Journal of Internet Services and Applications, 2022, 1:1, doi: 10.5753/jisa.2022.2520

Characterization of the Mobile User Profile Based on Sentiments and Network Usage Attributes

Leonardo P. de Morais [**Universidade Federal de Goiás** | ***leonardo.morais@egresso.ufg.br***]
Roger Immich [**Universidade Federal do Rio Grande do Norte** | ***roger@imd.ufrn.br***]
Nádia Félix Silva [**Universidade Federal de Goiás** | ***nadia.felix@ufg.br***]
Thierson Couto Rosa [**Universidade Federal de Goiás** | ***thierson@ufg.br***]
Vinicius da Cunha Martins Borges [**Universidade Federal de Goiás** | ***vcmborges@ufg.br***]

*Universidade Federal de Goiás - UFG - Alameda Palmeiras, Quadra D, Câmpus Samambaia, Instituto de Informática,
Goiânia, GO, 74001-970, Brazil*

Received: 21 February 2022 • **Accepted:** 18 August 2022 • **Published:** 27 December 2022

Abstract Providing resources to meet user needs in futuristic mobile networks is still challenging since the network resources like spectrum and base stations do not increase in the same proportion as the accelerated growth of network traffic. Because of this, human/user behavior attributes can assist resource management in dealing with these challenges, which pick up aspects of how the user impacts the usage of mobile networks, such as network usage, the content of interest, urban mobility routines, social networks, and sentiment. A user profile is a combination of user/human behavior attributes. Such profiles are expected to be a knowledge for softwarization enablers to improve the management of future wireless networks fully. Nevertheless, the correlation between human sentiment and wireless and mobile network usage has not been deeply investigated in the literature about the mobile user profile. This work aims to define the user profile using a transfer learning approach for the sentiment classification of WhatsApp messages. A real-life experiment was conducted to collect users' attributes, namely the WhatsApp messages and network usage. A new data analysis methodology is proposed that consists of a frequent item-set pattern mining (FP-Growth) based on Association Rules, the Chi-squared statistical test, and descriptive statistics. This methodology assesses the correlation between sentiment and network usage in a profound way. Results show that the users participating in the experiment form three groups. The first group, with 55.6% of the users, contains users who present a strong relation between negative sentiment and low network usage and also a strong relation between positive sentiment and high network usage. The second group contains 25.9% of the users and is composed of users who present a strong relation between positive sentiment and high network usage. The third group contains 18.5% of the users for whom the correlation between sentiment and network usage is still statistical significant, but the strength of this relation is much more weak then in the other two groups. Thus, 81.5% of the users (the first two groups) present a strong relation between user sentiment captured from WhatsApp messages and the network traffic generated by them.

Keywords: Future Mobile Networks, Sentiment Analysis, User Profile, Association Rules, Frequent Item-set Mining

1 Introduction

The next generations of wireless and mobile telecommunications should guarantee a satisfactory and personalized experience to users in a variety of scenarios, which proves to be challenging [Wang *et al.*, 2017; Taleb *et al.*, 2017; Singh and Sharma, 2019]. These networks have to support heterogeneous wireless connections and distinct radio communication technology to provide high bandwidth rates, low latency, and seamless mobility [Qamar *et al.*, 2019]. Furthermore, they also have to enable new categories of services, allowing new business opportunities for the network operators [Giraldo-Rodríguez *et al.*, 2015; Afolabi *et al.*, 2018]. In this context, the fifth generation of mobile networks (5G) provides support for a plethora of applications, such as Virtual Reality (VR) [Erol-Kantarci and Sukhmani, 2018] and Internet of Things (IoT) [Li *et al.*, 2018], and also new scenarios, namely the development of autonomous cars [Raissi *et al.*, 2019] and Intelligent Transportation System (ITS) [Akabane *et al.*, 2019] as well as broad support in health area [Thuemmler *et al.*, 2018].

Software Defined Radio (SDR), Software Defined Networking (SDN), and Network Function Virtualization (NFV) are network softwarization technologies essential to enable 5G networks as well as newer versions. These technologies enable to control and program the mobile network resources and services flexibly and intelligently. This is performed using machine learning techniques, software applications, and virtualized network functions in the hardware commodity of the cloud, in which the networks will not be dependent on physical, closed, and monolithic network functions [Cho *et al.*, 2014; Bouras *et al.*, 2017; Yousaf *et al.*, 2017; Neto *et al.*, 2021].

Despite network softwarization technologies and novel higher capacity radio communication technologies, there are several challenging situations for the future mobile networks to support high communication quality, which include: a wide range of applications with different requirements, the constant upsurge in network traffic, the increasing number of connected devices as well as the demand for personalized services, and the need to reduce costs, just to name a few examples. In light of this, specific aspects of user behavior can assist resource management in dealing with these challenges. These aspects pick up characteristics of how the user impacts the usage of mobile networks, such as network usage, signal quality, wireless technology (e.g., Wi-Fi, LTE), and the content of interest. Nevertheless, attributes of human behavior, such as the urban mobility routines, social networks, sentiment, places (e.g., work and study routines), and times of entertainment, can also influence these user behavior aspects and, consequently, services and resources management policies [Taleb *et al*., 2017; Akabane *et al*., 2018; dos Santos and Lopes, 2019].

A user profile is the summary/combination of user and human behavior attributes, such as users' interests, characteristics, and preferences. User profiling is the system of collecting, organizing, and inferring the user profile information [Naboulsi *et al*., 2016; Zhao *et al*., 2019; Eke *et al*., 2019]. It is important to notice that different user profiles within the same network cell can demand distinct network resources depending on the traffic and mobility of the users, as well as their contextual and behavioral demands. The Mobile Network Operators (MNOs) can take advantage of the knowledge of the user profiles based on human behavior to manage the network resources and services better. This allows exploring the flexibility provided by the network softwarization technologies to allocate personalized and optimized resources. For instance, the SDN control plane can be defined intelligently according to user profiles. Therefore, the profiles will guide how to control the network resources. In doing that, MNOs can lead to greater satisfaction of mobile users without decreasing monetary gains or increasing their costs [Bouras *et al*., 2017]. By exploring the users' attributes and their personal relationships, the accuracy of predicting future events on wireless and mobile networks can be significantly improved [Kabir *et al*., 2018].

Several works investigate the user profile to optimize network configurations. Most of these works focus on characteristics of the user behavior [Naboulsi *et al*., 2014; Giraldo-Rodríguez *et al*., 2015; M. Borges *et al*., 2015; Furno *et al*., 2017; Zhao *et al*., 2017; Qiao *et al*., 2018; Vamvakas *et al*., 2019; Ullah and Binbusayyis, 2022], such as call detail records, mobile data traffic usage/demand, content demand, wireless technology, and used applications. Some works focus on human behavior based on frequently visited websites, online social friendship, user mobility, targeted ads, and anomaly detection [Tandon and Chan, 2009; Wu *et al*., 2017; Leng *et al*., 2015; Li *et al*., 2019; Chan *et al*., 2019; Sakouhi and Akaichi, 2021; Awwad, 2021; Ullah and Binbusayyis, 2022]. Moreover, ringer modes and personality traits are other attributes of human behavior that are investigated in characterizing the user profile [Peltonen *et al*., 2020; Komninos *et al*., 2021]. Nevertheless, no previous work has considered including the correlation between user sentiment and network usage to define the user profile. The term sentiment has several psychological and physiological definitions. They range from subjectively accessible emotions to somatosensory experiences, ideas, and beliefs. These internal sensations organize the human being's mental life and are also vital signs of well-being [Nummenmaa *et al*., 2018]. Human beings are daily subjected to several emotions that give rise to sentiments that can be analyzed and classified.

The correlation between user sentiment and the amount of data transmitted/received in the network is a possible way of showing how patterns of human behavior can be related to the use of wireless and mobile networks. By using the knowledge of user sentiment and details of the network usage, the users' behavior patterns can be mapped and grouped into similar profiles. To take advantage of this, this work investigates whether the users' sentiment is related to the use of the wireless network. More specifically, we investigate whether users with a positive sentiment tend to consume more or less data or if users with a negative sentiment tend to isolate themselves and consequently consume fewer data. Therefore, the definition of these profiles will be of great value for the decision-making process of allocating and optimizing the use of network resources.

This work aims to define new mobile user profiles for next-generation wireless networks through the classification of sentiment they express through WhatsApp messages. It studies the correlation of sentiment with the network usage of wireless and mobile networks. To do that, text pattern description and recognition techniques based on frequent itemset mining, association rules algorithms, and Transfer Learning (Artificial Intelligence) will be applied. To the best of our knowledge, this is the first work to characterize mobile user profiles, considering the sentiment and the network usage.

The main contributions of this work are as follows: (1) transfer learning sentiment classification of WhatsApp messages; (2) the creation of a database combining factual information on the use of the wireless network and WhatsApp messages, collected from the volunteers participating in the research; (3) a data analysis methodology based on statistic methods and the FP-Growth pattern mining algorithm as well as user profiles based on the correlation of wireless usage data and user sentiments.

The remainder of this work is structured as follows. Section 2 presents the related work. The description of the main concepts and background are given in Section 3. The proposed characterization of the user profiling and the performed assessments are described in Section 4. Furthermore, the conclusion and future work are presented in section 5.

2 Related Work

User profile works that consider the user sentiment for the optimization of mobile networks are scarce. This section analyses the few related works found in this research subject.

