.71	13.38
LSA ALL EMOTION WORDS	9.77	**90.22**	17.57
OUR APPROACH	**59.38**	16.07	**23.88**

presented approaches ~~were~~are evaluated on the test set of 1,000 newspaper headlines (Section 4). From Table 3 we observe that our lexicon-based approach achieves the best precision, **59.38%**, while it is importantly behind from the LSA ALL EMOTIONS WORDS approach in terms of the recall. In overall, we succeed the best F1 score with an increase of **6.31%** from the second best approach.

6.2. Results based on a machine learning process

With the hybrid features presented in Section 3.2, we evaluate the classifier families presented in Section 5.1, either individually or as part of an ensemble. Specifically in the *documents features vectors* we use~~d~~ the tf weighting scheme for estimating the weights of each term, w_{ki}, in a document, d_i, as it leads to a better F1 score in relation to the binary and $tf - ifd$ weighting schemes.

Here, we present the results for all tested~~the best performing~~ classifiers. In the training process we use~~d~~ the 250 headlines from the SemEval-2007 dataset,

Table 4: Initial results of emotional classification

	Avg. Prec.	**Avg. Rec.**	**Avg. F1**
NB	34.20%	39.20%	33.70%
BayesNet	13.00%	36.00%	19.10%
SVM	31.80%	36.50%	30.70%
LR	34.00%	34.80%	**34.20%**
J48	33.20%	39.80%	30.50%
RF	34.40%	37.50%	31.00%
NBTree	33.50%	36.70%	20.50%
LADTree	53.70%	40.00%	27.50%
J48 Bagging	33.60%	40.20%	30.10%
RF Bagging	32.00%	38.10%	28.90%
J48 AdaBoost	33.20%	39.80%	30.50%
RF AdaBoost	37.70%	40.00%	30.70%
NB AdaBoost	31.20%	35.20%	31.50%
J48, NB - MV (MP)	34.70% (35.80%)	39.10% (39.60%)	32.40% (33.30%)
J48, NBTree - MV (MP)	35.00% (33.20%)	37.40% (39.60%)	26.00% (30.10%)
J48, LADTree - MV (MP)	39.90% (34.40%)	39.60% (39.90%)	29.70% (30.10%)
J48, LADTree, RF - MV (MP)	38.40% (34.40%)	40.00% (40.40%)	29.70% (31.40%)
J48, RF - MV (MP)	32.40% (34.20%)	37.90% (**40.60%**)	30.40% (31.70%)
J48, RF, NB - MV (MP)	35.20% (35.90%)	40.30% (39.80%)	33.60% (33.60%)
J48, LADTree, BayesNet - MV (MP)	**54.40%** (34.40%)	40.01% (39.90%)	27.50% (30.10%)

and for testing the rest 1,000 headlines from the same set of data.

From Table 4 we observe that the best precision is obtained with the ensemble classifier 'J48, LADTree, BayesNet - Majority Voting' classifier, ~~33.98%~~**54.40%**. The ensemble approach (J48, RF - Maximum Probability) results to the best recall, ~~27.02%~~**40.60%**, while ~~also~~the Logistic Regression classifier outperforms in respect to the F1-score, ~~26.28%~~**34.20%**.

6.3. Hybrid approach - Hypothesis testing

In this section, the hybrid approach, and consequently the main hypothesis, is tested. Regarding the machine learning process, we use the model extracted by the ensemble classifier, while we build upon the lexicon-based approach which uses the WordNet-Affect lexicon to extract the set of representative words.

Table 5: Overall results with the Hybrid approach

	Avg. Prec.	Avg. Rec.	Avg. F1
EF + J48 + RF - Maximum Probability	43.70%	**44.90%**	38.10%
SF + J48 + RF - Maximum Probability	42.50%	35.10%	31.70%
EF + SF + J48 + RF + Maximum Probability	**47.50%**	42.00%	**42.40%**
NB TRAINED ON BLOGS	12.04%	18.01%	13.22%
CCG-BASED SYSTEM	42.68%	23.70%	28.97%

Table 5 shows the weighted average precision, recall, and F1-score. In overall, we obtain the best performance when we have both sentiment and emotional features as input in the machine learning process, i.e., ~~34.47%~~**42.40%**. Despite the fact that without the consideration of the emotional and sentiment features the best performance is obtained by simply using the Logistic regression algorithm (**34.20%**), under the hybrid scheme the best performance is achieved by following an ensemble approach (Table 5 shows the best obtained results after experimenting with the different machine learning approaches). Comparing the hybrid approach with our lexicon-based and machine learning processes, we observe that in overall there is an increase of ~~8.19%~~**8.20%** in the F1-score. So, with the hybrid approach the overall performance of the emotional detection process is significantly improved, which confirms the hypothesis testing.

