Journal of Intelligent Information Systems (2021) 57:369–394
https://doi.org/10.1007/s10844-021-00647-8

Mining emotion-aware sequential rules at user-level from micro-blogs

Marjana Prifti Skenduli[1] · **Marenglen Biba**[1] · **Corrado Loglisci**[2] · **Michelangelo Ceci**[2] · **Donato Malerba**[2]

Received: 15 February 2021 / Revised: 18 May 2021 / Accepted: 19 May 2021 /
Published online: 22 June 2021

Abstract

Social Media have enabled users to keep inter-personal relationships, but also to voice personal sensations, emotions and feelings. The recent literature reports on the potential of technologies based on emotion detection and analysis. However, the understanding of user generated emotional content is a challenging task because it requires the extraction of textual units of interest and the search for potential knowledge nuggets, such as those on the correlation between emotions conveyed over time. In this paper, we study this array of problems through the discovery of structured information on the emotions, which is more difficult than the mere recognition of individual mentions. We propose a framework to discover forms of implication between emotions through high-utility sequential rules. Apart from being emotion-aware and time-aware, these rules have the ability to handle numeric information concerning the quantities of expressed emotions, contrary to the classical association rules designed only for binary data. The application on micro-blogs concerning politics shows the viability of the framework to real-world scenarios and its potential to capture user-level emotional behaviours.

✉ Marjana Prifti Skenduli
marjanaprifti@unyt.edu.al

Marenglen Biba
marenglenbiba@unyt.edu.al

Corrado Loglisci
corrado.loglisci@uniba.it

Michelangelo Ceci
michelangelo.ceci@uniba.it

Donato Malerba
donato.malerba@uniba.it

[1] University of New York Tirana, Tirana, Albania, USA

[2] Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro", via Orabona 4, I-70125 Bari, Italy

Keywords Pattern mining · Sequential rules · Emotion analysis · Social media analysis

1 Introduction

Social media and and content-based communication technologies, such as micro-blogs, are increasingly playing an important and growing role in different disciplines, such as Cognitive sciences, Social sciences, Healthcare, Finance/Marketing etc. Their popularity is in large part defined by the platform's inherent simplicity, thus establishing diverse avenues for people to broadcast opinions, to reach out each other or even foster brand-customer relationships. Indeed, micro-blogs are fundamentally human, in that they voice people's perceptions and feelings dealing with the universal experiences of love and hate, life and loss, shame and pride, etc. Of course, the more people willing to use micro-blogs, the more appeal the social media would hold and the richer the content generated for analysis purpose.

In a continued effort to gain valuable hidden information, which can be later used to facilitate decision making processes, researchers have designed computational solutions to work on messages with emotional content, in order to recognize emotional status expressed by the authors and further analyze it. However, while the recognition of emotions has received large consideration (Kang et al. 2018; Mohammad and Kiritchenko 2015), the same cannot be said for the analysis, whose viability in sophisticated content-based technologies, such as sentiment metering, recommendation systems (Simsek and Karagoz 2020), political orientation monitoring, is only recently receiving attention. A research stream which seems gaining growing interest is focused on the evolution of the emotional content and on computational methods able to study how the emotions expressed in the social postings change over time (Sano et al. 2019). Indeed, the emotions that an individual conveys may change because of multiple factors, for instance, due to the interaction with others, such as when an *influencer* user posts something that trigger emotional behaviours of his/her *followers*, or due to feelings or personal experiences (Akiyama et al. 2017; Yang et al. 2017; Ali et al. 2020).

However, the psychological processes underlying the emotions are so complex that one could hardly establish precisely, through the analysis of emotional content messages, the causes that stimulate a change of the emotional status. Instead, what appears feasible is the study of the correlation and implication of the emotions by leveraging the information on the simultaneous occurrence of the emotions. Indeed, the co-presence provides statistical evidence of the correlation and therefore may provide valid arguments on the interaction between emotions and on the implication. Thus, emotions which occur simultaneously over time with a certain regularity or that are manifested regularly together appear worthwhile to be investigated than those which are more episodic. This makes users which frequently post messages to be more preferred for analysis purposes than those who post sporadically.

All the above considerations are addressed in this paper, in which we propose a computational solution designed to discover forms of correlation and implication of emotions by finding patterns of concomitant emotional status and time-separated emotional status over social postings. To do this, three issues need to be considered.

First, the emotions might not be uniformly expressed over the postings, in the sense that users might post a considerable number of messages with the intention of expressing a particular emotional status, while they behave normally when feel other emotions. This means that we should deal with uneven distributions of the occurrences of the emotions over time. To do this, in this paper, we use a quantification mechanism able to distinguish the different contributions of each emotion and assign higher weights to the emotions with more occurrences (or more mentions) than those which are less frequent (or less mentioned).

Second, to analyze the evolution of the emotional status of a user, one should process the emotional content of the messages she/he authors over time. This needs a representation formalism able to depict both the co-occurrences (e..g, two emotions are expressed together) and successions (e.g., one emotions is expressed after another one). To do this, in this paper, we consider the itemset-based sequence representation in Skenduli et al. (2018), which, additionally, allows decoupling of the analysis from specific temporal granularities. This solution enables us to represent flows of messages which have been deployed with very different temporal distribution (e.g., a user posts messages daily, another one hourly) with the same time-related representation.

Third, the analysis of messages with emotional content turns out more challenging than the one concerning other kinds of textual snippets because the emotional status is personal and requires subjective forms to be expressed, which are different, for instance, from the ways we find formal narratives or texts of general interest. This means we could rely neither on purely lexical resources nor on general purpose solutions of natural language processing, but we had to build models able to learn the way the emotions are reported/conveyed. This is strongly based on the characteristics of the native language of the authors and therefore we could not utilize models of any language. In this paper, we resort to machine learning approaches which allow us to build models able to recognize emotions regardless of language (language agnostic) and without necessarily incorporating lexical resources.

The rest of this paper is organized as follows. Section 2 illustrates the contributions and underlying motivations, while Section 3 discusses literature close to the investigated problem and presents contributions. Section 4 provides a description of the methodology through three main steps. The application to the social messages posted by politicians is reported in Section 5, where we discuss the discovered High-Utility Sequential Rules (HUSR) at the level of individual user. Finally, we draw conclusions and remarks in Section 6.

2 Contributions

The problem investigated in this paper underpins three sub-problems, namely analyzing social messages with emotional content, identifying emotional status and working on concomitant emotions and time-separated emotions. These raise technical challenges which we face with contributions in document classification, deep learning, emotion psychology and pattern mining described as follows.

Social messages are textual snippets that cannot be processed as they are. By nature, they are prone to sparseness, low quality of the writing, and often contain abbreviations, neologisms, typos and slang forms. The social messages of micro-blogs specifically are very short and have limited content, which suggests us to process these as sentences, rather than in the form of documents. Indeed, documents are typically characterized by richer content, written in correct language and are often formal and impersonal. A pre-processing step is therefore necessary and crucial to get an abstract, time-related and emotion-aware representation. We implement it by combining a natural language processing pipeline and a neural network architecture, which, together, allows us to tackle the emotion identification as a sentence-based classification task.

The classification is a predictive task which requires learning models from training emotive sentences, which typically are produced by means of a (semi-automatic) annotation activity. There is large availability of such textual corpora mostly for widely-covered languages (such as, English and Chinese), while the same cannot be said for rare idioms. In this work, we consider the Indo-European languages and, in particular, Albanian, which is a very interesting language characterized by words that have multiple meanings (Skenduli et al. 2018). This is an issue we try to address through a neural network architecture.

Neural models are different from traditional machine learning methods, in that a neural model does not rely on previously extracted features, since features are identified during the (deep) training process. The use of these solutions has already been proved successful for classifying messages based on sentiment polarity (dos Santos and Gatti 2014), while very few attempts have been made for emotion-based categories. To the best of our knowledge, the current paper represents the first research which adopts deep learning methods to recognize emotion categories from social messages written in rare idioms.

To discover forms of correlation and implication between emotions, we rely on the frequent itemset mining framework (FIM), whose purpose is that of searching itemsets and evaluate their statistical evidence over databases (Han et al. 2004). FIM works on the emotion categories assigned to the messages, resulting from the pre-processing step. Thus, emotions and concomitants of emotions would play the roles of items and itemsets respectively. However, FIM is limited to handle only the binary information on the presence/absence of the emotions and it is not able to capture the quantitative information on the number of occurrences of an emotion in a message. To do that, we resort to the high-utility itemset mining (HUIM) (Gan et al. 2018), which allows us to associate quantities to the emotions and represent the differences in terms of occurrences, even when emotions do not co-occur frequently together. Although the utility-based itemsets appear to be an adequate solution to capture patterns of concomitant emotions, it may not handle time-separated emotions. This is the reason why we resort to the High-Utility Sequential Rules (HUSR) (Zida et al. 2015a), which would capture both sequential patterns of emotions and quantitive information, therefore, unearth forms of implication between emotions.