The Big Data Driven (BDD) model's primary goal was to collect data from the user and the network in general to build a Big Data database to optimize 5G networks [Zheng *et al*., 2016]. This database was then used to analyze the user's behavior, allowing the tailoring of the network according to it. This big data-driven model refereed to Mobile Network Operators (MNO) with long and short-term strategies to make decisions about resources and services based on the analysis and interpretations of relevant network information. The user sentiment was used for two purposes. Firstly, the user sentiment is employed in order to define the user profile based on the Quality of Experience (QoE) modeling, for example, how various non-technical factors exist that may influence QoE results, such as device type, user emotion, habit, and expectation. Secondly, it was adopted to predict where and how users may use the mobile network. For instance, predicting a marathon event, where some streets become highly dense scenarios (large crowds of people). Therefore, resulting in highly congested traffic in these places during the event period.

The authors proposed three case studies to demonstrate that the model can overcome the challenges of 5G networks, namely a study of the traffic behavior of the user, and the implementation of a personalized cache service, and the assessment of the end-users quality of experience. In the latter case, the authors use machine learning techniques to classify user sentiment and deliver a personalized service. Nevertheless, these use cases are conceptual proposals where it is possible to notice some issues. For example, the article does not mention the adopted algorithms and machine learning techniques and why they were chosen in QoE Modeling, specifically for sentiment analysis or user emotion detection. Nonetheless, that work does not investigate the user profile based on the correlation between the user sentiment and network traffic usage. It does not explain how the dataset was created to make it possible to infer user sentiment, mobility, and social network. We aim to investigate the definition of a new mobile user profile based on the correlation between user sentiment and network traffic usage in a more generic way. Thus, we provide knowledge that can aid the MNO in optimizing the network resources and services more precisely.

The proposed model in [Chen *et al.*, 2015] offered personalized services through cloud computing and affective computing recognizing the user's sentiment/emotion in the user's home in order to provide enhanced QoE, called Emotion-aware Mobile cloud Computing (EMC). The user emotion can be recognized from text, image, video, and other emotional data (multi-dimensional emotional data) collected by mobile terminals using affective computing techniques. Furthermore, they also seek to reach a trade-off between local cloudlet (mobile devices) and remote cloud (computers with higher computing capacity) in order to perform smaller or larger tasks, respectively, i.e., Less 5G Support and Larger Terminal Workload/More 5G Support and Smaller Terminal Workload.

On the one hand, if there is more bandwidth-intensive delivery supported by 5G, it will be more effective to collect and transmit a large variety and volume of emotional data, especially for multimedia data with big volumes. Thus, the personalized services for the users are more complete and efficient. On the other hand, the QoE level may be relatively lower with less 5G support and a larger terminal workload, which is weakly dependent on 5G, therefore ensuring that EMC offers the basic services even in a poor wireless network resource scenario, such as medical emergency handling and telehealth education. However, the authors did not reveal which affective techniques were used to detect the user's emotion. This work also lacks to explain the results of such techniques better. In addition, the user profile is poor and superficially described, i.e., how can the home care services based on emotions be adapted according to the user emotion for each user's mental status and diseases? Besides, the user profile is based only on a single attribute (user emotion); the authors do not investigate a correlation between user emotion and network traffic usage to define mobile user profiles.

The User Aware Edge Caching (UAEC) mechanism predicts the intensity of human emotions through social media (e.g., Twitter) to optimize the cache allocation [Kabir *et al.*, 2018]. It can do so by classifying and assessing the characteristics that are related to human behaviors. UAEC also optimizes the content of cache servers based on the sentiment. The authors adopted a combination of Hidden Markov Models (HMM) and Machine Learning for the development of their algorithm for the user's sentiment classification. However, it is not clear the reasons for this combination nor why these techniques were chosen. Moreover, this work uses only Twitter as input data. Furthermore, this work does not seek a correlation between user sentiment and network usage.

As can be seen, few related works take into account the sentiment analysis in the user profile. Besides, these works do not mention the used technique to classify the sentiment or the analyzed social network, they do not attempt to combine or relate the sentiment to network usage. To the best of our knowledge, our work is the first one that seeks to define user profiles based on the relationship between user sentiments derived from their messages and network usage.

Our work differs from the works [Zheng *et al.*, 2016; Chen *et al.*, 2015; Kabir *et al.*, 2018] in some aspects. Firstly, we investigate if there is a correlation of sentiment with the network usage of the wireless and mobile networks that enables the definition of new user profiles. Secondly, we adopt WhatsApp messages to collect the user's sentiments. In Instant Message (IM) applications like WhatsApp and Telegram, users better control who their interlocutors are. Messages are sent directly to one person or a group of contacts known by the user. This characteristic of IM communication allows users to be more spontaneous in their messages. This last claim is supported by the literature [Quan-Haase and Young, 2010; Karapanos *et al.*, 2016]. Thirdly, we apply a sentiment classifier based on Bidirectional Encoder Representations from Transformers (BERT) for the Portuguese language, which is based on the transfer learning concept. Fourthly, we create a novel dataset combining real information on the use of the wireless network and sentiment. Furthermore, we propose a new data analysis methodology based on statistical methods, frequent item-set pattern mining, and association rules to infer mobile user profiles based on the correlation between wireless usage data and user sentiments.

3 Concepts and Background

This section describes the main concepts adopted in this work. First of all, the sentiment analysis approach is discussed. Secondly, frequent item-set patterns mining and association rules from a dataset are also discussed.

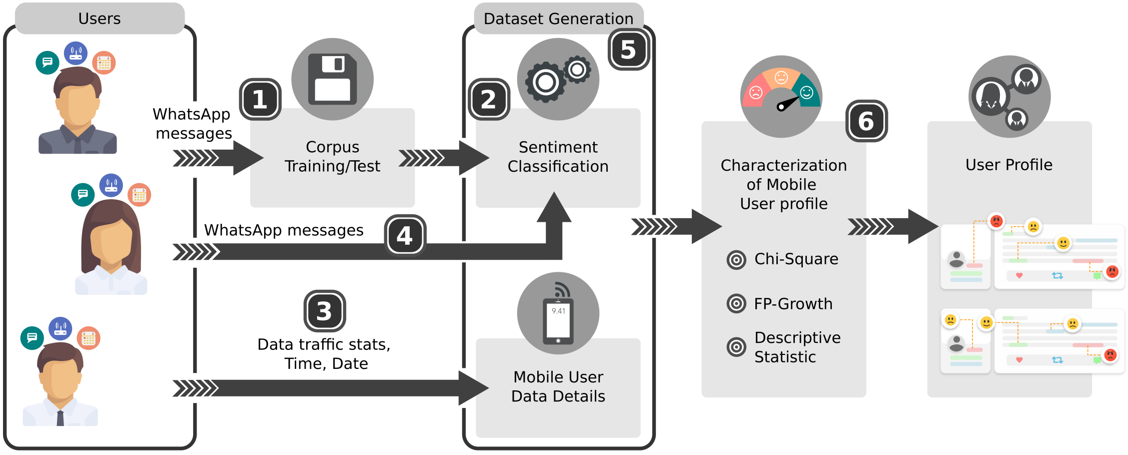


Figure 1. High-level architecture for characterizing the mobile user profiles

3.1 Sentiment Analyses Approach

Sentiment Analysis (SA) is one of the advances in the Natural Language Processing (NLP) task. It can be defined as analyzing and identifying the polarity/sentiment expressed in a piece of text. This text can be from different sources such as WhatsApp messages or social media posts [Liu, 2012]. Various methods and models [Díaz-Galiano *et al.*, 2019; Basile *et al.*, 2019] have helped in evolving the state-of-the-art in SA, such as classification systems based on lexicon [Junior *et al.*, 2021] and Recurrent Neural Networks (Convolutional Neural Networks and Long Short Term Memory (LSTM) with static word embeddings are the most common choice).

Recent advances in Deep Learning (DL) have led to breakthroughs in many NLP tasks, and sentiment analysis is one of them [Peters *et al.*, 2018; Devlin *et al.*, 2018; Clark *et al.*, 2020; dos Santos Neto *et al.*, 2020]. Bidirectional Encoder Representations from Transformers (BERT) is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right, as well as proper contexts in all layers [Devlin *et al.*, 2018]. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as Whatsapp sentiment analysis. This is the transfer learning concept, in which the fine-tuning phase involves copying the weights from a pre-trained network and tuning them on the downstream task. In the other words, the language model is learned from a large dataset (unlabeled), and this knowledge is transferred to be adjusted with a labeled dataset (much smaller than the first).

Transfer learning leverages data from distinct domains to train a model with better generalization properties. BERT adopts a fine-tuning approach that requires almost no specific architecture for the end task. This is a desirable property because an intelligent agent should minimize the use of prior human knowledge in the model design. Hence, BERT is one of the critical innovations in the recent development of contextualized representation learning.

3.2 Frequent Itemset Mining and Association Rules

Frequent pattern mining searches for recurring relationships in a given data set. This section introduces the basic concepts of frequent pattern mining to discover interesting associations and correlations between a set of items (itemset) in transactional datasets. We explain how we use these concepts to derive relations between user sentiment and the amount of traffic data in section 4.4.