Compared with the already existing systems, we observe that the hybrid approach ~~even though is falling behind from the CCG-BASED SYSTEM in terms of the precision,~~ has an increase of **4.82%** in the precision value, ~~10.7%~~**18.3%** in the recall, while also the overall performance predominates by ~~5.5%~~**13.43%**. Deepening even more in each primary emotion (Table 6), we observe that almost in all cases (except of the anger emotion) the F1-score is quite satisfactory. Joy, followed by the emotion of sadness performed the best with ~~52.60%~~**56.20%** and ~~45.45%~~**46.00%** F1 score, respectively. The emotion of anger yields the worst results in relation to all other emotions with only ~~14.00%~~**7.60%** F1 score, which is also the case when we only use~~d~~ the lexicon-based approach (Table 2). Concerning the ~~fear and~~ disgust emotion~~s they~~ it ha~~d~~s a moderate performance

Table 6: Performance over the six emotional categories (EF + SF + J48 + RF - MP)

	Prec.	**Rec.**	**F1**
Anger	18.80%	4.80%	7.60%
Disgust	37.50%	24.00%	29.30%
Fear	52.00%	25.30%	34.10%
Joy	65.60%	49.20%	**56.20%**
Sadness	44.50%	47.50%	**46.00%**
Surprise	24.30%	50.80%	32.80%
Avg.	47.50%	42.00%	42.40%

with ~~29.60%~~**29.30%** F1 score.

The results obtained either with the lexicon-based, machine learning, or hybrid approach highlight the difficulties exist in detecting with high accuracy human emotions out of texts. This can be justified by the fact that in the written word we cannot easily consider important aspects of humans expressions, such as the sarcasm or the more serious or cold tone. Such aspects are mainly transmitted to the interlocutor via the vocal tone which is characterized from the corresponding emotional nuance. Despite the above obstacles, in overall, the results obtained with the hybrid approach indicate that if we succeed to improve each approach separately the overall performance will be further improved. For instance, the intrusion of more semantic aware features in the machine learning approach (Chatzakou et al., 2015) ~~can~~ may will lead to an enhanced performance.

7. Study II: capturing a wider emotions spectrum

In the second study setup we proceed~~ed~~ with the detection of a wider spectrum of emotions by considering social emotions and those that characterize general affective states in addition to the primary ones. For the experiments reported in this section we use the dataset collected from Twitter.

Due to the confirmation of the hypothesis testing (see Section 6.3), here we proceed with a hybrid approach for the emotions detection process. Initially,

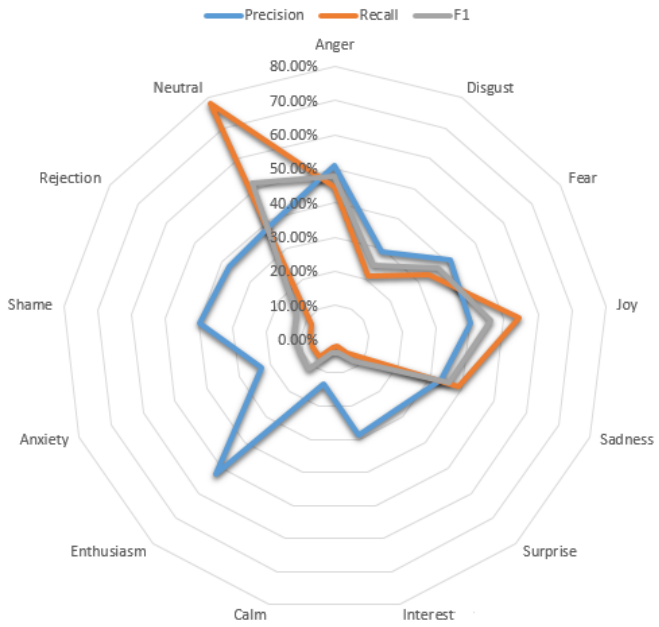


Figure 6: Results of a lexicon-based approach upon 12 emotion categories

we present the results obtained with the lexicon-based approach as its sentiment and emotional features will then be used as input in the machine learning process, while then we proceed with the evaluation of the hybrid approach.