HUSR are discovered from social messages posted by users, taken individually. The messages are collected by time-windows so that we can grasp individual emotional behaviours and even have indication on the confidence degree with which an emotional status may be expressed after others. This is different from working on the social messages of communities of users, where the resulting HUSR may provide little significant information as several users may convey different emotional status and this may result in emotional behaviours without statistical evidence.

3 Related work

The main novelty of the present work consists in the analysis of social postings with emotional content through a frequent itemset mining framework, which, despite the potential viability in sophisticated emotive content-based technologies, has attracted attention only recently. In particular, we record a growing interest on methods that discover knowledge in the form of implication between emotions through association rules and causal rules. Diaz-Garcia et al. (2020) have investigated the association between sentiments and persons in the context of US political elections. They first discover association rules from named entities mentioned in Tweets, then they take only the rules whose consequents have names of politicians and finally generalize, through a sentiment-based pre-defined taxonomy, the entities occurring on the rule antecedents. In Hai et al. (2011), the authors use association rules for associative classification. In particular, they are interested in obtaining the most common characteristics regarding certain groups of words that can represent an opinion (class), identified as the consequent of a rule. Two other methods using the association rules, look for patterns of sentiments on Twitter for two different purposes, respectively (Bing et al. 2014; Mamgain et al. 2016). The first assigns sentiments to the items and then perform the Apriori algorithm. However, the above-mentioned method does not account the time-variability of the emotional content and consider only the simplified information on the presence/absence of the sentiments, on the contrary of our work presented here.

Other kinds of rules have been also explored. In Tzacheva et al. (2020), the authors propose to extract action rules as form of actionable recommendations. Action rules describe the possible transition of objects and, in that work, action rules have been used to show how sentiments of Twitter data change to become more positive. The authors in Dehkharghani et al. (2014) introduce sentimental causal rules to determine the polarity degrees of Twitter data. In particular, a constraint-based technique, called Local Causal Discovery, is used to understand the causality among different aspects of products and sentiments towards these aspects. Another category of rules, that is, those of classification, has been addressed for the predominant area of emotion/sentiment recognition on emotive texts (Wen and Wan 2014; Yuan et al. 2014; Berka 2020).

Studies on the implication between emotions have been presented from Yada et al. (2017) and Fan et al. (2020) and Gao et al. (2015). Yada et al. (2017) report a bootstrapping method to automatically build emotion causes from conjunctive phrases that are similar to initial emotion causes, which however have to be manually configured. Fan et al. (2020) defines a linguistic approach to incrementally build a graph of transition between emotions by exploiting a parsing procedure. Gao et al. (2015) implement a rule-based system upon a pre-defined emotional model. The paper exploits the conditions that trigger emotions and extracts the corresponding cause events in fine-grained emotions from the results of events, actions of agents and aspects of objects. Commonly, these three approaches rely on manual intervention or expert knowledge encoding, which proves the difficulty of the study of the implication with purely automatic and data-driven approaches, such as those based on itemset mining.

Another characteristic which often recurs over the existing works on association rules and causal rules of sentiments/emotions is the assumption of the sole presence/absence within postings, with no consideration on the different emphasis with which the emotions can be manifested. This is what we aim to do with the framework on high-utility sequential rules, whose adaptation on emotional analysis, and more generally, on social media is quite novel. Huang et al. (2015) resort to the high utility patterns (which are itemset configuration from which sequential rules are derived) for the task of topic detection from microblogging text streams. More precisely, high utility patterns are used to identify the most representative and non-redundant co-occurrences of topic words, where the representativeness corresponds to a frequency-based utility measure, while non-redundancy estimates the co-presence of two patterns in the texts. A quite similar application of high-utility pattern has been proposed in Choi and Park (2019), where the authors focus on the detection of emerging topics in Twitter by implementing a notion of utility which accounts for the growth in appearance frequency of the words over tweets. To the best of our knowledge, there is not much research on emotive texts even with high-utility itemsets only, while there is a recognizable interest on temporal/sequential data. For instance, Gan et al. (2018) combines utility and correlation to extract non-redundant correlated itemset in purchase behaviours.

4 Methodology

The primary focus of this work is to present and assess a novel approach for fine-grained emotion analysis, mining high-utility sequential rules on micro blogging data. The proposed method relies on three important tasks: (1) Data pre-processing; (2) Emotion detection (sentence-based classification); (3) Emotion-aware sequential rule mining.

4.1 Data pre-processing

Data preprocessing is crucial to ensuring the successful implementation of a machine learning task, especially for the social messages/micro blogging domain, whose textual data are subject to what is dubbed as the ”curse of dimensionality” (as the dimensionality increases, the space of all possible input examples increases so fast that the available data become sparse), leading to increased sparsity and noise. In order to bring our raw data into a form that is predictable and analyzable, we design/conceptualize a fully fledged workflow adapted to the needs of our approach. This workflow starts with a preliminary preprocessing task that aims to clean and scale the real-world datasets, as they prepare to move through the pipeline. Preprocessing is handled over a set of basic but well-chosen techniques which are far from being exclusive but yet they remain domain gnostic (tailored to the characteristics of our microblogging data). As illustrated in Fig. 1, we start with the normalization of the text, aimed at transforming lexical variants into their canonical form. Next we propose noise removal which entails the removal of special characters, digits and punctuation marks that affect the accuracy of our analysis. Then we proceed with lowercasing the textual data and padding the posts to a maximum length (e.g. 250 characters). The curated Facebook posts are further stored in a database from where the annotation of data begins. The annotation procedure is performed in three different steps by three different expert annotators. Initially, two annotators label one half of the training corpus independently from each other. Then we measure their mutual rater agreement, using proportional kappa metric as an indicator. In the second step, the goal is to increase inter-rater reliability. Therefore, the third annotator compares both annotated versions and resolves the discrepancies, informing accordingly the first two annotators. In the final step, the annotators proceed with the annotation of the remaining data, making sure they work completely independently by consulting the third annotator.

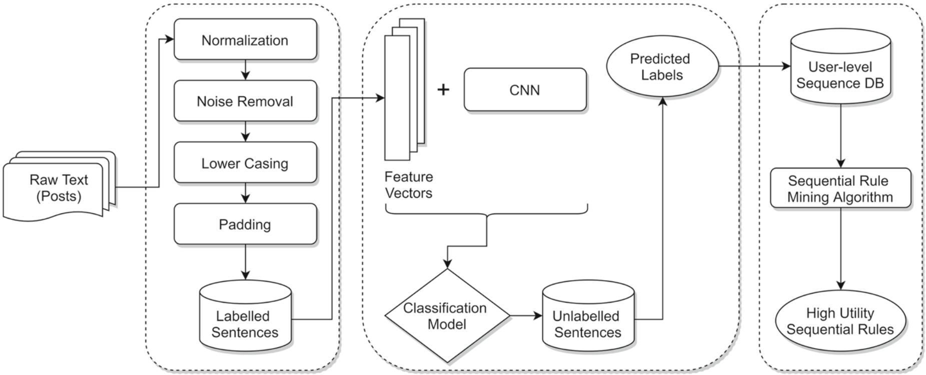

Fig. 1 Workflow of the proposed approach

4.2 Emotion detection

After being subjected to the preprocessing task, the textual data are eligible to undergo an important second task, that of emotion detection. The emotion detection task can be cast as a sentence-based classification problem. One of the essential features of sentence-based classification is that it assigns social messages to pre-defined classes/categories of emotions, as formulated in the Theories of Emotion in Psychology. In our case, we adopt the Ekman model introduced in Ekman (1993).

Current research suggests that deep learning based models have significantly surpassed classical machine learning approaches in various text classification tasks including but not limited to sentiment/emotion analysis. State-of-the-art approaches developed to tackle the emotion detection task, rely on the exploitation of either the intra-sentential context or the inter-sentential one, or both (Yang and Cardie 2014). In doing so, neural network architectures starting from Feed Forward Networks, going on to Recurrent Neural Networks (RRN-s), Convolutional Neural Networks (CNN-s), Transformers and more, have proven to be way more accurate compared to classical rule-based methods. In our analysis we choose to exploit emotion signals/indicators at the sentence level, trying to learn the conveyed emotion while taking into account the context in which they occur, rather than the overall discourse. We employ a neural network architecture based on a Convolutional Neural Network, a model inspired from Kim (2014), that has since become a baseline standard for text classification architectures. Though, the original model has been extended to accommodate our multi-class problem in the Albanian language setting (Skenduli and Biba 2020). We basically train a simple CNN with one layer of convolution on top of feature vectors obtained from previously labelled sentences. Each word is regarded as a feature, transforming therefore a sentence into a concatenation of low-dimensional vectors. As to the choice of the data representation model, we refer to our previous work[1], where we assess the impact of several word embedding techniques (from shallow to deep contextualized, including pre-trained language models (LMs)) on the fine-grained emotion analysis task. Based on the obtained results, we decide to use multi-lingual word embeddings, pre-trained on Common Crawl and Wikipedia corpora using fastText. These models were trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives (Grave et al. 2018). As shown in Fig. 1 the output of the pre-trained embedding layer is fed into the convolutional layer, which in turn filters the embedded word vectors using multiple widths. Next, they go through a ReLu activation and maxpooling operation. Finally, the max-values from the convolutional layer are passed to a last, fully-connected, classification (*softmax*) layer. The resulting classification model is used to automatically predict the probability distribution of emotion labels over the set of unlabelled sentences.