Let $I = \{i_1, i_2, \cdots, i_{|I|}\}$ be a set of possible items and $\mathcal{T} = \{T_1, T_2, \cdots, T_{\mathcal{T}}\}$ be a multiset[1] of transactions were each transaction T in $\mathcal{T}$ is a non-empty subset of I. In the context of association rule mining, a set of items is usually referred to as *itemset* [Han *et al.*, 2012], and from now on, we use this term in this article. Thus, each T in $\mathcal{T}$ is an item set. Each transaction T is uniquely identified by a tid. A transaction T is said to contain an itemset X if $X \subseteq T$. The list of transactions that contain a given itemset X is called $tidlist$ of X and is denoted as $\mathcal{L}_{\mathcal{T}}(X)$ [Veloso *et al.*, 2002].

The *support* $s(X)$ of an itemset X is the ratio between the number of transactions in $\mathcal{T}$ that contain X and the number of transactions in $\mathcal{T}$, i.e.

$$s(X) = \frac{|\mathcal{L}_{\mathcal{T}}(X)|}{|\mathcal{T}|}. \tag{1}$$

An itemset X is a *frequent itemset* if $s(X)$ is greater or equal to a user-specified minimum support threshold ($minsup$) [**?**].

An *association rule* is an implication of the form $X \rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \emptyset$ [**?**]. X is referred to as the *antecedent* itemset of the association rule $X \rightarrow Y$, while Y is the *consequent* itemset. The support $s(X \rightarrow Y)$ of a rule $X \rightarrow Y$ is defined as:

$$s(X \rightarrow Y) = s(X \cup Y), \tag{2}$$

i.e., the number of transactions in $\mathcal{T}$ containing the itemset $X \cup Y$, which is equivalent to the joint probability of X and Y. The *confidence* $c(X \rightarrow Y)$ of a rule $X \rightarrow Y$ is defined as:

$$c(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)} \tag{3}$$

[1] Multiset is a modification of the concept of a set that, allows for multiple instances for each of its elements.

which corresponds to the conditional probability of Y given X.

An association rule r is said to be *strong* if $s(r) \geq minsup$ and $c(r) \geq minconf$, where $minsup$ is the minimum support threshold and $minconf$ is the minimum confidence threshold [Han *et al.*, 2012]. Equation 3 shows that the confidence of rule $X \rightarrow Y$ can be easily derived from the support of X and $X \cup Y$. Once the support counts of X, Y, and $X \cup Y$ are found, it is straightforward to derive the corresponding association rules $X \rightarrow Y$ and $Y \rightarrow X$ and check whether they are strong. Thus, the problem of mining strong association rules depends on that of finding frequent itemsets. For this reason, association rule mining, in general, involves two steps: 1 - mining frequent itemsets and 2 - mining association rules from the frequent itemsets. In this work, we used the Orange software[2] for mining both the frequent itemsets and the strong association rules on our dataset.

4 Characterization of the Mobile User Profile

This section describes the proposal and the assessment of the characterization of the mobile user profiles. To this end, wireless and mobile network usage and messages from WhatsApp were collected. Frequent Item-Set Pattern Mining and Association Rules techniques are used to recognize patterns in stored data, thus defining profiles of mobile users and seeking to verify whether there is a correlation between sentiments of mobile users and the amount of traffic they demand on the network. Figure 1 summarizes our methodology taken to characterize the profiles of mobile users.

Our methodology consists of the following steps:

1. To build a collection of WhatsApp messages collected from WhatsApp private groups in order to train a sentiment classifier (Sub-section 4.1);
2. To train the classifier in the collection of WhatsApp messages and to assess its effectiveness (Sub-section 4.2);
3. To collect the mobile users network usage (data traffic consumption, date, time, user id) (Sub-section 4.3);
4. To collect WhatsApp messages from mobile users (Sub-section 4.3) and to apply the classifier trained in step 2 to label the WhatsApp messages from the individual and private chats according to the sentiment they convey (Sub-section 4.4);
5. To generate a set of transactions by combining user sentiment derived from the classifier with normalized mobile user data (Sub-section 4.4);
6. To mine information to characterize the profiles of the mobile user using association rule (Sub-section 4.5).

These steps are described in detail in the following subsections.

4.1 Building a Training Corpus for Sentiment Classification

Our methodology's first step consists of obtaining a training corpus of WhatsApp messages. We chose messages from WhatsApp because it is the most popular global mobile messenger application of the world and one of the three most popular social networks worldwide [Clement, J., 2019]. It offers advantages in the analysis of sentiments as it is mainly employed to communicate with closer ties compared to more public platforms such as Facebook, Twitter, and Instagram [Karapanos *et al.*, 2016]. When people send a specific message to an individual or group of closest people more privately, they are looking to communicate something meaningful, genuine, tangible, and personal to the recipient, e.g., a sentiment. Moreover, WhatsApp is much more temporal since Facebook, Twitter, and Instagram are designed to foster an image or promote a thing of the person, their group, or their brand, whatever/whenever the person is posting. Hence, the sentiment classification on WhatsApp can pick up the current and most accurate sentiment of the mobile user. Therefore, this can benefit the characterization of the mobile user profiles based on sentiment attributes.

To derive a sentiment classifier for WhatsApp messages, we first needed to build a collection of messages from that mobile application to work as a training corpus. The construction of this collection was necessary since Sentiment Analysis in WhatsApp is not widespread in literature [Joshi, 2019; Rupavathy *et al.*, 2018] and, to the best of our knowledge, there is no *corpus* with WhatsApp sentences in Portuguese labeled with sentiment polarity that we could use as the training set.

Most of the existing investigations about sentiment analysis in WhatsApp show datasets that have more extreme sentiments rather than neutral, which suggests the polarized nature of the WhatsApp chat [Joshi, 2019; Resende *et al.*, 2019; Jain *et al.*, 2019; Dahiya *et al.*, 2020]. Therefore, this work takes only into consideration positive and negative sentiment classes.

We built our training corpus from private group chats with many people, thus allowing more sentence diversity. Approximately six thousand (6,000) messages from the WhatsApp groups were randomly selected to compose our training corpus. There are two types of volunteers in this work, namely volunteer mobile users and volunteer annotators. The former are volunteers who participate in the experiment about definition of mobile user profiles, the latter participate as corpus annotators. Besides, there is no overlapping between these types of volunteer sets. In the first moment, three volunteer annotators manually labeled the same messages as positive (4,450 messages) or negative (1,550 messages) of the training corpus. With the labeled data in hand, the labels with two-out-of-three annotators in agreement were kept (tiebreaker criteria). In addition, 500 different messages were used as the test corpus in the second moment. The same volunteer annotators of the training corpus manually labeled the test corpus (obtaining a total of 177 and 323 as negative and positive classes, respectively). Furthermore, the messages of the test corpus are distinct from those of the training corpus and were also written by distinct volunteers.

[2] https://orangedatamining.com/

Table 1. Examples of WhatsApp Messages

Messages	Class
Dá pra ver a irmã dele chorando tadinha. (You can see his sister crying.)	Negative
Palavra abençoada. (Blessed word.)	Positive
Tenho uma entrevista hj a tarde, mas comecei a estudar. (I have an interview today, but I started studying)	Positive
Pensei que você já tinha visto desculpas. (I thought you had already seen, sorry).	Negative

The resulting Kappa was 0.9140 [Krippendorff, 2011], which means the labeling has a high concordance between the annotators.

Table 1 lists some examples of labeled sentences.

4.2 Deep Neural Training/Test based on Transfer Learning

Most top-performing submissions to SemEval 2020 Tasks[3] [Hossain *et al.*, 2020; Sharma *et al.*, 2020; Patwa *et al.*, 2020] about humour and sentiment classification adopted a pre-trained language model such as BERT [Devlin *et al.*, 2018], RoBERTa [Liu *et al.*, 2019], XLNET [Yang *et al.*, 2020] and ALBERT [Lan *et al.*, 2020] models, and then fine-tune on the training set of the task. We adopted a similar approach to derive our sentiment classifier. However, since our training set and the messages to be classified are written in Portuguese, we used a BERT version pre-trained for the Portuguese language [Souza *et al.*, 2019]. This is the only transformer-based model pre-trained exclusively for the Portuguese language on a large dataset publicly released. It was pre-trained on the Brazilian Web as Corpus dataset (BrWaC) [Wagner Filho *et al.*, 2018], a dataset composed of 2.7 billion tokens, which were crawled and filtered from more than 60 million Brazilian Portuguese pages. We fine-tuned this version of BERT using the training corpus we obtained as described in Section 4.1.

In order to evaluate the effectiveness of our BERT-based classifier, we used the test corpus described in previous subsection. The sentiment classifier achieves high values in all effectiveness measures that are usually employed in deep learning of sentiment analysis, such as Overall Accuracy (87%), Precision (85%), Recall (86%), and F1-Score (86%). With more details for the negative class, we achieved a Precision of 80%, a Recall of 84%, and an F1-Score of 82%, while, for the positive class, a Precision of 91%, a Recall of 88%, and an F1-Score of 90%.

Figures 2(a) and 2(b) show that the most number of sentences have approximately 0 to 50 characters for the Training Corpus and Testing Corpus. Hence, the WhatsApp sentences are mostly short; they are more challenging than Twitter messages. For instance, the average sentence size was 23 characters for the Training Corpus, which is smaller than the average sentence length of tweets (i.e., 53 characters after changing characters to the limit [Boot *et al.*, 2019]).