7.1. Emotions detection with the Lexicon-based approach

Based on the process presented in Section 5.2, first we build the set of the representative words for the social emotions and those that characterize general affective states (use of the WordNet-affect lexical database, Section 3.3) in addition to the primary ones. Then, a lexicon-based approach is followed to detect the wider spectrum of emotions out of tweets. Figure 6 depicts the precision, recall, and F1-score.

We observe that the emotion of *anger* performs best with **47.72%** F1 score, while the *interest* emotion shows the lowest performance, **3.88%**, followed by the *calm* emotion. The low performance of both such emotions could suggest that the sets of representative words for the affect states *interest* and *calm* are

not well defined and therefore, a more thorough analysis should be carried out. The overall performance is **23.61%**, with **35.08%** precision, and **23.74%** recall, which can be considered as quite satisfied as it is also in alignment with the human annotations who had shown difficulties in clearly determine the dominant emotion in the considered texts (based on the inter-rater agreement score, Section 4). In addition to the inherent difficulties exist in automatically recognizing a wider spectrum of emotions out of text sources, further difficulties are posed from the lack of structured written word, e.g., grammar and syntactic flaws due to informal and fast writing, in social networks and especially in Twitter (i.e., limited size of posted texts).

7.2. Detecting emotions with the Hybrid approach

In this part of the analysis we experimented with the best performing machine learning algorithms (based on the first experimentation setup, Section 6.2) in conjunction with sentiment and/or emotional features extracted with the lexicon-based approach. By applying the hybrid approach for detecting the wider spectrum of emotions we will also test whether the hypothesis testing (Section 5.3)) is applied in this case too, i.e., similar to the primary emotion detection process. Table 7 shows the results emerged with a 10-fold cross validation process in the whole Twitter dataset. From this table we observe that the best precision is achieved with the Logistic Regression algorithm after the inclusion of both sentiment and emotional features in the machine learning process, while the best recall value is obtained based on the J48 algorithm combined with the emotional features. The best overall performance, **34.50%**, is obtained after the inclusion of only the emotional features in the machine learning process and under an ensemble scheme. The best performance, **23.15%**, is obtained after the inclusion of both sentiment and emotional features in the machine learning process and under an ensemble scheme.

In overall, the performance of the wider spectrum of emotions is falling behind by **7.9%** compared to the first experimental setup where only a set of primary emotions is considered (that is 42.40%). This can be justified by the

Table 7: Machine learning based results over both basic & social emotions

	Avg. Prec.	**Avg. Rec.**	**Avg. F1**
EF + NB	32.90%	20.00%	22.30%
EF + J48	34.70%	**39.00%**	34.20%
EF + SVM	34.00%	36.30%	33.80%
EF + LR	37.30%	38.80%	32.20%
EF + J48 + RF - MP	34.60%	37.90%	**34.50%**
SF + NB	32.20%	32.20%	32.00%
SF + J48	29.90%	34.50%	30.40%
SF + SVM	33.40%	35.80%	32.20%
SF + LR	35.70%	38.00%	32.50%
SF + J48 + RF - MP	30.50%	34.80%	31.00%
EF + SF + NB	33.70%	22.10%	23.40%
EF + SF + J48	32.80%	37.30%	33.50%
EF + SF + SVM	34.50%	36.80%	34.20%
EF + SF + LR	**40.10%**	38.80%	32.20%
EF + SF + J48 + RF - MP	33.00%	37.40%	33.70%

fact that the lack of a sufficient number of texts for every emotion can result to less satisfactory classifiers (Provost, 2000). More specifically, from Table 8 we observe that the best F1-score is achieved with the *anger* and *joy* emotions, which based on Figure 2 are among the emotions with the highest number of instances in the dataset, while the *shame* emotion, which has a quite limited number of instances, achieves the lowest F1-score. Concerning the *fear* and *disgust* emotions even though there is a limited number of instances within the dataset, their performance is much better and also quite similar to the F1-score obtained with the lexicon-based approach. Such a result confirms that the existence of a well-defined set of representative words for each emotion can importantly affect the whole emotion detection process.