In conclusion, our choice for a CNN-based model is motivated by its simplicity and ability to learn to recognize patterns in a high-dimensional space. Additionally, our choice to incorporate pre-trained word vectors into our CNN model, is motivated by the fact that pre-training is crucial for resource constrained languages like Albanian and also steers models to richer and more insightful learning, significantly improving their accuracy compared to baseline models.

[1] http://womencourage.acm.org/2020/wp-content/uploads/2020/07/womENcourage_2020_paper_11.pdf//

4.3 Emotion-aware sequential rule mining

In our quest to give an extra edge to our designed approach we decide to further analyze the emotion related information gained post classification task. We aim to discover patterns of concomitant emotional status and time-separated emotional status over social postings, which will indeed reveal forms of correlation and implication between emotions. This is achieved by introducing a third task, that of sequential rule mining of utility-annotated emotions conveyed through social messages. In order to prepare the input to this task, we develop a Java based conversion tool that implements our representation formalism, which not only frees up the analyses from temporal granularities but more importantly it doesn't necessarily imply the use of lexical resources. This conversion tool is instrumental in designing a language agnostic solution that can easily be adopted for a wide range of high to low resource languages. Additionally, this tool is powerful because it turns data attributes into time-related and emotion-aware numerical representations that satisfy the input requirements of a sequential rule mining algorithm. More specifically, the input to the conversion tool consists of automatically predicted emotion labels corresponding to the posts of a specific user/author, along with their respective timestamps. These non-lexical data undergo a complex conversion process that aims to output sequences, out of which we build the whole user-level sequence database.

Our sequence database $SDB = \langle s_1, s_2, \ldots, s_p \rangle$ collects the emotional status recognized from the messages posted by a user. In particular, it encloses a list of sequences, each assigned to an identifier 1, 2, ..., p. A sequence s_p is a list of itemsets of (labels of) emotional status, automatically predicted from messages posted during a user-defined time window and arranged according to their chronological onset over time. Intuitively, each itemset associated to a unique time-stamp, may contain several emotional status (occurred in the same time-stamp), but cannot have the same emotional status appearing more than once. Each emotional status is associated to a positive integer value, referred to as its utility.

A concrete illustration is reported in the following example:

$\{\langle\{joy[4], anger[1], shame[1]\}, t_i\rangle, \langle\{joy[1], anger[1], shame[1]\}, t_{i+1}\rangle, \langle\{joy[1]\}, t_{i+2}\rangle, \langle\{joy[1]\}, t_{i+3}\rangle\}$

where, for instance, at the time-stamp t_i, three labels co-exist, namely, the emotional status joy occurs four times, $anger$ and $shame$ occur once, respectively.

So, each row of the sequence database represents a list of ordered transactions happening during the selected time window. Transactions, in turn, represent itemsets that adhere to a total chronological order. Whereas, items (here emotion labels) within an itemset are not necessarily ordered and moreover they should be distinct (no item can appear twice). It is important to note that due to the required distinctiveness of items within an itemset, we can maintain only a total order. Additionally, in order to better model authors' posting behaviour, the time window of sequences is defined based on days of consecutive posting activity, rather than on calendar days, omitting therefore days with no user-posting activity.

4.3.1 The proposed solution

To further our analysis, we investigated several frameworks ranging from FIM to HUIM which have been successfully applied in several domains, but they fail to capture and quantify emotions conveyed through social messages, and more importantly they provide no measure of confidence that the rule will be followed. Consequently, our emotion analysis problem lends itself to a different, yet less explored family of algorithms, namely High-Utility Sequential Rule (HUSR) mining algorithms.

High-utility sequential rule mining is an important data mining task with a wide application focus. However, to the best of our knowledge, its main real-world applications, in any event go back to product recommendation and market basket analysis. We have been inspired from the later application to draw similarities between customer transactions and user generated emotions and consequently adopt high-utility sequential rule mining onto the field of emotion analysis. In particular, in the market basket analysis context, a typical sequence database consists of sequences of customer transactions, where each transaction consists of items bought, and each item is annotated with an individual utility value denoting the sale profit. In our emotion analysis context, we have a different working scenario, in that there are sequences of emotions expressed from the author of social messages over time, and this solicits changes to the notion of utility. Thus, instead of ”sale profit” (of sold items), we introduce ”occurrences” of the emotional status, which are the emotion-labels assigned to the posts.

The notion of HUSR is derived by the one of *sequential rule* (Lo et al. 2009), which is an implication in the form of *if antecedent then consequent* ($X \rightarrow Y$), where antecedent and consequent have no element in common. In a sequential rule, if the conjunction of the items on *antecedent* (X) occurs in a sequence, the conjunction of the items on *consequent* (Y) is likely to occur afterward with a given *confidence* or probability in the same sequence. Moreover, not all the sequential rules are of interest, but only those that are verified in a significant number of input sequences. However, HUSR differ from sequential rules by the property to consider the utilities specific to the corresponding input sequences and represent them with a summarizing value. Therefore, a high-utility sequential rule is denoted with three statistical parameters, taking into account that not all the HUSR are considered as valid, but only those that exceed three user-defined thresholds, namely, *min_confidence* $\in$ [0,1], *min_support* and *min_utility* $\in \mathbb{R}^+$, respectively, minimum thresholds for confidence, support and utility.

To discover HUSR present in at least *min_support* sequences SDB, we account for the *support* of each rule $X \rightarrow Y$. It can be determined as the portion of the sequences of SDB in which the conjunction $X \cup Y$ is present.

The confidence denotes the strength of the implication *if* X *then* Y and it can be determined as the ratio of the number of sequences of SDB in which the conjunction $X \cup Y$ is present out of the number of sequences of SDB in which only the antecedent X is present.

Finally, an utility value is associated to each HUSR and can be determined as the sum of the utilities associated to the individual emotional status involved in X and Y.

To discover HUSR, we could inject the information of the utilities of the emotional status in the search space of the sequential rules and resort to any algorithm of discovery of sequential rules, with the difference that each rule is denoted with the utility value. The typical procedure generates rules with more and more longer antecedents and consequents, that is, by incrementally inserting new items into either antecedents or consequents of valid rules, starting from those with one item on the antecedent and one item on the consequent.

However, to do that, most of algorithms rely on the anti-monotonicity property of the support, which would limit us to keep only those sequential rules that meet the support (and confidence). Indeed, we can build longer rules from (shorter) rules that meet *min_support* and *min_confidence*, but we can avoid to build longer rules when the (originating) shorter rules do not exceed *min_support*. This is not quite the case with the utility properties, where we cannot clearly define whether it is anti-monotonic or not. In fact, careful inspection of the search space of rules, reveals that it is possible to encounter two different cases: (1) expansions of a shorter rule, whose utility decreases due to the fact that they are supported in a smaller or in an equal number of sequences and (2) expansions of a shorter rule, whose utility increases because they are supported in a greater or in an equal number of sequences.

A promising idea is that of associating ’approximated’ utilities to the rules and explore the space accordingly. This has been proposed in Zida et al. (2015b), which we have adopted in the current work. That method uses the sequence estimated utility measure to prune unpromising items and rules. It also leverages two efficient data structures to compute the three statistical parameters for each rule: the *utility-table* structure that is used to quickly calculate both the support and utility of rules and the *bit vector* to calculate the confidence.

5 Experiments

In this section we implement the proposed computational solution by means of the HUSRM algorithm, in order to test and assess it on two different types of real-world data sets: one multi-user data set namely D1 and six single-user data sets namely D2, D3, D4, D5, D6 and D7. Having access to a large choice of user generated data we could sample either multi-user data sets or single-user data sets, and run experiments along three different strands:

1. Evaluate high-utility rules en masse (all data sets)
2. Evaluate high-utility rules individually (selected/most representative data sets)
3. Compare and contrast emotional traits expressed from a single user versus a multi-user pool

We will also report experimental results and discuss conclusions drawn from both a qualitative and quantitative perspective.

5.1 Data sets and experimental settings

Our experiments are conducted on seven real-world data sets, built from micro-blogging data, more specifically from live Facebook posts. The collection of micro-blogging data, was performed by means of RestFB[2] - a simple and flexible Facebook Graph API client written in Java language. It is an open source tool released under the terms of the MIT License. We had to reuse the source code and setup a solid framework that allowed us to fetch posts out of public community pages belonging to Albanian public figures. We fetched around 60K posts belonging to 119 Albanian politicians, shortlisted among the most active ones on Facebook and the most popular according to the general perception of the social media audience.