4.3 Setup for Data Gathering of Mobile User Profiles

Figure 3 represents an overview of the cloud infrastructure scenario. This scenario was deployed to characterize user-profiles and correlate user sentiment with wireless network usage data. The cloud infrastructure was implemented through the Internet to collect network usage data. Each participant installed an application on his/her smartphone to connect to the created network and have the data collected. Thus, this scenario allows more mobility for the volunteers who participate in the research since it eliminates the need for a single fixed wireless access infrastructure. Therefore, this scenario allows data collection to happen anywhere and anytime, making data collection easier.

Each volunteer user installed the *Drony*[4] application on his/her smartphone. This application is a proxy client that performs Internet access from the volunteer's smartphone with an Internet outbound server located in the cloud, which was configured on the *Google Cloud Platform*[5]. The *Squid Proxy 10*[6] solution was implemented to collect access data. It is worth noting that this server does not store any information that identifies the volunteer in order to guarantee data privacy. All data that somehow identified the volunteer were discarded after processing the information, such as spreadsheets that linked the volunteers' emails with the identifier of a user. The server only generates internet access logs containing: *Date*, *Hour*, *User ID* (User about whom the information was collected), *Network Usage* in bytes (download+upload) for an one-hour period.

Table 2 shows the statistics of the collected data in the cloud infrastructure scenario. The numbers elucidate the total collecting time and the number of participants and quantify the collected information.

Table 2. Data collected in the Cloud Infrastructure Scenario.

Description	Quantity
Participants	27
Collecting time	3 months
Number of WhatsApp collected sentences	162314

It is worth mentioning that WhatsApp messages are not stored on servers as well as there is end-to-end encryption when these messages are transmitted/received so that WhatsApp and third parties cannot read them anyway. Therefore, we collected the WhatsApp messages directly from people that voluntarily agreed to participate in the research.

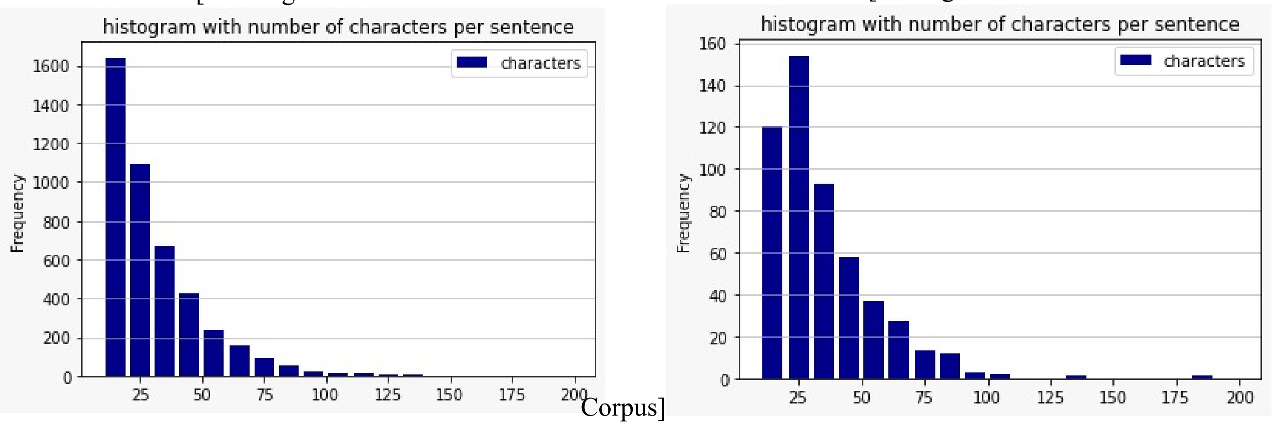
This work is supported by a research project submitted, evaluated, and approved by the Federal University of Goiás ethics committee to allow volunteer involvement. The volunteer participants in this research authorized the use of the WhatsApp messages by signing the consent term, which defined the rights and duties of volunteers and researchers.

[3] `https://alt.qcri.org/semeval2020/`

[4] https://play.google.com/store/apps/details?id=org.sandroproxy.drony

[5] https://cloud.google.com/

[6] http://www.squid-cache.org/



[Training Corpus]

[Testing Corpus]

Figure 2. Histogram: Frequency × Size of sentences (characters).

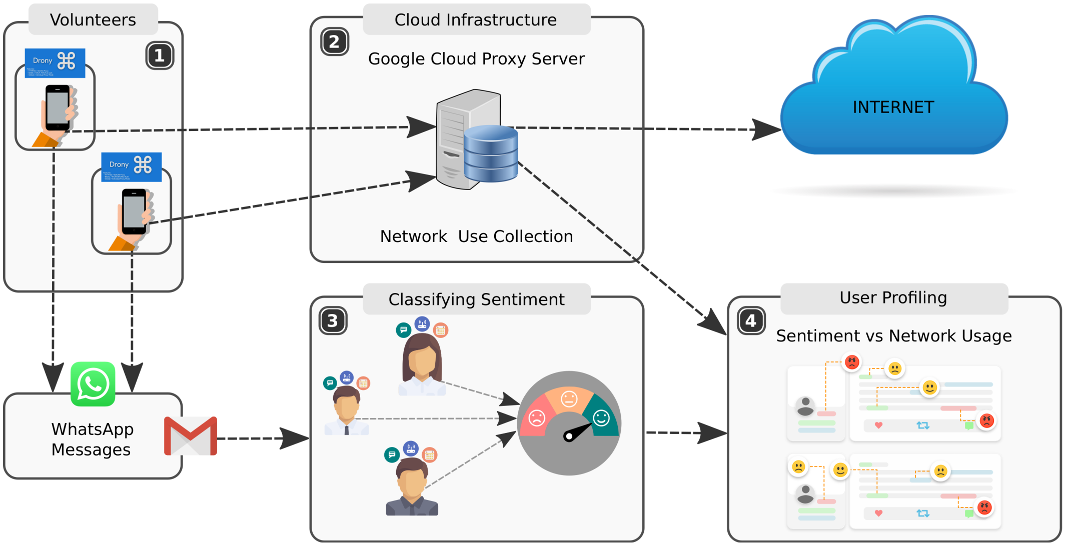


Figure 3. Overview of the Cloud Infrastructure Scenario.

Nevertheless, this term establishes that the WhatsApp messages must not be, under no circumstances, publicly released to preserve the volunteer's privacy. In a perfect scenario for data collection, it would be desirable that the volunteers did not know that they were being monitored. However, using material collected from social media and mobile messenger chats entails a number of legal obligations and complex questions of research ethics. Among the researchers' obligations are informing the research objective (defining user profiling) to the participants and collecting the user's acceptance. However, we did not inform volunteers that we would use their WhatsApp messages for sentiment analysis. Therefore, we believe it can not influence the message sentiment.

WhatsApp messages from individual and closed chats and network usage on the wireless network were collected from 27 participants/volunteers. It is worth mentioning that the majority of participants are young university students. It is important to stress that no information identifying the user is stored to mitigate data privacy issues. All data that somehow identified the participating users were discarded after processing.

4.4 Dataset Generation

A two-phase process was conducted to derive the dataset, which was used to analyze the correlation between users' sentiment and their network consumption. In the first phase, we derive an initial dataset. The initial dataset consists of two different auxiliary datasets, which were generated to characterize the user profiles. The first one contains the sentiment from the WhatsApp corpus described previously. The second one contains the network usage collected from the mobile user devices. In the second phase, we conduct a normalization process over the initial dataset, generating the final dataset effectively used in the correlation analysis.

Phase-one: data harvesting and integration

We set up the cloud infrastructure (Figure 3) and collected data about the network usage of all volunteer users for the entire collection period. We also collected the individual and private WhatsApp chats sent by the volunteers (to a specific email address) during this period since this kind of WhatsApp chat increases the chances of people communicating a genuine and tangible sentiment [Karapanos *et al*., 2016]. We joined the data from two auxiliary datasets (network usage and Whatsapp message collection) as follows:

- The data traffic consumption (network usage) is collected (step 1) and stored (step 2), summing up data download and upload (Megabytes) in both Wi-Fi and LTE network interfaces for a one-hour period.
- Each message was submitted to the BERT model sentiment analysis described previously, and the model's output label (positive or negative) is assigned to the message (step 3).
- A global sentiment label is assigned to the user for each hour by choosing the most frequent sentiment label of messages sent by her/him in the one-hour period. For instance, suppose that on the day 2019/10/10, in the one hour starting at 12:00, User15 sent 20 messages. If the majority of the messages (15 messages) were labeled as *negative* by the sentiment classifier based on transfer learning in this period of time, then a transaction is created in the initial dataset for User15 with a *negative* label in the one hour as shown in the first line of Table 3.
- A single user id is employed for every user, such as User15 or User8, to collect and to store WhatsApp messages in corpus and network usage in the traffic consumption dataset for that specific user. The data traffic consumption and the global sentiment label of a volunteer were gathered in the same transaction combining the records with the same value of user-id, hour, and day fields from both auxiliary datasets (step 4). Hence, the association between WhatsApp messages and users' traffic consumption for the initial dataset generation is based on the user id, date, and hour fields.

Table 3. Concatenated file with sentiment and mobile user network usage in the initial dataset.