The difficulties of detecting a wider spectrum of emotions from texts is not only reflected in the average F1-score, but is also highlighted in the fair agreement between the annotators through the crowdsourcing process. The tendency of achieving low values in the F1-score (e.g., in *calm* and *shame* emotions) is more intensive in the case of the social emotions and those that characterize general affective states. Apart from the reasons presented previously (i.e., not

Table 8: Performance of the 12 emotional categories (EF + J48 + RF - MP)

	Prec.	**Rec.**	**F1**
anger	40.90%	39.70%	**40.30%**
disgust	33.30%	12.10%	17.80%
fear	26.30%	23.80%	25.00%
joy	46.80%	55.90%	**51.00%**
sadness	36.80%	33.80%	35.30%
surprise	48.90%	18.30%	26.70%
enthusiasm	31.00%	21.90%	25.70%
rejection	14.30%	2.70%	4.50%
shame	0.00%	0.00%	0.00%
anxiety	13.20%	8.00%	10.00%
calm	6.30%	1.20%	2.00%
interest	18.60%	9.40%	12.50%
Avg.	34.60%	37.90%	34.50%

well-defined set of representative words and low number of instances within the dataset), this tendency can also be justified by the fact that social emotions are importantly dependent on social appraisals and concerns, i.e., they depend on others' thoughts, feelings, or actions as instantiated, in contrast to the primary emotions (Hareli and Parkinson, 2008). So, to better convey the wider spectrum of emotions we should also consider a user's social network by analyzing the type of relations among the involved members within a conversation, as well as the emotions exhibited by all members with close relations, e.g., members with whom a user has recently communicated with. Finally, to improve the overall performance of the machine learning process a more well-established ground truth dataset should be constructed by further improving the crowdsourcing process. For instance, the use of widely recognized crowdsourcing platforms (e.g., CrowdFlower[17], Amazon mechanical turk[18]), the selection of the highest rated annotators (based on such platforms), or the further filtering of the annotators (e.g., removal of annotators who completed too quickly the annotation

[17] https://www.crowdflower.com/
[18] https://www.mturk.com/mturk/welcome

process), will help to build a more 'correct' ground truth dataset.

8. Case study: correlations between implicit and explicit emotional states

Previous study setups build upon datasets where independent annotators identified the underlying emotions. Since humans perceive emotions in a subjective manner mis-annotations are highly possible, so, here, we proceeded with a case study at which we monitored participants' emotions both implicitly and explicitly. Our objective is to examine the existence of possible similarities or differences among participants explicit emotion declaration and the emotions as these are identified with an automatic emotional analysis process. Next we briefly describe the input and output of the case study.

Input. Twofold input: (i) participants' texting conversations on Facebook, and (ii) the experienced emotions in such conversations explicitly declared by them.

Output. The expected output is to investigate whether similar patterns exist among implicit and explicit emotions declarations.

8.1. Participants and procedure

The study had two parts. *In the explicit emotion description part*, participants described their emotion experience while interacting online using the following emotional states: anger, anxiety, calm, disgust, enthusiasm, fear, joy, interest, rejection, sadness, shame, and surprise, similar to the emotions analyzed in Section 7. We developed a platform at which participants for every texting interaction on Facebook they had to declare the time, the duration of such interaction, and the experienced emotion by selecting from the predefined list of emotions. Interactions lasting less than 10 minutes were not recorded to avoid burdening participants and also to increase likelihood of detecting emotional events. Participants were instructed to report the experienced emotions the quicker the possible after an interaction was conducted to avoid problems of trying to recall an event.

In the implicit emotion detection part, participants' provided access to their personal account on Facebook and so, we had access to the inbox and chat messages. Such exchanged messages were then used to detect the experienced emotions for the same time period that the participants explicitly declared their emotions. Overall, the available information at hand was: (i) the exchanged message itself, (ii) publishing and exchanging date/time, and (iii) sender of the message. All messages were maintained into a database in an encrypted form following standard privacy practices. Participants read a page of debriefing the whole process and they informed about the privacy policy concerning their private data management.