The fetched posts where captured and stored in a local SQL database, from where we could extract and build several data sets to fit the needs of our experimental study. Originally, our extracted data contains the following attributes: *source_id*, *post_id*, *caption description*, *icon*, *link*, *message* and *created_date*. However, after executing the preprocessing steps explained in Section 4 and subjecting our data to the neural network classification model, we proceed with an ablated version of our original data sets, analyzing instances with only two attributes: *emotion_class* and *created_date*. Table 1 encompasses detailed information about data sets’ properties. As shown in this table, we expect data set D1 to be prone of a higher temporal density, while the rest of data sets promise to reveal insights widely lingered

[2] https://restfb.com//

Table 1 Data set properties

Data set ID	No of instances	Description	Timestamp interval / Time coverage
D1	2352	Multi-user	Jan 2018-Mar 2018
D2	158	Single-user	Oct 2013-Mar 2018
D3	321	Single-user	Jan 2013-Mar 2018
D4	1002	Single-user	Nov 2011-Mar 2018
D5	1490	Single-user	May 2011-Mar 2018
D6	1648	Single-user	Dec 2011-Mar 2018
D7	3285	Single-user	Dec 2011-Mar 2018

across time axes. More details on the emotive post counts and their temporal distribution are depicted in Fig. 2.

Implementation of this study is done in Java SE (version 15) using the *SMPF* open-source, high-utility sequential rule mining libraries (Fournier-Viger et al. 2014) and MySQL Workbench (version 6.3.10), on a 64-bit Windows 10 Pro N workstation with 16 GB RAM. The preprocessing workflow and the neural network architecture for the sentence-based classification are implemented in Google Colaboratory (Google Colab) with NVIDIA Tesla K80 GPU using Python (version 3.8).

5.2 Experimental evaluation

In order to assess the proposed approach in terms of effectiveness in representing user generated emotions and prediction of future emotional trails, we come up with novel heuristics

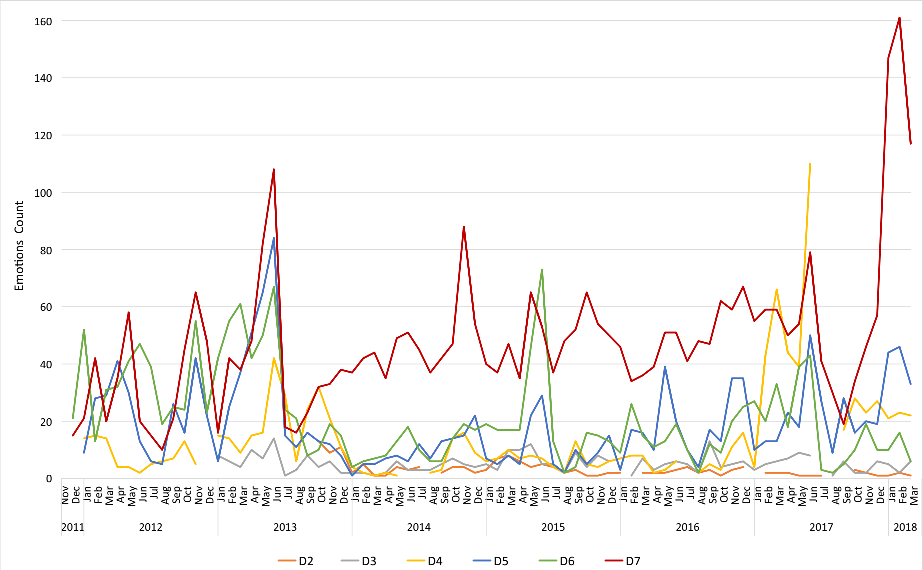

Fig. 2 Temporal distribution of conveyed emotions in terms of emotive post counts for single-user data sets D2-D7

that rely mainly on standard evaluation metrics but also extend to include human-oriented interpretations. The usage of a high-utility sequential rule mining algorithm, imposes us to exploit the mining output which consists of high-utility sequential rules followed by three important metrics: rule support, rule confidence and rule utility, respectively denoted as *#SUP*, *#CONF* and *#UTIL*. More specifically, by interpreting these three metrics we can introduce the notion of interestingness among discovered rules. In doing so, we strictly rely on the utility (profit) of a rule converting it into a scoring function that allows us to pick the most important sequential rules and further evaluate them by means of the support of a rule and confidence of a rule. The later one is interpreted as the strength or usefulness, which in turn provides a measure of the probability that the rule in question will be followed.

5.2.1 Experimental results

Parsing emotional textual content is a rather unique and complex task, however the comprehensive set of experiments that we present in this section aim to uncover interesting insights through the mining of emotion-aware sequential rules deduced from user-generated microblogs. As previously explained in Section 4, the explicit output of HUSRM algorithm consists of a text file, where every line defines a high-utility sequential rule. Rules can be broken down into two components: the antecedent, consisting of left side concomitant items and the consequent which constitutes the right side items, followed by the numerical value of three important metrics: *#SUP*, *#CONF* and *#UTIL*. We run a total of 100 experiments on seven benchmark data sets varying minimum utility threshold *min_utility* (a positive integer) in the [2, 24] interval and the minimum confidence threshold *min_confidence* (a double value) in the [0.2, 0.9] interval. Below we present a breakdown of experimental results arranged in three main groups.

5.2.2 Evaluation of high-utility rules en masse

In this strand of experiments we collectively analyze all the high-utility sequential rules generated from data sets D1-D7, comprehensively analyzing the data, while keeping an eye on the data set pertinence for more elaborated conclusions. We start by identifying the rules with the highest utility, then those with the highest support followed by the ones with the highest confidence. The results are tabulated and the details are debriefed below. As shown in Table 2, it can be clearly observed that the rule with the highest utility *#UTIL* overall, is $\{shame\} \rightarrow \{joy\}$ that belongs to the single-user data set D7. As to the rules with the highest confidence, Table 3 lists all the rules that satisfy this criteria and they belong to data sets D1 and D4. It is clearly noted that among the rules with the highest confidence, *fear* or *disgust* is the most likely emotion to occur afterward with an optimal confidence of *#UTIL*=1.

Table 2 High-utility sequential rules with the highest rule utility (en masse)

Dataset ID	Antecedent	Consequent	$\#CONF$	$\#UTIL$	$\#SUP$ %
D5	SHAME	JOY	0.46	272	40%
D6	JOY	SADNESS	0.30	99	30%
D7	SHAME	JOY	0.50	**283**	44%

Table 3 High-utility sequential rules with the highest rule confidence (en masse)

Dataset ID	Antecedent	Consequent	$\#CONF$	$\#UTIL$	$\#SUP$ %
D1	JOY, ANGER, SADNESS, GUILT	DISGUST	1	8	1.15%
D1	JOY, ANGER, DISGUST, SHAME	FEAR	1	8	1.15%
D1	JOY, DISGUST, SHAME	FEAR	1	8	1.15%
D1	ANGER, DISGUST, SHAME	FEAR	1	8	1.15%
D4	ANGER, DISGUST, SHAME	FEAR	1	14	1.15%
D4	JOY, ANGER, DISGUST, SHAME	FEAR	1	14	1.15%
D4	JOY, DISGUST, SHAME	FEAR	1	14	1.15%
D4	JOY, ANGER, SADNESS,GUILT	DISGUST	1	14	1.15%

As to the rules with the longest antecedent overall, we extracted the following results that can be examined from Table 4. Among the extracted rules, the most useful rule with the highest rule utility and support is $\{joy, fear, sadness, shame\} \rightarrow \{anger\}$, pointing to an interesting sequential relationship between the noted user generated emotions. Conversely, the extract of the rules with the longest consequent is dominated by $\{sadness, disgust\}$ or $\{fear, anger\}$ emotions duplets and has a length not greater than 2 items (emotions) (see Table 5). Here, the most useful and interesting rule with the highest rule utility and support is $\{joy, anger, guilt\} \rightarrow \{sadness, disgust\}$.

Table 6 reports all the rules consisting on a single item antecedent and a single item consequent. The rules with the highest metrics value are noted in bold for every data set that is part of this en masse analysis, leading us to some interesting insights: (1) There are no *if single item antecedent then single item consequent* rules in the multi-user data set D1; (2) $\{shame\}$ or $\{joy\}$ emotions are predominantly showing in the antecedent side of the most important rules; iii) Data sets D5 and D7 have in common a rule that has single emotion on both antecedent and consequent side, that is $\{shame\} \rightarrow \{joy\}$. The rest of data sets are disjoint.

The experiments on the real-world data sets have been set on the properties of the input data sequences, varying two input thresholds *min_utility* and *min_confidence*. As a result we have filtered down all the distinct high-utility sequential rules, summarizing the distribution of the seven Ekman emotions on the antecedent and on the consequent side, as depicted in the following radar charts (Fig. 3 a, and b). Interestingly, the antecedent radar chart is denser than the consequent one and reveals that the ubiquitous emotion among all data sets is $\{joy\}$ followed by $\{sadness\}$. According to Fig. 3b $\{sadness\}$ reappears as a ubiquitous emotion, this time on the consequent side. As to the top ranking emotions, there is evidence that the antecedent side is dominated from $\{guilt\}$ and the consequent side from $\{anger\}$, hence pointing towards emotions that have a negative polarity.

For comparative purposes, in Table 7 we report sequential rules obtained from dataset D1, using our algorithm HUSRM and an alternative sequential rule mining algorithm namely ERMiner. We found that rules generated with ERMiner and those with our proposed approach are not directly comparable even though there is a slight overlapping if we consider same rules generated. This is due to the fact that our proposed framework is based on the utility score which is conceptually different from what is used in ERMiner, which in turn uses simply frequency-based thresholds.