Date	Hour	User ID	Network usage	Sentiment
2019/10/10	12 PM	User15	9873	Negative
2019/10/10	12 PM	User8	348654	Positive

Phase-two: data normalization

The initial dataset obtained at the end of phase one is used as input in this phase. This normalization phase was adopted because it can improve the quality of the data and, consequently, the results of the Frequent Item-Set Pattern Mining and Association Rules. Normalization aims to adjust values measured on different scales to a notionally common scale, thus decreasing the granularity of the data to increase the efficiency of the Association Rules algorithms [Malik *et al.*, 2010]. The normalization of each transaction of the initial dataset was conducted as follows:

1. The date field was converted to a weekday/weekend: Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday. For example, the date 2019/6/2 has been converted to Wednesday.
2. The field corresponding to the one hour was converted to a period of the day[7]: Morning, Afternoon, and Night, as follows:
    - **Morning** - Greater than or equal to 5 AM and less than 12 AM;
    - **Afternoon** - Greater than or equal to 12 AM and less than or equal to 6 PM;
    - **Night** - Greater than 6 PM and less than 12 PM.
3. The data traffic consumption attribute was normalized using the following steps. First, all records were ordered, considering the ascending order value of the attribute [Hautamaki *et al.*, 2004]. Next, the records were divided into two equal parts, where the record that separated these two parts contained a value equal to 558 Megabytes (Mb) so that the value that was greater than or equal to 558 Mb would be normalized to high value and smaller normalized to a low value.

Table 4 shows some transactions of the initial dataset and how their fields were normalized and transformed in the fields of the corresponding transactions of the final dataset. As shown in Table 4, each transaction output by the normalization phase is a tuple from $W \times P \times U \times C \times S$, where W corresponds to the set of names for the days of the week, $P = \{Morning, Afternoon, Night\}$, U is set of volunteers in our experiment, $C = \{low, high\}$ and $S = \{positive, \; negative\}$. Thus, in our case, the set of items $I = W \cup P \cup U \cup C \cup S$.

Table 4. Example of output of the normalization phase (Final Dataset).

Day of Week	Period	User ID	Network usage	Sentiment
Thursday	Afternoon	User15	Low	Negative
Thursday	Afternoon	User8	Low	Positive

4.5 Dataset Analyses

The statistic is a practical, efficient, and straightforward approach for user profiling [Zhao *et al.*, 2019]. Besides, as there is no knowledge about the relationship between sentiment and network usage attributes, we employ descriptive and non-parametric statistics to characterize the user profiles as well as an algorithm for frequent itemset mining based on Association Rules, which is also based on statistics of analyzed attributes.

Overall Dataset Analysis

First of all, we used the Pearson Chi-square test for statistical independence to verify if user sentiment and the intensity of network usage are two independent events. The Pearson Chi-square test for independence determines whether there is a statistical significant difference (i.e., a difference that is probably not just due to chance) between the expected frequencies and the observed frequencies in one or more categories [Zhai *et al.*, 2018]. The null hypothesis for this test is that the occurrence of two events is statistically independent. Thus, if the test fails to demonstrate the null hypothesis, we can conclude that both events are related to each other, although the structure of this relation is not revealed by the test.

Table 5 some statistics for the two attributes of interest (sentiment and network usage) in our dataset. The left most

[7] We discarded the messages and information about network usage in the period between 12 PM and 5 PM due to the shallow user activity in this period.

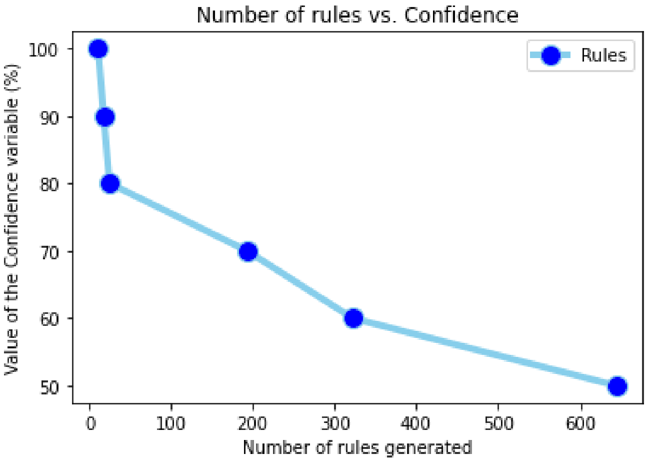


Figure 4. Comparative number of rules x confidence values.

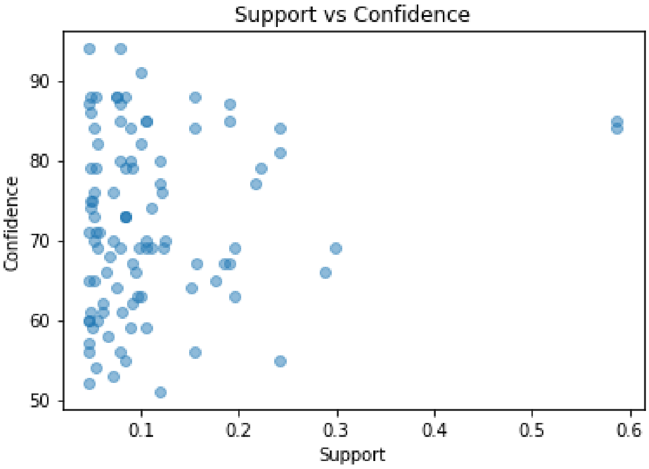


Figure 5. Comparative of confidence x support.

value in each cell is the actual count. The values in parentheses show the expected counts. The values in square brackets show the Chi-square value for each cell. The Total row shows the column totals for the observed values, and the Total column shows the row totals for the observed values.

Table 5. Chi-square test for sentiment and network usage attributes.

Data/Sentiments	**Positive**	**Negative**	**Total**
High	1191 (981.61) [44.67]	205 (414.39) [105.81]	1396
Low	235 (444.39) [98.66]	397 (187.61) [233.71]	632
Total	1426	602	2028

Equation 4 presents the formula for calculating the Chi-squared value for the dataset, which is the sum of the Chi-square values in the cells. The Chi-square value in each cell is the square of the difference between the observed frequency (f_o) and the expected frequency (f_e) divided by f_e.

$$\chi^2 = \sum_{k=1}^{n} \frac{(f_o - f_e)^2}{f_e} \tag{4}$$

We obtained the *critical value* from the Chi-square distribution based on the degrees of freedom and on the significance level. The critical value is the one expected if the two variables are independent. In our case the degree of freedom equals to one ($(r-1)(c-1)$), where r and c are, the number of rows and the number of columns in Table 5, respectively. The significance level α is 0.05. Thus, the critical value in the Chi-square distribution is 3.841. The Chi-square value χ^2 for our dataset is 482.8539, which is much greater than the critical value (3.841). Hence, there is a statistical significant difference between the frequencies, discarding the hypothesis that the sentiment and the network usage are two independent events. Thus, there is a relation between sentiment and network usage.

In other to further investigate the strength of the relation between sentiment and network usage in our dataset, we applied The FP-Growth algorithm on the the entire normalized dataset. This dataset is the final file in the format exemplified in Table 4. We used the the FP-Growth algorithm implemented in the Orange framework[8].

Figure 4 shows the graph comparing the number of rules with values of confidence. The entered values in the variable confidence are 50%, 60%, 70%, 80%, 90% and 100%, generating 652, 341, 179, 22, 16 and 3 rules, respectively. It is observed that from 80% there is a drop in the number of rules, below 30.

Figure 5 presents the comparative graph showing all generated rules with the confidence of 50% or more and support of 0.05 or more. It is observed that the highest density of rules is obtained for support values between 0.05 and 0.1. The rules with confidence above 80% and support close to 0.6 present the sentiment positive as antecedent and high network usage as consequent.

Table 6 shows the rules with the most significant values of confidence and support. The rule *positive* → *high* has support equal to 0.59, meaning that sentiment positive cooccurs with high network usage in almost 60% of the records of the dataset. The confidence value is 83% which means that in the records where positive sentiment occurs, 83% of them also have high network usage. On the other hand, the rule *negative* → *low* has support equal to 0.19, which means that negative sentiment and low network usage cooccur in almost 20% of the dataset records. Also, in 66% of the records where the negative sentiment occurs, the low network usage also occurs. These values show a strong relation between positive sentiment and high network usage and a non-negligible relation between negative sentiment and low network usage.

Table 6. Rules with greatest Support and Confidence

Antecedent	**Consequent**	**Support**	**Confidence**
Sentiment=Positive	Network=High	0.59	83%
Sentiment=Negative	Network=Low	0.19	66%

4.5.1 *User Profile Analysis

In other to determine the user profiles considering both sentiment detected in WhatsApp messages and network usage, we applied the FP-Growth algorithm to the dataset records of each individual user. We also computed for each user the percentage of each attribute (sentiment type and intensity of

[8] https://orange.biolab.si

network usage). We only consider association rules with support greater or equal to 0.05 and confidence equal or greater than 50%.

The generated association rules allowed to identify three user groups. The first group is shown in Table 7 and contains 55.6% of the total users. The percentage columns in the Table show high values of both positive sentiment and high network usage. Also, there is a predominance of rules with high confidence and high support that involve the attributes of positive sentiment and high network usage, which shows a strong relation between positive sentiment and high network usage. Another aspect that characterizes users in this group is that in spite of the percentages of both negative sentiment and low network activity being low, they are similar to each other for most users and the confidence of rules formed by these two components are high (always above or equal 60%) which shows that there is also a strong relation of the negative sentiment and low network usage in this group, whenever they occur. We also computed the Chi-square test for statistical independence on the first group, obtaining a value of 403.31 (See Table 10 in the Appendix) which is much greater than the critical value (3.841).