The recruitment of the participants was based on two different pathways. The first one included students from the Aristotle University of Thessaloniki (computer science department) and the University of Crete (department of psychology), where both are established in Greece. Additionally, we advertised the study to the webpage of our research group in the Aristotle University to reach people of different ages and educational backgrounds. In total 36 persons participated in both the implicit and explicit parts of the case study ranging from 19 to 35 years old. Five of them were excluded from the analysis as they had limited interactions. So, the final sample consisted of 31 persons in total, where 24 of them were females and 7 males. The experiment lasted 10 days in order to conclude to a sufficient number of data, either concerning the explicit emotion declarations or participants' texting activity on Facebook, and to also ensure participants' active participation during the entire study.

8.2. Overcoming language barriers

Participants' native language was Greek, so two main stages of data processing took place before the beginning of the research, i.e., *transliteration* and *translation*.

~~**Transliteration process.**~~ *Transliteration process.* Transliteration is the conversion of the Greeklish text to the Greek language, i.e., the practice of transforming a written text from one writing system to another considering all

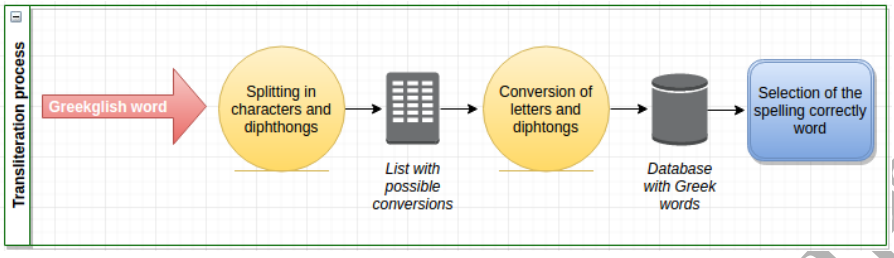


Figure 7: Transliteration process

the rules that are related to such a procedure. Two main rules are the *vocal* and the *spelling*. For instance, considering the vocal rules the Latin letter "o" can be used for representing the Greek letters "ο, ω". With the above example we observe that the letter conversion does not build upon the correct spelling as the only constrain is to *sound the same*. In the spelling case, which requires the spelling to be correct, the Greek letter "ω" will be solely represented by the Latin letter "w" and not with the Latin letter "o" as previously indicated. Table 9 shows some representative examples of spelling and vocal letter conversions.

Table 9: Indicative examples of Latin to Greek conversion

Latin letter	**Greek letter**
e	ε, αι
i	ι, η, υ, ει, οι
h	η, χ
o	ο, ω
x	χ, ξ

To transform all the Greeklish texts to the corresponding Greek ones, we developed an automatic transliteration system (Figure 7).

In the transliteration process we consider~~ed~~: (i) all the possible conversions of each single letter, and (ii) the existing diphthongs, i.e., the combination of two vowels which sound as one and the second vowel is always 'ι' or 'υ'. E.g., indicative diphthongs are the 'ει', 'οι', and 'ου'. After converting a Greeklish word with all possible ways (e.g., the possible conversions of the word 'paizo'

are the words: 'παιζο', 'παιζω'), then we should select the right one, i.e., the one that is spelling correctly (in our example is the word 'παίζω'). To proceed with the right selection we rely on the Wikipedia Greek database, which consists from a large number of Greek words.[19]

~~**Translating process.**~~ *Translating process.* After the transliteration process, we translate the Greek texts to English by using the Google Translate API which gives reliable results on sentiment analysis (Balahur and Turchi, 2014).

8.3. Understanding of tendency via analyzing Facebook conversations

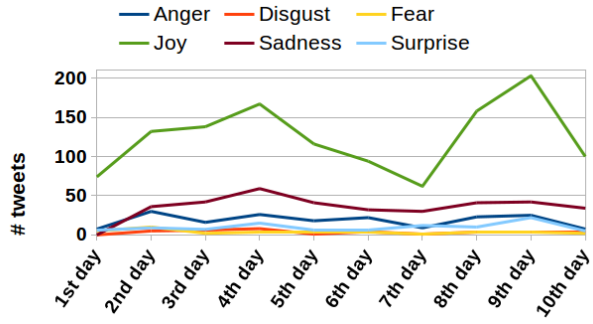
Next, we proceed~~ed~~ with the detection of participants' emotions in the 10-days study. We build upon a lexicon-based approach due to its slight superiority on the study conducted in Section 7.