Table 4 High-utility sequential rules with the longest antecedent

Dataset ID	Antecedent	Consequent	$\#CONF$	$\#UTIL$	$\#SUP$ %
D1	JOY, ANGER, SADNESS, GUILT	DISGUST	1.0	7	1.15%
D1	JOY, ANGER, DISGUST, SHAME	FEAR	1.0	5	1.15%
D3	JOY, ANGER, SADNESS, DISGUST	FEAR	0.3	5	2.44%
D4	JOY, ANGER, SADNESS, GUILT	DISGUST	1.0	7	1.15%
D4	JOY, ANGER, DISGUST, SHAME	FEAR	1.0	5	1.15%
D6	JOY, FEAR, DISGUST, GUILT	ANGER	1.0	5	0.74%
D6	JOY, FEAR, SHAME,GUILT	ANGER	1.0	5	0.74%
D6	FEAR, DISGUST, SHAME,GUILT	ANGER	1.0	5	0.74%
D6	**JOY**, **FEAR**, **SADNESS**, **SHAME**	**ANGER**	**0.5**	**25**	**2.96%**
D6	JOY, DISGUST, SHAME, GUILT	FEAR, ANGER	0.3	6	0.74%
D6	JOY, DISGUST, SHAME, GUILT	FEAR	0.3	5	0.74%
D6	JOY, DISGUST, SHAME, GUILT	ANGER	0.3	5	0.74%
D6	JOY, DISGUST, SHAME, GUILT	SADNESS	0.3	5	0.74%
D7	JOY, SADNESS, DISGUST, GUILT	ANGER	1.0	5	0.43%
D7	JOY, DISGUST, SHAME, GUILT	ANGER	1.0	5	0.43%
D7	SADNESS, DISGUST, SHAME, GUILT	ANGER	1.0	5	0.43%
D7	JOY, SADNESS, SHAME, GUILT	ANGER	0.5	5	0.43%

Table 5 High-utility sequential rules with the longest consequent

Dataset ID	Antecedent	Consequent	*#CONF*	*#UTIL*
D4	**JOY, ANGER, GUILT**	**SADNESS, DISGUST**	**0.50**	**7**
D4	ANGER, GUILT	SADNESS, DISGUST	0.33	6
D6	JOY, DISGUST, SHAME, GUILT	FEAR, ANGER	0.33	6
D6	JOY, DISGUST, GUILT	FEAR, ANGER	0.33	5
D6	DISGUST, SHAME, GUILT	FEAR, ANGER	0.33	5
D6	DISGUST, GUILT	FEAR, ANGER	0.33	4

Table 6 High-utility sequential rules with single antecedent and single consequent

Dataset ID	Antecedent	Consequent	*#CONF*	*#UTIL*
D2	**JOY**	**ANGER**	**0.35**	**21**
	SHAME	GUILT	0.33	2
	ANGER	DISGUST	0.25	4
	ANGER	SHAME	0.25	4
	DISGUST	JOY	0.25	2
	DISGUST	ANGER	0.25	2
	DISGUST	SHAME	0.25	2
D3	**DISGUST**	**ANGER**	**0.30**	6
	JOY	**SHAME**	0.29	**25**
	JOY	ANGER	0.24	21
	SHAME	JOY	0.24	9
	ANGER	DISGUST	0.27	7
	DISGUST	JOY	0.20	4
	DISGUST	SHAME	0.20	4
D4	**SHAME**	**FEAR**	**0.40**	4
	GUILT	DISGUST	**0.40**	4
	SHAME	**ANGER**	0.37	**22**
	SHAME	DISGUST	0.33	20
	FEAR	DISGUST	0.33	4
D5	**SHAME**	**JOY**	**0.46**	**272**
	SHAME	DISGUST	0.42	240
	JOY	DISGUST	0.36	186
	SHAME	SHAME	0.34	185
D6	**GUILT**	**ANGER**	**0.43**	6
	JOY	**DISGUST**	0.30	**99**
D7	**SHAME**	**JOY**	**0.50**	**283**
	GUILT	**ANGER**	**0.50**	4
	SHAME	DISGUST	0.49	268

Table 7 High-utility sequential rules with the highest *#UTIL*, generated from ERminer versus HUSRM algorithm for dataset D1. Rules reported below are obtained for a fixed value of *min_confidence* = 0.6

ERminer		HUSRM	
Antecedent	Consequent	Antecedent	Consequent
FEAR, GUILT	JOY	JOY, ANGER, SADNESS, GUILT	DISGUST
FEAR, DISGUST, GUILT	JOY	JOY, ANGER, SADNESS, GUILT	DISGUST
ANGER, SADNESS, SHAME	JOY	JOY, ANGER, DISGUST, SHAME	FEAR
SADNESS, SHAME	JOY	JOY, ANGER, SADNESS, GUILT	DISGUST
JOY, ANGER, DISGUST, SHAME	FEAR	JOY, ANGER, DISGUST, SHAME	FEAR
JOY, DISGUST, SHAME	FEAR	JOY, DISGUST, SHAME	FEAR
ANGER, DISGUST, SHAME	FEAR	JOY, ANGER, SADNESS, GUILT	DISGUST
DISGUST, SHAME	FEAR	ANGER, SADNESS, SHAME	JOY
JOY, DISGUST, SHAME	ANGER	ANGER, DISGUST, SHAME	FEAR
DISGUST, SHAME	ANGER	JOY, ANGER, DISGUST, SHAME	FEAR
JOY, ANGER, SADNESS, GUILT	DISGUST	JOY, DISGUST, SHAME	FEAR
ANGER, SADNESS, GUILT	DISGUST	JOY, ANGER, SADNESS, GUILT	DISGUST
JOY, DISGUST, SHAME	FEAR, ANGER	ANGER, SADNESS, SHAME	JOY
DISGUST, SHAME	FEAR, ANGER	ANGER, DISGUST, SHAME	FEAR

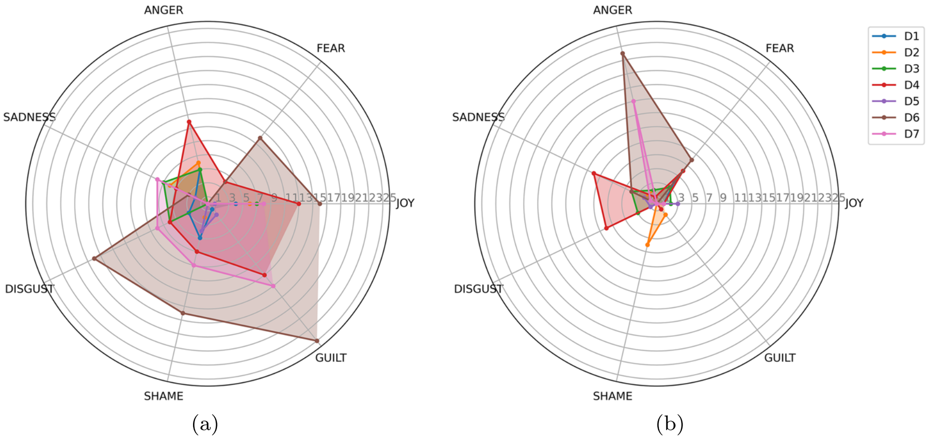

Fig. 3 Emotion frequency based on the overall number of distinct high-utility sequential rules: **a** Antecedent side, **b** Consequent side

5.2.3 Evaluation of high-utility rules individually

In this strand of experiments we take a deep look at a representative excerpt of data sets. We present our findings along two different perspectives: i) the quantitative perspective backed from statistical-based evidence that we provide from high-utility sequential rules, and ii) the qualitative perspective building upon a careful human oriented interpretation of high-utility sequential rules for each data set. Under the qualitative (human oriented interpretation) perspective we craft two different experimental settings keeping one of the thresholds (*min_utility* or *min_confidence*) fixed and varying the other one. In the first setting, namely Setting 1 we keep *min_confidence* fixed and vary *min_utility*. Conversely, in Setting 2 we keep *min_utility* fixed and vary *min_confidence*.

Results for Data set D1

Quantitative results In terms of quantitative results, high-utility sequential rules obtained from data set D1, suggest that the most frequently encountered emotions in the antecedent side are $anger$ and $shame$ occurring 83% of the time each, vs the consequent side where the most frequent emotion is $fear$ occurring 67% of the time (also in Figure 4a). The longest high-utility sequential rule for D1 - the multiuser data set is $\{joy, anger, sadness, guilt\} \rightarrow \{disgust\}$. While the shortest one is $\{joy, anger, shame\} \rightarrow \{fear\}$. Interestingly enough, this rule is also the rule with the highest utility *#UTIL*. Additionally, D1 generates a collection of high-utility sequential rules, where $anger$ and $shame$ are the most dominating emotions, followed by joy and $fear$.

Qualitative Results In terms of qualitative results, high-utility sequential rules obtained from data set D1 (multi-user) suggest the following conclusions.