The second group of users is shown in Table 8 and is composed of 25.9% of the users. As occurs in the first group, users in this group are characterized by the rule *positive → high* with both high support and high confidence values for most of users. In fact, the rules of type *positive → high* in the second group present confidence and support values superior to those of the same type on the first group of users. However, contrary to the first group, both the relations between negative sentiment and any aspect of network usage and the relations between low network usage and any sentiment type are inconclusive. They do not present high support nor high confidence. Also, the number of occurrence of both negative sentiment and low network usage are smaller among users of the second group. Thus, we can say that users in the second group present strong relation between positive sentiment and high network usage. The Chi-square value for this group (see Table 11 in the Appendix) is 37.2375 which is still high, but smaller than that in the first group.

The third group of users is shown in Table 8 and it is composed of 18.5% of the users in the dataset. User sentiment is still correlated with network usage in the third group, since the Chi-square value for this group is 6.671, and consequently superior to the critical value. However, the Chi-square value is much smaller then in the two first groups. Also, the proportion of records with negative sentiment in the third group is similar to those in the first group but, contrary to the first group the negative sentiment is associated to high network usage. In fact, the proportion of records with low network usage is very small for most users in the third group. Thus, users in the third group tend to be characterized by a high network usage and this usage is much less related to sentiment than in the other first two groups. Thus, we can conclude that despite sentiment and network usage being correlated in the third group, the strength of this relation is the weakest when compared to the other two groups.

he analysis of the three mobile user groups showed a strong relation between user sentiment and network usage in the first two groups which correspond to 81.5% of the users. Thus, we can conclude that sentiment captured from WhatsApp messages is an important feature to characterize network usage. In our case, three distinct user profiles were identified, each corresponding to one of the three group discussed above.

Many aspects of the next generation of wireless and mobile networks can be optimized with these mobile user profiles inferred in this work, such as spectrum demand, antenna placement, and cache pre-fetching by forecasting traffic demand according to the sentiment of most users [Wang *et al.*, 2014; Wei *et al.*, 2014; Cho *et al.*, 2014; Srinivasan and Murthy, 2019]. Therefore, network softwarization technologies can better allocate and manage the network resources through a decision support system aware of the user profiles, making available only resources that will be used. This will free surplus resources, opening the possibility of eliminating the existence of idle resources and providing better planning of how and where these resources will be consumed.

MNOs can be aware of the evolution of their users' behavior that will most be related to the resource demand, which helps allocate its investments and deployments more dynamically, improving both CAPEX (CAPital EXpenditures) and OPEX (OPerating EXpenditures). Furthermore, this enables a better distribution of resources to where they are used the most and thus prioritizing specific users. Hence, based on this knowledge, stochastic optimization models for the network can be proposed, which is based on the degree of certainty (confidence and support values of frequent item-set mining patterns). Therefore, the networks can be optimized as a whole, providing a better quality of service to its users once they understand their needs more precisely.

The dynamic spectrum allocation needs a short period to make decisions. The spectrum allocation service can make a scheduler based on statistical information (confidence) for short or long periods. Thus, if the spectrum scheduler made decisions taking into account mobile users' profiles defined in this research work as being sentiment and consumed traffic, the network could offer an optimized and personalized service to its users. Based on the stochastic optimization that takes advantage of the knowledge of profiles, the scheduler can predict who demands high/low traffic during day periods (Morning, Afternoon, and Night), and it would be possible to distribute the spectrum band to a group of users in shorter periods (1 hour or period of the day) [Hayashida *et al.*, 2019; Srinivasan and Murthy, 2019]. The proposed time duration for defining/updating the user profiles in this work is one hour. This process considers the collection of texts, collection of data on network usage, classification of user sentiment, and definition of profiles.

5 Conclusions and Future works

One of the most significant challenges of next-generation wireless and mobile networks is maintaining quality and user satisfaction while providing personalized services, considering the growth in the number of devices and applications with different performance requirements. In this context, the present work inferred knowledge about human behavior that is related to the use of the network, defining mobile user pro-

Table 7. Rules per User in the First Group of Mobile Users.

User	Antec.	Conse.	Sup.	Conf.	Pos.(%)	Neg.(%)	High.(%)	Low.(%)
user1	Positive Negative	High Low	0.51 0.20	0.79 0.60	67	33	64	36
user2	Positive Negative	High Low	0.39 0.31	0.84 0.81	61	39	47	53
user8	Positive Low	High Negative	0.68 0.20	0.97 0.92	69	31	79	21
user10	Positive Negative	High Low	0.50 0.23	0.81 0.66	66	34	63	37
user11	Positive Low	High Negative	0.68 0.16	0.93 0.77	73	27	79	21
user15	Positive Negative	High Low	0.61 0.24	0.95 0.88	73	27	65	35
user21	Positive Low	High Negative	0.56 0.16	0.89 0.71	65	35	77	23
user32	Positive Negative	High Low	0.50 0.25	0.80 0.66	67	33	67	33
user34	Positive Low	High Negative	0.40 0.30	0.72 0.66	52	48	57	43
user36	Positive Low	High Negative	0.82 0.07	1.00 1.00	82	18	91	9
user37	Positive Low	High Negative	0.55 0.25	0.88 0.76	63	37	68	32
user46	Positive Low	High Negative	0.68 0.15	0.91 0.71	76	24	79	21
user50	Negative High	Low Positive	0.36 0.37	0.92 0.92	61	39	41	59
user53	Positive Negative	High Low	0.47 0.29	0.95 0.92	69	31	51	49
user54	Positive Negative	High Low	0.62 0.20	0.92 0.78	75	25	69	31

Table 8. Rules per User in the Second Group of Mobile Users.

User	Antec.	Conse.	Sup.	Conf.	Pos.(%)	Neg.(%)	High.(%)	Low.(%)
user3	Positive	High	0.88	0.97	91	9	95	5
user5	Positive	High	1.00	1.00	100	0	100	0
user14	Positive	High	1.00	1.00	100	0	100	0
user19	Positive	High	0.78	0.94	84	16	85	15
user29	Positive	High	0.79	0.98	82	18	97	3
user33	Positive	High	0.63	0.89	71	29	84	16
user39	Positive	High	0.75	0.90	85	15	85	15

Table 9. Rules per User in the Third Group of Mobile Users.

User	Antec.	Conse.	Sup.	Conf.	Pos.(%)	Neg.(%)	High.(%)	Low.(%)
user12	Positive	High	0.33	1.00	50	50	100	0
	Negative	High	0.66	1.00				
user13	Positive	High	0.57	0.88	60	40	73	27
	Negative	High	0.21	0.60				
	Low	Negative	0.14	0.67				
user16	Positive	High	0.84	1.00	85	15	96	4
	Negative	High	0.11	0.75				
	***Negative	***Low	0.03	1.00				
user52	Positive	High	0.81	0.93	87	13	95	5
	Negative	High	0.13	1.00				
	Low	Positive	0.05	1.00				
user55	Positive	High	0.80	1.00	83	17	100	0
	Negative	High	0.20	1.00				

files. The softwarization technologies and stochastic/linear optimization models can apply this knowledge to manage better the services and resources of next-generation mobile and wireless network services, such as spectrum allocation and caching.

Related works that consider human behavior sentiment analysis as an attribute for optimizing wireless and mobile networks are scarce. These few studies that evaluate the sentiment analysis in the user profile are very superficial. The main research question was to characterize user profiles, correlating the user's sentiment with network usage data: can the user's sentiment be related to network data traffic? A data analysis methodology based on statistic methods and the FP-Growth pattern mining algorithm was developed to validate the hypothesis mentioned above for user profiling.

The results of the chi-square test and FP-Growth for frequent mining items and association rules for most users show that the dependence between sentiment and network usage is very high for more than 81.5% of the users. Furthermore, three mobile user profiles were inferred. The first group is formed by users whose positive sentiment is strong related to high network usage, and negative sentiment is strong related to low network usage. The second group is characterized by users for whom the positive sentiment is strongly related to high network usage, but the association rules were inconclusive about the relation between the negative sentiment and low network usage. The third group is composed by user who use the network intensively but for whom sentiment is only loosely related to network usage. However, this group corresponds to a minority (18.5%) of the users.

When analyzing the obtained results, it can be considered that the research hypothesis is confirmed. The user's sentiments is related to the demand for network traffic. However, it is important to highlight that the statistical tools we used are not sufficient to determine any causal relation between sentiment and network usage.

As future work, we intend to develop stochastic optimization models for decision support systems of dynamic spectrum allocation as well as for caching recommender systems that will take into account the inferred user profiles in this work (sentiment x network usage).

Appendix

Table 10. Chi-square test for sentiment and network usage attributes for the first group of users. The sum of Chis-square values equals to 403.31 which is superior to the critical value (3.841).

Data/Sentiments	Positive	Negative	Total
High	895 (711.76) [47.17]	157 (340.24) [98.69]	1052
Low	220 (403.24) [83.27]	376 (192.76) [174.19]	596
Total	1115	533	1648

Table 11. Chi-square test for sentiment and network usage attributes for the second group of users. The sum of Chis-square values equals to 37.2375 which is superior to the critical value (3.841).

Data/Sentiments	Positive	Negative	Total
High	226 (214.12) [0.66]	34 (45.88) [3.08]	260
Low	12 (23.88) [5.91]	17 (5.12) [27.59]	29
Total	238	51	289

Table 12. Chi-square test for sentiment and network usage attributes for the third group of users. The sum of Chis-square values equals to 6.6715 is superior to the critical value (3.841).