In total 10,539 messages were exchanged during the study. Figure 8 presents the evolution of the three types of emotions: primary emotions (Subfigure 8a), social ones and those that characterize general affective states (Subfigure 8b). The dominant emotion ~~was~~is *joy*, followed by the emotion of *sadness*. Social emotions and those that characterize general affective states are relatively rare with the most popular to be the *anxiety* emotion. Finally, Figure 9 shows that a large portion of the exchanged messages has no emotion, which is expected as users on Facebook often discuss about details of daily life without necessarily expressing any emotion or about abstract concepts, e.g., politics (Wang et al., 2013).

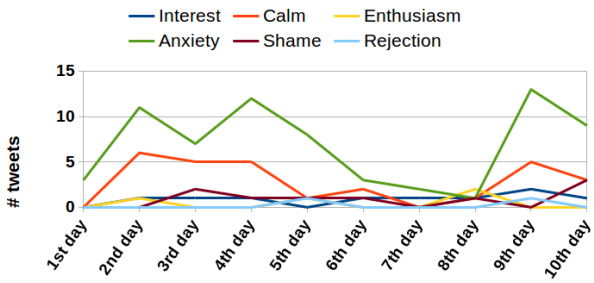
8.4. Comparing implicit and explicit emotions

To be able to compare participants implicit and explicit emotions we proceed~~ed~~ with a *6-hour window*, where a chat is active only for 6 hours after the exchange of the first message. So, then we compare~~d~~ the emotions detected (i.e., with the lexicon-based approach) during such 6-hour windows with the explicit participants emotion declarations for the same time period.

[19] https://dumps.wikimedia.org/



(a) Basic emotions



(b) Social & affective states

Figure 8: Graphical representation of emotions during the 10-days study

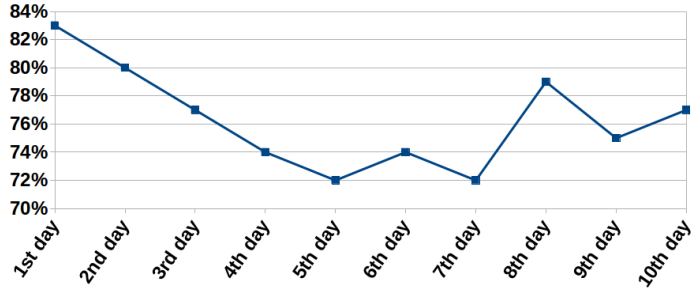


Figure 9: Distribution of neutral state during the 10-days study

Figure 10 overviews the emotions distribution based on both the participants' emotions declarations and the emotions obtained implicitly via the lexicon-based process. We observe that mostly more positive emotions are expressed