Setting 1 keeping a fixed value of *min_confidence* = 0.6 and varying *min_utility* ∈ [2,7]:

- Emotion $fear$ appears only on the consequent side

- Emotions *anger* or *sadness* or *shame* or *guilt* appear only on the antecedent side
- Whenever {*anger*, .., *shame*} appear among the antecedents they always produce the emotion *fear*
- Whenever {*joy*, *anger*, *shame*} appear on the antecedent side they always produce the emotion *fear*
- Whenever *disgust* is on the antecedent side the consequent is *fear*
- Whenever *anger* and *sadness* co-occur among the antecedent items, then they can trigger *disgust* or *joy*

Setting 2 keeping a fixed value of $min\_utility = 4$ and varying $min\_confidence \in [0.6\text{-}0.9]$:

- Emotion *fear* appears only on the consequent side, whereas *anger* or *sadness* or *shame* or *guilt* emotions do not appear at all on the consequent side
- When *disgust* appears on the consequent side the *#UTIL* tends to be higher
- Whenever {*anger*, *sadness*, *guilt*} co-occur on the antecedent side then the consequent is always *disgust*
- Whenever {*joy*, ..., *disgust*, *shame*} co-occur on the antecedent side then the consequent is always *fear*

Results from Data set D6

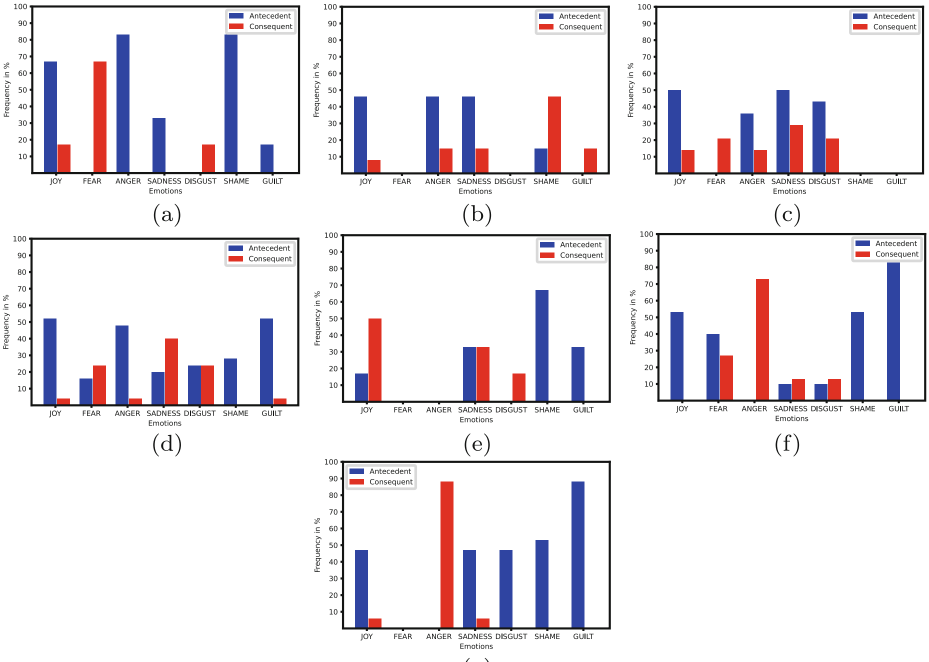

Fig. 4 Emotion frequency on the antecedent and consequent based on the total distinct rules per data set in percentage (a) D1, (b) D2, (c) D3, (d) D4, (e) D5, (f) D6, (g) D7

Quantitative Results In terms of quantitative results, high-utility sequential rules obtained from data set D6, suggest that the most frequently encountered emotions on the antecedent side are $guilt$ and joy occurring 83% and 53% of the time respectively, vs the consequent side where the most frequent emotion is $anger$ occurring 73% of the time (see Fig. 4f). The longest high-utility sequential rule for D6 - a singleuser data set, is $\{joy, fear, disgust, shame, guilt\} \rightarrow \{anger\}$. While the shortest one is $\{guilt\} \rightarrow \{anger\}$. $\{joy\} \rightarrow \{sadness\}$ is the rule with the highest utility *#UTIL*. Additionally, D6 generates a collection of high-utility sequential rules, where $guilt$ and $anger$ are the most dominating emotions, followed by joy and $shame$.

Qualitative Results In terms of qualitative results, high-utility sequential rules obtained from data set D6 suggest the following conclusions:

Setting 1 keeping a fixed value of *min_confidence* = 0.3 and varying *min_utility* ∈ [4,10]:

- Emotion $anger$ appears on the consequent side when *#CONF* is high.
- Emotion $anger$ does never appear on the antecedent side
- When *#CONF* goes below 0,35 emotion $fear$ appears on the consequent side
- When fine-tuning *min_utility* the most present emotion on the consequent side is $anger$
- joy or $disgust$ or $shame$ or $guilt$ do not appear on the consequent side
- When joy and $fear$ co-occur, the consequent emotion is always $anger$
- When $guilt$ is on the antecedent side, then the consequent is always $anger$
- When $fear$ is among the antecedent items, then the consequent is always $anger$
- When $disgust$, $shame$, $guilt$ co-occur on the antecedent side, they lead to emotion $fear$ or $anger$ or $sadness$ or $fear$, $anger$ on the consequent side
- Co-occurring joy, $fear$, $guilt$ on the antecedent side lead to highest value of rule confidence *#CONF*=1

Setting 2 keeping a fixed value of *min_utility* = 6 and varying *min_confidence* ∈ [0.4-0.7]:

- The presence of emotion $guilt$ on the antecedent side decreases *#UTIL*
- The presence of emotion DISGUST on the antecedent side increases *#CONF*
- $anger$ or $sadness$ are the only emotions on the consequent side
- Co-occurring emotions $shame$ and $guilt$ trigger two different emotions $anger$ or $sadness$
- Emotion $anger$ does not appear on the antecedent side
- $\{joy, fear, disgust, shame, guilt\} \rightarrow \{anger\}$ is the rule with the highest confidence *#CONF*=1.

Results from Data set D7

Quantitative Results In terms of quantitative results, high-utility sequential rules obtained from data set D7, suggest that the most frequently encountered emotions in the antecedent side is $guilt$ occurring 88% of the time, vs the consequent side where the most frequent emotion is $anger$ occurring again 88% of the time (see Fig. 4g). The longest high-utility sequential rule for D7 - a singleuser data set, is $\{joy, sadness, disgust, shame, guilt\} \rightarrow \{anger\}$. While the shortest one is $\{shame\} \rightarrow \{joy\}$. Interestingly enough, this rule is also the rule with the highest utility *#UTIL*. Additionally, D1 generates a collection of high-utility sequential rules, where $anger$ and $guilt$ are the most dominating emotions.

Qualitative Results In terms of qualitative results, high-utility sequential rules obtained from data set D7 suggest the following conclusions:

Setting 1 keeping a fixed value of *min_confidence* = 0.4 and varying *min_utility* ∈ [4,10]:

- Emotions *joy* or *sadness* on the consequent side lead to a very high *#UTIL* but low *#CONF*
- Whenever *disgust* appears on the consequent side, then the *#CONF*=1
- While increasing *min_utility* in the interval [2-10], the *#CONF* of the rules decreases
- Emotion *fear* does not appear at all (both sides)
- Emotion *anger* appears only as a consequent emotion
- Emotion *disgust* or *shame* or *guilt* do not appear as consequent emotions
- Emotion *shame* as a antecedent item triggers two different emotions *joy* or *sadness*
- Coexisting emotions *joy* and *guilt* lead to emotion *anger*
- Coexisting *sadness* and *guilt* lead to emotion *anger*
- Whenever *guilt* is on the antecedent side, then the consequent is always *anger*
- For values of $\min\_utility > 6$ emotion *disgust* does not appear while *sadness* shows only on the consequent side
- Increasing utility, the number of emotions on both sides shrinks to one on each side

Setting 2 keeping a fixed value of *min_utility* = 6 and varying *min_confidence* ∈ [0.4-0.7]:

- The increase of *min_confidence* leads to rules with higher *#CONF*
- Increasing the *min_confidence* leads to the increase of the number of antecedents and the consequent is dominantly emotion *anger*
- Increasing *min_confidence*, emotion *joy* or *sadness* is eliminated from the consequent side, and *anger* becomes the predominant emotion on the consequent side
- Emotion *fear* does not appear at all on both sides
- Emotion *anger* does not appear on the antecedent side
- At low values of *min_confidence* emotion *shame* triggers *joy* or *sadness*, at higher *min_confidence* the presence of *shame* on the antecedent side leads to the appearance of emotion *anger*

5.2.4 Comparison of high-utility rules of an individual user versus a multi-user pool

The experimental work involving data sets D1-D7, produced a significant amount of results, posing a need to formalize them further into conclusions and generalize results to extended contexts of the same nature. In that sense, we take a comparative approach here, considering the multi-user data set D1 as a baseline and the remaining six single-user data sets D2-D7 as competitive targets.

The results tabulated in Tables 8 and 9, allow us to assess the sequential rules drawn from data sets D1-D7, in terms of their confidence and utility metrics. We focus our comparative analysis on data sets' top three ranking HUSR-s for each metric. There are several observations we can make.