Data/Sentiments	Positive	Negative	Total
High	70 (67.38) [0.10]	14 (16.62) [0.41]	84
Low	3 (5.62) [1.22]	4 (1.38) [4.94]	7
Total	73	18	91

Acknowledgment

This article presented a research effort carried out in the ambit of the international research project between Brazil and France, entitled EMBRACE (EMBRACE - lEveraging huMan Behavior Resource AlloCations ModEls - https://team.inria.fr/embrace/). This work is partly funded by INRIA, FAPESP (Grant #2018/23011-1), FAPEMIG, CAPES (Finance Code 001), FAPEG and Fundacao Auracaria.

6 Declarations

6.1 Availability of data and materials

This work is supported by a research project submitted, evaluated, and approved by the Federal University of Goiás ethics committee to allow volunteer involvement. The volunteer participants in this research authorized the use of the WhatsApp messages by signing the consent term, which defined the rights and duties of volunteers and researchers. Nevertheless, this term establishes that the WhatsApp messages must not be publicly released in order to preserve the volunteer's privacy.

6.2 Acknowledgments

This work has been partially supported by the CAPES (Brazil).

6.3 Competing interests

The authors declare that they have no competing interests

6.4 Funding

This work is partly funded by INRIA, FAPESP (Grant #2018/23011-1), FAPEMIG, CAPES (Finance Code 001), FAPEG and Fundacao Auracaria.

6.5 Authors' contributions

Leonardo developed the work (implementation and testing of the computer code) and investigation, carried out the formal analyses and validation, data curation, and wrote the original draft.

Roger assisted in the visualization, conceptualization, writing, reviewing, and editing. He assisted in future mobile networks.

Nádia was the co-advisor who helped investigate, work on methodology resources, and write the original draft. She assisted in sentiment analysis and text pattern mining.

Thierson helped in conceptualization, investigation, formal analyses, data curation validation, writing, reviewing, and editing. He assisted in statistic methods, sentiment analysis and text pattern mining subjects.

Vinicius was the advisor that worked in the investigation, Project management, conceptualization, methodology, writing the original draft, and supervision. He assisted in future mobile networks.

References

Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., and Flinck, H. (2018). Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys Tutorials*, 20(3):2429–2453. DOI: 10.1109/COMST.2018.2815638.

Akabane, A. T., Immich, R., Madeira, E. R. M., and Villas, L. A. (2018). imob: An intelligent urban mobility management system based on vehicular social networks. In *IEEE Vehicular Networking Conference (VNC)*, pages 1–8. DOI: 10.1109/VNC.2018.8628436.

Akabane, A. T., Immich, R., Pazzi, R. W., Madeira, E. R. M., and Villas, L. A. (2019). Exploiting vehicular social networks and dynamic clustering to enhance urban mobility management. *Sensors*, 19(16):3558. DOI: 10.3390/s19163558.

Awwad, A. M. A. (2021). Visual emotion-aware cloud localization user experience framework based on mobile location services. *International Journal of Interactive Mobile Technologies*, 15(14). DOI: 10.3991/ijim.v15i14.20061.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics. DOI: 10.18653/v1/S19-2007.

Boot, A. B., Sang, E. T. K., Dijkstra, K., and Zwaan, R. A. (2019). How character limit affects language usage in tweets. *Palgrave Communications*, 5(1):76. DOI: 10.1057/s41599-019-0280-3.

Bouras, C., Kollia, A., and Papazois, A. (2017). Sdn nfv in 5g: Advancements and challenges. In *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, pages 107–111. DOI: 10.1109/ICIN.2017.7899398.

Chan, C. A., Yan, M., Gygax, A. F., Li, W., Li, L., Chih-Lin, I., Yan, J., and Leckie, C. (2019). Big data driven predictive caching at the wireless edge. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. DOI: 10.1109/ICCW.2019.8756663.

Chen, M., Zhang, Y., Li, Y., Mao, S., and Leung, V. C. M. (2015). Emc: Emotion-aware mobile cloud computing in 5g. *IEEE Network*, 29(2):32–38. DOI: 10.1109/MNET.2015.7064900.

Cho, H., Lai, C., Shih, T. K., and Chao, H. (2014). Integration of sdr and sdn for 5g. *IEEE Access*, 2:1196–1204. DOI: 10.1109/ACCESS.2014.2357435.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. DOI: 10.48550/arXiv.2003.10555.

Clement, J. (2019). Whatsapp - statistics and facts. Accessed 18 September 2019. Available at: `https://www.statista.com/topics/2018/whatsapp/`.

Dahiya, S., Mohta, A., and Jain, A. (2020). Text classification based behavioural analysis of whatsapp chats. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 717–724. DOI: 10.1109/ICCES48766.2020.9137911.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. DOI: 10.48550/arXiv.1810.04805.

Díaz-Galiano, M. C., Vega, M. G., Casasola, E., Chiruzzo, L., Cumbreras, M. Á. G., Cámara, E. M., Moctezuma, D., Montejo-Ráez, A., Cabezudo, M. A. S., Tellez, E. S., Graff, M., and Miranda-Jiménez, S. (2019). Overview of TASS 2019: One more further for the global spanish sentiment analysis corpus. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 550–560. CEUR-WS.org. Available at: `http://ceur-ws.org/Vol-2421/TASS_overview.pdf`.

dos Santos, R. P. and Lopes, G. R. (2019). Thematic series on social network analysis and mining. journal of internet services and applications. volume 10, pages 1–4. SpringerOpen. DOI: 10.1186/s13174-019-0113-z.

dos Santos Neto, M. V., da Silva Amaral, A. D., da Silva, N. F. F., and da Silva Soares, A. (2020). Deep learning brasil - NLP at semeval-2020 task 9: Sentiment analysis of code-mixed tweets using ensemble of language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1233–1238. International Committee for Computational Linguistics. DOI: 10.18653/v1/2020.semeval-1.164.

Eke, C. I., Norman, A. A., Shuib, L., and Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access*, 7:144907–144924. DOI: 10.1109/ACCESS.2019.2944243.

Erol-Kantarci, M. and Sukhmani, S. (2018). Caching and computing at the edge for mobile augmented reality and virtual reality (ar/vr) in 5g. In Zhou, Y. and Kunz, T., editors, *Ad Hoc Networks*, pages 169–177, Cham. Springer International Publishing. DOI: 10.1007/978-3-319-74439-1_1.

Furno, A., Naboulsi, D., Stanica, R., and Fiore, M. (2017). Mobile demand profiling for cellular cognitive networking. *IEEE Transactions on Mobile Computing*, 16(3):772–786. DOI: 10.1109/TMC.2016.2563429.

Giraldo-Rodríguez, C., Fontenla-González, J., Pérez-Garrido, C., Mhiri, S., Gil-Castiñeira, F., and González-Castaño, F. J. (2015). Tsa: Terminal-supported 5g network optimization. In *2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 209–216. DOI: 10.1109/WiMOB.2015.7347963.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition*. Morgan Kaufmann Publishers. Available at: `http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1`.

Hautamaki, V., Karkkainen, I., and Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 430–433 Vol.3. DOI: 10.1109/ICPR.2004.1334558.

Hayashida, T., Okumura, R., Mizutani, K., and Harada, H. (2019). Possibility of dynamic spectrum sharing system by vhf-band radio sensor and machine learning. In *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 1–6. DOI: 10.1109/DySPAN.2019.8935871.

Hossain, N., Krumm, J., Gamon, M., and Kautz, H. (2020). Semeval-2020 task 7: Assessing humor in edited news headlines. DOI: 10.48550/arXiv.2008.00304.

Jain, D., Garg, A., and Saraswat, M. (2019). Sentiment analysis using few short learning. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 102–107. DOI: 10.1109/ICIIP47207.2019.8985855.

Joshi, S. (2019). *Sentiment Analysis on WhatsApp Group Chat Using R*, pages 47–55. Springer Singapore, Singapore. DOI: 10.1007/978-981-13-6347-4_5.

Junior, A. B., F. da Silva, N. F., Rosa, T. C., and C. Junior, C. G. (2021). Sentiment analysis with genetic programming. *Information Sciences*. DOI: 10.1016/j.ins.2021.01.02.

Kabir, A., Iqbal, M., ul Abidin Jaffri, Z., Rathore, S. A., Kitindi, E. J., and Rehman, G. (2018). User aware edge caching in 5g wireless networks. In *IJCSNS International Journal of Computer Science and Network Security*, volume 18, pages 25–32. Accessed 10 January 2019. Available at: `http://paper.ijcsns.org/07_book/201801/20180104.pdf`.

Karapanos, E., Teixeira, P., and Gouveia, R. (2016). Need fulfillment and experiences on social media: A case on facebook and whatsapp. *Computers in Human Behavior*, 55:888–897. DOI: 10.1016/j.chb.2015.10.015.

Komninos, A., Frengkou, A.-E., and Garofalakis, J. (2021). Hush now! context factors behind smartphone ringer mode changes. *Pervasive and Mobile Computing*, page 101332. DOI: 10.1016/j.pmcj.2021.101332.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Available at:`https://repository.upenn.edu/asc_papers/43/`.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*. DOI: 10.48550/arXiv.1909.11942.