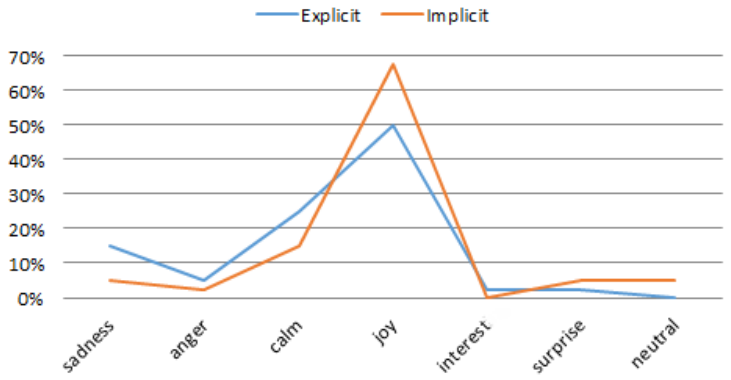


Figure 10: Distributions of explicit and implicit emotions declarations

from participants, while some especially negative emotions, e.g., *disgust*, *fear*, and *rejection*, are totally absent from both implicit and explicit emotion declarations. Overall, in both implicit and explicit emotions statements the predominant emotion is *joy* covering at about half of the cases, followed by the *calm* emotion. The higher presence of positive emotions can be justified by the fact that people in OSNs often try to preserve their 'face' in the online space, i.e., to create positive impressions about themselves in their social network (Chou and Edge, 2012). The overall accuracy, namely the proportion of true results among the total number of cases examined, is **47.50%**. From the 52.50% of the mis-classified cases almost the 38% of them ~~was~~is due to the incapability of the lexicon-based approach to correctly recognize the *calm* emotion, which is also in alignment with the results presented in Section 7. Overall, the analysis on both Twitter (Section 7) and Facebook (Section 8) social networks demonstrates the challenges involved in detecting human emotions in texts. Such difficulties are potentially due to the need to further consider the emotional state of both the author of a text and his surrounding environment during such an analysis, or the lack of considering additional aspects, such as the sarcasm.

9. Conclusions

This work addresses the problem of detecting a wide spectrum of emotions from online text sources. Current research efforts mainly focus on detecting specific primary emotions without considering the social ones or those that characterize general affective states. Motivated by such lack of a wider emotional spectrum analysis, we proceeded with an approach which permits the detection of 12 emotions, i.e., Ekman's six primary emotions, three social ones, and three emotions that characterize general affective states. We built upon widely used emotional analysis approaches, i.e., lexicon-based and machine learning, either individually or under a hybrid scheme, and we exploited various features that consider both emotional and semantic information to better detect the emotions under consideration. The studies conducted on different text sources which span from news headlines to OSN sources, i.e., Twitter and Facebook, to ensure that our methods are valid for online texts with different structural attributes. This work also considers the up to now lack of the explicit human emotion declaration and validation in such studies and it deals with the challenging tasks for detecting people's emotions on the 'wild'. For this goal, a further case study was implemented at which we monitored explicit and implicit human emotional experiences on Facebook to examine whether there are any similarities among explicit declarations and the results obtained with an automatic emotional analysis process, showing quite promising results.

The conducted analysis highlighted the difficulties that exist in detecting a wider spectrum of emotions from texts. So, next research work is foreseen in exploiting more semantic aware features in an effort to capture more complex emotional attributes. Also, social-related aspects, e.g., users' personality, group-related or cultural factors, will be considered to better perceive the expressed social emotions and those that characterize general affective states. Additionally, we indent to conduct a wider case study (i.e., larger sample) which will be more suitable for providing valuable knowledge about specific aspects that characterize each one of the primary, social, or those that characterize general

affective states emotions. Finally, a more well-defined crowdsourcing process will be performed by using well established crowdsourcing platforms and also considering the annotators' credibility.

Reproducibility. For the reproducibility of our results we share the data collected from Twitter social network at: http://bit.ly/2bLgVUP.

Acknowledgements

Part of this work (especially work in Section 8) has been funded by the Network of Excellence in Internet Science (7th EU Framework Programme).

References

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M., and Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PloS one*, 12(2):e0171649.

Augustyniak, L., Kajdanowicz, T., Szymanski, P., Tuliglowicz, W., Kazienko, P., Alhajj, R., and Szymanski, B. K. (2014). Simpler is better? lexicon-based ensemble sentiment classification beats supervised methods. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 924–929.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, volume 10, pages 2200–2204.

Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Becker-Asano, C. and Wachsmuth, I. (2009). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49.

Bhattacharya, D. and Ram, S. (2015). Rt news: An analysis of news agency ego networks in a microblogging environment. *ACM Transactions on Management Information Systems*, 6(3):11:1–11:25.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chatzakou, D., Koutsonikola, V., Vakali, A., and Kafetsios, K. (2013). Micro-blogging content analysis via emotionally-driven clustering. In *Proceedings of Affective Computing and Intelligent Interacti, Humaine Association Conference on*, pages 375–380.

Chatzakou, D., Passalis, N., and Vakali, A. (2015). Multispot: Spotting sentiments with semantic aware multilevel cascaded analysis. In *Big Data Analytics and Knowledge Discovery*, pages 337–350.

Chatzakou, D. and Vakali, A. (2015). Harvesting opinions and emotions from social media textual resources. *Internet Computing, IEEE*, 19(4):46–50.