First, a multi-user oriented analysis seems to be less informing than a single-user one, as the resulting HUSR-s from D1 (the multiuser dataset) have scoring metrics which are predominantly lower than those of the single-user data sets. This is in large part attributed to the fact that several users together may convey different emotional status in a given time-window, which may not necessarily co-occur, leading to emotional behaviours with

Table 8 Top 3 High-utility sequential rules with respect to *#UTIL*

Data set ID	Antecedent	Consequent	*#CONF*	*#UTIL*
D1	JOY, ANGER, SHAME	FEAR	0.67	8
	JOY, ANGER, SADNESS, GUILT	DISGUST	1.00	7
	JOY, ANGER, DISGUST, SHAME	FEAR	1.00	5
D2	JOY	ANGER	0.35	21
	JOY, ANGER	SHAME	0.25	6
	JOY, ANGER	SADNESS	0.25	6
D3	JOY	DISGUST	0.29	25
	JOY	ANGER	0.24	21
	JOY, ANGER, DISGUST	SADNESS	0.50	14
D4	DISGUST	ANGER	0.37	22
	DISGUST	SADNESS	0.33	20
	JOY, ANGER, SHAME	FEAR	0.67	8
D5	SHAME	JOY	0.46	272
	SHAME	SADNESS	0.42	240
	JOY	SADNESS	0.36	186
D6	JOY	SADNESS	0.30	99
	JOY, FEAR, SADNESS	ANGER	0.32	29
	JOY, FEAR, SADNESS, SHAME	ANGER	0.50	25
D7	SHAME	JOY	0.50	283
	SHAME	SADNESS	0.49	268
	JOY, SHAME, GUILT	ANGER	0.50	8

insignificant or no statistical evidence at all. Second and not surprisingly, there are still pairs of users who share common emotional behaviours (expressed in terms of HUSR-s in common) as evidenced in Table 10. It is clearly noted that the users behind data sets D6 and D7 respectively have in common three distinct emotional patterns that always lead to *anger* related reactions (posts). Similarly, users behind D2 and D3 also have a strong correlation encapsulated in five emotion-rich HUSR-s in common. While, users behind D5 and D7 are loosely correlated by means of only one HUSR in common $\{shame\} \rightarrow \{joy\}$. Along similar lines, radar charts in Fig. 3 suggest notable similitude between data sets D6 and D7 (having a lot of overlapping areas in common) in terms of frequency, density and variety of emotions encountered in their corresponding high-utility sequential rules. Third, the HUSR mining approach seems to be more effective as the size of the data sets increases, outpacing the impact of their temporal density increase. Fourth, looking at Table 8, the shorter the HUSR, the higher becomes its utility score, implying more interesting associations behind these rules. In terms of rule utility, rules from data sets D7, D5 followed by D6 are topping the list, surpassing the baseline D1. In terms of rule confidence, rules from data sets D4, D6 and D7 are the ones that reach the maximum value of *#CONF* = 1, similar to the baseline D1 (see Table 9). Fifth, to further explore and uncover hidden data patterns from our user-level sequential rules, we propose a new metric called *cumulative rule utility*. For each user/dataset we calculate the cumulative sum of utilities of distinct rules, whose consequent emotion is joy, $fear$, $anger$, $sadness$, $disgust$, $shame$ and $guilt$ respectively. Results are

Table 9 Top 3 High-utility sequential rules with respect to *#CONF*

Data set ID	Antecedent	Consequent	*#CONF*	*#UTIL*
D1	JOY, ANGER, SADNESS, GUILT	DISGUST	1.00	7
	JOY, ANGER, DISGUST, SHAME	FEAR	1.00	5
	JOY, DISGUST, SHAME	FEAR	1.00	4
D2	JOY	ANGER	0.35	21
	JOY, ANGER, SADNESS	SHAME	0.33	4
	JOY, SHAME	GUILT	0.33	3
D3	JOY, ANGER, DISGUST	SADNESS	0.50	14
	ANGER, DISGUST	SADNESS	0.33	7
	JOY, ANGER, SADNESS, DISGUST	FEAR	0.33	5
D4	JOY, ANGER, SADNESS, GUILT	DISGUST	1.00	7
	JOY, ANGER, DISGUST, SHAME	FEAR	1.00	5
	JOY, DISGUST, SHAME	FEAR	1.00	4
D5	SADNESS, SHAME, GUILT	JOY	0.50	9
	SADNESS, GUILT	JOY	0.50	7
	SHAME	JOY	0.46	272
D6	JOY, FEAR, DISGUST, SHAME, GUILT	ANGER	1.00	6
	JOY, FEAR, DISGUST, GUILT	ANGER	1.00	5
	JOY, FEAR, GUILT	ANGER	1.00	4
D7	JOY, SADNESS, DISGUST, SHAME, GUILT	ANGER	1.00	6
	JOY, SADNESS, DISGUST, GUILT	ANGER	1.00	5
	JOY, DISGUST, SHAME, GUILT	ANGER	1.00	5

tabulated in Table 11, noting in bold the highest value for each dataset. We suggest to interpret high values of cumulative rule utility as indicators of emotions (emotive posts) that produce more effect. For example, for users behind datasets D3, D4 and D5, *sadness* is the emotion that produces more effect. Consequently, D5 reporting the highest value among the three, may be subjected to further analysis or profiling. Next, for D2 and D6 the emotion with the highest effect is *anger*, notably more influential for D6 due its higher *cumulative rule utility* value. Moreover, we believe this metric deserves further investigation as future

Table 10 Data sets having high-utility sequential rules in common

Data set pairs		High-Utility sequential rules
D6	D7	$\{GUILT\} \rightarrow \{ANGER\}$
D6	D7	$\{ANGER, DISGUST, SHAME, GUILT\} \rightarrow \{ANGER\}$
D6	D7	$\{DISGUST, SHAME, GUILT\} \rightarrow \{ANGER\}$
D2	D3	$\{JOY, ANGER\} \rightarrow \{SADNESS\}$
D2	D3	$\{JOY\} \rightarrow \{ANGER\}$
D2	D3	$\{ANGER\} \rightarrow \{SADNESS\}$
D2	D3	$\{SADNESS\} \rightarrow \{JOY\}$
D2	D3	$\{SADNESS\} \rightarrow \{ANGER\}$
D5	D7	$\{SHAME\} \rightarrow \{JOY\}$

Table 11 Cumulative rule utility based on the consequent emotion of distinct rules

SUM(rule_util) for each Consequent Emotion							
Data set ID	JOY	FEAR	ANGER	SADNESS	DISGUST	SHAME	GUILT
D1	4	**21**	0	0	7	0	0
D2	2	0	**23**	10	0	22	5
D3	13	12	27	**39**	35	0	0
D4	4	31	22	**53**	33	0	4
D5	288	0	0	**426**	185	0	0
D6	0	16	**157**	114	0	0	0
D7	**283**	0	70	268	0	0	0

work, because it may be used to measure users distance along with other traditional metrics and help discover sentiment communities.

6 Conclusions

The analysis of micro-blogging content through frequent pattern mining is in itself a significant and challenging problem because it combines the extraction of information from unstructured data and the discovery of regularities from structured data. In the task investigated in this paper, we tackled two additional complexity degrees, that is, the emotional content of social media postings and the use of ’quantitative’ patterns. Indeed, high-utility patterns have been considered to reduce the information loss concerning the different emphasis with which distinct emotions can be voiced. A further complexity degree lies in the study of the ’time-aware’ implication of the emotions, which allows us to unearth insights on the temporal consequence rather than the mere co-occurrence. This represents an innovation point compared to many works, equally focused on the implication, but tailored for classic association rules. Likewise association rules, the HUSR are very close to natural language interpretations in a straightforward way, even without having a-priori information about the problem.

Methodologically, the proposed computational solution is organized in three main steps. The first and second steps are in charge of i) handling the unstructured information of the social postings and ii) recognizing the mentions of the emotions. Clearly, this part resorts to emotion detection models that require supervised sentences specific of the language. However, this does not constraint the whole framework to be adapted for other languages. The third step implements a purely unsupervised data mining task, which does require no adaptation.

From the experimental viewpoint, we focused on the sequences of emotional status associated to individual users (user-level) rather than working on summarizing sequences (community-level). Indeed, this allows us to preserve the original occurrences of the emotions by reducing the risk of distribution drifts that aggregations of the occurrences would have introduced.

From the application viewpoint, we explored the viability of the approach in the context of social postings with political content, which turns out to be appealing and challenging since it is a strong collector of opinions and emotions. We further introduced a new metric, *cumulative rule utility*, which we believe will open new avenues of future work, pointing to user distance measuring and sentiment community discovery.