Leng, B., Liu, J., Pan, H., Zhou, S., and Niu, Z. (2015). Topic model based behaviour modeling and clustering analysis for wireless network users. In *2015 21st Asia-Pacific Conference on Communications (APCC)*, pages 410–415. DOI: 10.1109/APCC.2015.7412547.

Li, L., Erfani, S., Chan, C. A., and Leckie, C. (2019). Multi-scale trajectory clustering to identify corridors in mobile networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2253–2256, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3357384.3358157.

Li, S., Xu, L. D., and Zhao, S. (2018). 5g internet of things: A survey. *Journal of Industrial Information Integration*, 10:1–9. DOI: 10.1016/j.jii.2018.01.005.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers. DOI: 10.1007/978-3-031-02145-9.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. DOI: 10.48550/arXiv.1907.11692.

M. Borges, V. C., Cardoso, K. V., Cerqueira, E., Nogueira, M., and Santos, A. (2015). Aspirations, challenges, and open issues for software-based 5g networks in extremely dense and heterogeneous scenarios. *EURASIP Journal on Wireless Communications and Networking*, 2015(1):164. DOI: 10.1186/s13638-015-0380-8.

Malik, J. S., Goyal, P., and Sharma, M. (2010). A comprehensive approach towards data preprocessing techniques and association rules. Available at:`http://bvicam.in/INDIACom/news/INDIACom%202010%20Proceedings/papers/Group3/INDIACom10_279_Paper%20(2).pdf`.

Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. (2016). Large-scale mobile traffic analysis: A survey. *IEEE Communications Surveys and Tutorials*, 18(1):124–161. DOI: 10.1109/COMST.2015.2491361.

Naboulsi, D., Stanica, R., and Fiore, M. (2014). Classifying call profiles in large-scale mobile traffic datasets. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 1806–1814. DOI: 10.1109/INFOCOM.2014.6848119.

Neto, E. P., Silva, F. S. D., Schneider, L. M., Neto, A. V., and Immich, R. (2021). Seamless mano of multi-vendor sdn controllers across federated multi-domains. *Computer Networks*, 186:107752. DOI: 10.1016/j.comnet.2020.107752.

Nummenmaa, L., Hari, R., Hietanen, J. K., and Glerean, E. (2018). Maps of subjective feelings. *Proceedings of the National Academy of Sciences*, 115(37):9198–9203. DOI: 10.1073/pnas.1807390115.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. DOI: 10.18653/v1/2020.semeval-1.100.

Peltonen, E., Sharmila, P., Asare, K. O., Visuri, A., Lagerspetz, E., and Ferreira, D. (2020). When phones get personal: Predicting big five personality traits from application usage. *Pervasive and Mobile Computing*, 69:101269. DOI: 10.1016/j.pmcj.2020.101269.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. DOI: 10.18653/v1/N18-1202.

Qamar, Faizanand Hindia, M. H. D. N., Dimyati, K., Noordin, K. A., and Amiri, I. S. (2019). Interference management issues for the future 5g network: a review. *Telecommunication Systems*, 71(4):627–643. DOI: 10.1007/s11235-019-00578-4.

Qiao, Y., Xing, Z., Fadlullah, Z. M., Yang, J., and Kato, N. (2018). Characterizing flow, application, and user behavior in mobile networks: A framework for mobile big data. *IEEE Wireless Communications*, 25(1):40–49. DOI: 10.1109/MWC.2018.1700186.

Quan-Haase, A. and Young, A. L. (2010). Uses and gratifications of social media: A comparison of facebook and instant messaging. *Bulletin of Science, Technology & Society*, 30(5):350–361. DOI: 10.1177/0270467610380009.

Raissi, F., Yangui, S., and Camps, F. (2019). Autonomous cars, 5g mobile networks and smart cities: Beyond the hype. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 180–185. DOI: 10.1109/WETICE.2019.00046.

Resende, G., Melo, P., Reis, J. C. S., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019). Analyzing textual (mis)information shared in whatsapp groups. In *Proceedings of the ACM Conference on Web Science*, WebSci'19, pages 225–234. DOI: 10.1145/3292522.3326029.

Rupavathy, N., Belinda, M., Nivedhitha, G., and Abhinaya, P. (2018). Whatsapp sentiment analysis. *Journal of Computational and Theoretical Nanoscience*, 15(11-12):3462–3465. DOI: 10.1166/jctn.2018.7645.

Sakouhi, T. and Akaichi, J. (2021). Dynamic and multi-source semantic annotation of raw mobility data using geographic and social media data. *Pervasive and Mobile Computing*, 71:101310. DOI: 10.1016/j.pmcj.2020.101310.

Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., Pulabaigari, V., and Gamback, B. (2020). Semeval-2020 task 8: Memotion analysis – the visuo-lingual metaphor!. DOI: 10.48550/arXiv.2008.03781.

Singh, A. and Sharma, A. (2019). A multi-agent framework for context-aware dynamic user profiling for web personalization. In *Software Engineering*, pages 1–16, Singapore. Springer Singapore. DOI: 10.1007/978-981-10-8848-3_1.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*. DOI: 10.48550/arXiv.1909.1064.

Srinivasan, M. and Murthy, S. R. (2019). Efficient spectrum slicing in 5g networks: An overlapping coalition formation approach. *IEEE Transactions on Mobile Computing*, pages 1–1. DOI: 10.1109/TMC.2019.2908903.

Taleb, T., Mada, B., Corici, M., Nakao, A., and Flinck, H. (2017). Permit: Network slicing for personalized 5g mobile telecommunications. *IEEE Communications Magazine*, 55(5):88–93. DOI: 10.1109/MCOM.2017.1600947.

Tandon, G. and Chan, P. K. (2009). Tracking user mobility to detect suspicious behavior. In *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*, pages 871–882. DOI: 10.1137/1.9781611972795.75.

Thuemmler, C., Paulin, A., Jell, T., and Lim, A. K. (2018). Information technology – next generation: The impact of 5g on the evolution of health and care services. In Latifi, S., editor, *Information Technology - New Generations*, pages 811–817, Cham. Springer International Publishing. DOI: 10.1007/978-3-319-54978-$1_1$00.

Ullah, I. and Binbusayyis, A. (2022). Joint optimization of privacy and cost of in-app mobile user profiling and targeted ads. *IEEE Access*, 10:38664–38683. DOI:

10.1109/ACCESS.2022.3166152.

Vamvakas, P., Tsiropoulou, E. E., and Papavassiliou, S. (2019). Dynamic spectrum management in 5g wireless networks: A real-life modeling approach. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 2134–2142. DOI: 10.1109/INFOCOM.2019.8737443.

Veloso, A., Meira Jr, W., Carvalho, M. d., Pôssas, B., Parthasarathy, S., and Zaki, M. J. (2002). Mining frequent itemsets in evolving databases. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 494–510. SIAM. DOI: 10.1137/1.9781611972726.29.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4339–4344. Available at:`https://aclanthology.org/L18-1686`.

Wang, C., Haider, F., Gao, X., You, X., Yang, Y., Yuan, D., Aggoune, H. M., Haas, H., Fletcher, S., and Hepsaydir, E. (2014). Cellular architecture and key technologies for 5g wireless communication networks. *IEEE Communications Magazine*, 52(2):122–130. DOI: 10.1109/MCOM.2014.6736752.

Wang, Y., Li, P., Jiao, L., Su, Z., Cheng, N., Shen, X. S., and Zhang, P. (2017). A data-driven architecture for personalized qoe management in 5g wireless networks. *IEEE Wireless Communications*, 24(1):102–110. DOI: 10.1109/MWC.2016.1500184WC.

Wei, L., Hu, R. Q., Qian, Y., and Wu, G. (2014). Key elements to enable millimeter wave communications for 5g wireless systems. *IEEE Wireless Communications*, 21(6):136–143. DOI: 10.1109/MWC.2014.7000981.

Wu, C., Chen, X., Zhu, W., and Zhang, Y. (2017). Socially-driven learning-based prefetching in mobile online social networks. *IEEE/ACM Transactions on Networking*, 25(4):2320–2333. DOI: 10.1109/TNET.2017.2681121.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2020). Xlnet: Generalized autoregressive pretraining for language understanding. DOI: 10.48550/arXiv.1906.08237.

Yousaf, F. Z., Bredel, M., Schaller, S., and Schneider, F. (2017). Nfv and sdn—key technology enablers for 5g networks. *IEEE Journal on Selected Areas in Communications*, 35(11):2468–2478. DOI: 10.1109/JSAC.2017.2760418.

Zhai, Y., Song, W., Liu, X., Liu, L., and Zhao, X. (2018). A chi-square statistics based feature selection method in text classification. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 160–163. DOI: 10.1109/ICSESS.2018.8663882.

Zhao, S., Li, S., Ramos, J., Luo, Z., Jiang, Z., Dey, A. K., and Pan, G. (2019). User profiling from their use of smartphone applications: A survey. *Pervasive and Mobile Computing*, 59:101052. DOI: 10.1016/j.pmcj.2019.101052.

Zhao, S., Zhao, Y., Zhao, Z., Luo, Z., Huang, R., Li, S., and Pan, G. (2017). Characterizing a user from large-scale smartphone-sensed data. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, pages 482–487, New York, NY, USA. ACM. DOI: 10.1145/3123024.3124437.

Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., and Xiang, W. (2016). Big data-driven optimization for mobile networks toward 5g. *IEEE Network*, 30(1):44–51. DOI: 10.1109/MNET.2016.7389830.