Chou, H.-T. G. and Edge, N. (2012). "they are happier and having better lives than i am": the impact of using facebook on perceptions of others' lives. *Cyberpsychology, Behavior, and Social Networking*, 15(2):117–121.

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., and Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLoS ONE*, 9(3):1–6.

Derks, D., Bos, A. E. R., and von Grumbkow, J. (2008). Emoticons in computer-mediated communication: Social motives and social context. *Cyberpsychol and Behavior*, 11(1):99–101.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, pages 1–15.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ekman, P., Friesen, W. V., and Ellsworth, P. (1982). *What emotion categories or dimensions can observers judge from facial behavior?* New York: Cambridge University Press.

El-Alfy, E.-S. M., Thampi, S. M., Takagi, H., Piramuthu, S., and Hannen, T. (2015). *Advances in Intelligent Informatics*. Springer.

Fagerberg, P., Ståhl, A., and Höök, K. (2004). emoto: emotionally engaging interaction. *Personal and Ubiquitous Computing*, 8(5):377–381.

Farnadi, G., Sitaraman, G., Rohani, M., Kosinski, M., Stillwell, D., Moens, M., Davalos, S., and Cock, M. D. (2014). How are you doing? emotions and personality in facebook. In *EMPIRE Workshop of the 22nd International Conference on User Modeling, Adaptation and Personalization*, pages 45–56.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine learning*, 29(2-3):131–163.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. (2016). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214 – 224.

Hareli, S. and Parkinson, B. (2008). What's social about social emotions? *Journal for the Theory of Social Behaviour*, 38(2):131–156.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Inkpen, D., Keshtkar, F., and Ghazi, D. (2009). Analysis and generation of emotion in texts. In *International Conference on Knowledge Engineering Principles and Techniques*, pages 3–13.

Kafetsios, K. and Nezlek, J. (2012). Emotion and support perceptions in everyday social interaction: Testing the "less is more hypothesis" in two different culture. *Journal of Social and Personal Relationships*, 29(2):165–184.

Khan, A. Z., Atique, M., and Thakare, V. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89.

Khan, F. H., Bashir, S., and Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245 – 257.

Kim, S. M., Valitutti, A., and Calvo, R. A. (2010). Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Artificial Intelligence Research*, 50(1):723–762.

Kleinbaum, D. G. and Klein, M. (2010). *Introduction to Logistic Regression*. Springer New York.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., and Ureña-López, L. A. (2013). Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934 – 3942.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th international workshop on Semantic Evaluation Exercises*, pages 437–442.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *Artificial Intelligence Research*, 55:95–130.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480 – 499.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241 – 4251.

Nie, C. Y., Wang, J., He, F., and Sato, R. (2015). Application of j48 decision tree classifier in emotion recognition based on chaos characteristics. In *International Conference on Automation, Mechanical Control and Computational Engineering*, pages 1847–1850.

Parkinson, B., Fischer, A. H., and Manstead, A. S. (2005). *Emotion in Social Relations: Cultural, Group, and Interpersonal Processes*. Psychology Press.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Prabowo, R. and Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143 – 157.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI Workshop on Imbalanced Data Sets*, pages 1–3.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4):723–742.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, 67(3):525–539.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., and Harabagiu, S. M. (2012). Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Roseman, I. J., Spindel, M. S., and Jose, P. E. (1990). Appraisals of emotion eliciting events: Testing a theory of discrete emotions. *Personality and Social psychology*, 59(5):899–915.

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

Smith, P. and Lee, M. G. (2013). A ccg-based approach to fine-grained sentiment analysis in microtext. In *AAAI Spring Symposium: Analyzing Microtext*, volume SS-13-01, pages 80–86.

Sreeja, P. and Mahalakshmi, G. (2016). Comparison of probabilistic corpus based method and vector space model for emotion recognition from poems. *Asian Journal of Information Technology*, 15(5):908–915.

Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1556–1560.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., and Zhou, M. (2016). Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185.

Valitutti, R. (2004). Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2012). Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the ASE/IEEE International Conference on Social Computing and ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 587–592.

Wang, Y.-C., Burke, M., and Kraut, R. E. (2013). Gender, topic, and audience response: An analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31–34.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Zareapoor, M. and Seeja, K. (2015). Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business*, 7(2):60.