However, the psychological processes underlying how one voices emotions on social media are so complex that, in this work, we were forced to take some exemplifications. The first one is the size of the time-window, whose value strongly depends on the nature of the input messages. We plan to apply the proposed approach to social messages with different topics (e.g., tweets related to Covid pandemic) and explore the influence of the several time-window sizes. Another exemplification is the one that assigns only one emotion to one sentence, while each sentence can express different emotions. We believe that this aspect can be further studied through solutions of Multi-label classification (de Almeida et al. 2018). Finally, social messages are produced at a high-rate and should be acquired as elements of a potentially unbounded flow, which requires the adaptation to the streaming discovery of rules (Ceci et al. 2009) or the integration of background emotion hierarchies to discover multi-level rules (Loglisci and Malerba 2009).

Data Availability The source code and the datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Akiyama, K., Kumamoto, T., Nadamoto, A. (2017). Emotion-based method for latent followee recommendation in twitter. In Indrawan-Santiago, M., Steinbauer, M., Salvadori, I.L., Khalil, I., Anderst-Kotsis, G. (Eds.) *Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2017, Salzburg, Austria, December 4-6, 2017* (pp. 121–125): ACM.

Ali, S.M., Noorian, Z., Bagheri, E., Ding, C., Al-Obeidat, F.N. (2020). Topic and sentiment aware microblog summarization for twitter. *Journal of Intelligent Information System*, *54*(1), 129–156.

Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. *Journal of Intelligent Information System*, *55*(1), 51–66.

Bing, L., Chan, K.C.C., Ou, C.X. (2014). Public sentiment analysis in twitter data for prediction of a company's stock price movements. In *11th IEEE International Conference on e-Business Engineering, ICEBE 2014* (pp. 232–239). Guangzhou.

Ceci, M., Appice, A., Loglisci, C., Caruso, C., Fumarola, F., Malerba, D. (2009). Novelty detection from evolving complex data streams with time windows. In Rauch, J., Ras, Z.W., Berka, P., Elomaa, T. (Eds.) *Foundations of Intelligent Systems, 18th International Symposium, ISMIS 2009, Prague. Proceedings, Lecture Notes in Computer Science*, (Vol. 5722 pp. 563–572): Springer.

Choi, H.-J., & Park, C.H. (2019). Emerging topic detection in twitter stream based on high utility pattern mining. *Expert Systems with Applications*, *115*, 27–36.

de Almeida, A.M.G., Cerri, R., Paraiso, E.C., Mantovani, R.G., Junior, S.B. (2018). Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, *320*, 35–46.

Dehkharghani, R., Mercan, H., Javeed, A., Saygin, Y. (2014). Sentimental causal rule discovery from twitter. *Expert Systems with Applications*, *41*(10), 4950–4958.

Diaz-Garcia, J.A., Ruiz, M.D., Martín-Bautista, M.J. (2020). Non-query-based pattern mining and sentiment analysis for massive microblogging online texts. *IEEE Access*, *8*, 78166–78182.

dos Santos, C.N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In Hajic, J., & Tsujii, J. (Eds.) *COLING 2014, 25th international conference on computational linguistics, proceedings of the conference: Technical papers, august 23-29, 2014, dublin, ireland* (pp. 69–78): ACL.

Ekman, P. (1993). Facial expression and emotion. *The American psychologist*, *48*, 384–92.

Fan, C., Yuan, C., Du, J., Gui, L., Yang, M., Xu, R. (2020). Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (pp. 3707–3717).

Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., Tseng, V.S. (2014). Spmf: a java open-source pattern mining library. *The Journal of Machine Learning Research*, *15*(1), 3389–3393.

Gan, W., Lin, J. C.-W., Fournier-Viger, P., Chao, H.-.C., Fujita, H. (2018). Extracting non-redundant correlated purchase behaviors by utility measure. *Knowl. Based Syst.*, *143*, 30–41.

Gan, W., Lin, J. C.-W., Fournier-Viger, P., Chao, H.-C., Hong, T.-P., Fujita, H. (2018). A survey of incremental high-utility itemset mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, *8*(2).

Gao, K., Xu, H., Wang, J. (2015). A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, *42*(9), 4517–4528.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Hai, Z., Chang, K., Kim, J. (2011). Implicit feature identification via co-occurrence association rule mining. In Gelbukh, A.F. (Ed.) *Computational linguistics and intelligent text processing - 12th international conference, cicling 2011. proceedings, part I, Lecture Notes in Computer Science*, (Vol. 6608 pp. 393–404). Tokyo: Springer.

Han, J., Pei, J., Yin, Y., Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, *8*(1), 53–87.

Huang, J., Peng, M., Wang, H. (2015). Topic detection from large scale of microblog stream with high utility pattern clustering. In Kacimi, M., Preda, N., Ramanath, M. (Eds.) *Proceedings of the 8th Workshop on Ph.D. Workshop in Information and Knowledge Management, PIKM 2015* (pp. 3–10). Melbourne: ACM.

Kang, X., Ren, F., Wu, Y. (2018). Exploring latent semantic information for textual emotion recognition in blog articles. *IEEE CAA J. Autom. Sinica*, *5*(1), 204–216.

Kim, Y. (2014). Convolutional neural networks for sentence classification.

Lo, D., Khoo, S.-C., Wong, L. (2009). Non-redundant sequential rules - theory and algorithm. *Information Systems*, *34*(4-5), 438–453.

Loglisci, C., & Malerba, D. (2009). Mining multiple level non-redundant association rules through two-fold pruning of redundancies. In Perner, P. (Ed.) *Machine Learning and Data Mining in Pattern Recognition, 6th International Conference, MLDM 2009. Proceedings, Lecture Notes in Computer Science*, (Vol. 5632 pp. 251–265). Leipzig: Springer.

Mamgain, N., Pant, B., Mittal, A. (2016). Categorical data analysis and pattern mining of top colleges in india by using twitter data. In *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 341–345).

Mohammad, S.M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301–326.

Sano, Y., Takayasu, H., Havlin, S., Takayasu, M. (2019). Identifying long-term periodic cycles and memories of collective emotion in online social media. *PLOS ONE*, *14*(3), 1–17.

Simsek, A., & Karagoz, P. (2020). Wikipedia enriched advertisement recommendation for microblogs by using sentiment enhanced user profiles. *Journal of Intelligent Information System*, *54*(2), 245–269.

Skenduli, M.P., & Biba, M. (2020). Classification and clustering of emotive microblogs in albanian: Two user-oriented tasks. In Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (Eds.) *Complex Pattern Mining: New Challenges, Methods and Applications* (pp. 153–171). Cham: Springer International Publishing.

Skenduli, M.P., Biba, M., Loglisci, C., Ceci, M., Malerba, D. (2018). User-emotion detection through sentence-based classification using deep learning: A case-study with microblogs in albanian. In Ceci, M., Japkowicz, N., Liu, J., Papadopoulos, G.A., Ras, Z.W. (Eds.) *Foundations of Intelligent Systems - 24th International Symposium, ISMIS 2018, Proceedings, Lecture Notes in Computer Science*, (Vol. 11177 pp. 258–267). Limassol: Springer.

Skenduli, M.P., Loglisci, C., Ceci, M., Biba, M., Malerba, D. (2018). An empirical evaluation of sequential pattern mining algorithms. In Barolli, L., Xhafa, F., Javaid, N., Spaho, E., Kolici, V. (Eds.) *Advances in Internet, Data & Web Technologies, The 6th International Conference on Emerging Internet, Data & Web Technologies, EIDWT-2018, Lecture Notes on Data Engineering and Communications Technologies*, (Vol. 17 pp. 615–626). Tirana: Springer.

Tzacheva, A.A., Ranganathan, J., Bagavathi, A. (2020). Action rules for sentiment analysis using twitter. *Int. J. Soc. Netw. Min.*, *3*(1), 35–51.

Wen, S., & Wan, X. (2014). Emotion classification in microblog texts using class sequential rules. In Brodley, C.E., & Stone, P. (Eds.) *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 187–193). Québec City: AAAI Press.

Yada, S., Ikeda, K., Hoashi, K., Kageura, K. (2017). A bootstrap method for automatic rule acquisition on emotion cause extraction. In *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017* (pp. 414–421).

Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 325–335). Baltimore: Association for Computational Linguistics.

Yang, J., Wang, Z., Di, F., Chen, L., Yi, C., Xue, Y., Li, J. (2017). Propagator or influencer?: A data-driven approach for evaluating emotional effect in online information diffusion. In Diesner, J., Ferrari, E., Xu, G. (Eds.) *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 836–843). Sydney: ACM.

Yuan, M., Ouyang, Y., Sheng, H. (2014). Investigating association rules for sentiment classification of web reviews. *Journal of Intelligent Fuzzy Systems*, *27*(4), 2055–2065.

Zida, S., Fournier-Viger, P., Wu, C.-W., Lin, J.C.-W., Tseng, V.S. (2015). Efficient mining of high-utility sequential rules. In Perner, P. (Ed.) *Machine Learning and Data Mining in Pattern Recognition - 11th International Conference, MLDM 2015, Proceedings, Lecture Notes in Computer Science*, (Vol. 9166 pp. 157–171). Hamburg: Springer.

Zida, S., Fournier-Viger, P., Wu, C.-W., Lin, J.C.-W., Tseng, V.S. (2015). Efficient mining of high-utility sequential rules. In *International workshop on machine learning and data mining in pattern recognition* (pp. 157–171): Springer.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.