How well do Spaniards sleep? Analysis of Sleep Disorders based on Twitter mining

Daniel Suárez*, Oscar Araque† and Carlos A. Iglesias†
Intelligent Systems Group
Universidad Politécnica de Madrid
Madrid, Spain
Avenida Complutense, 30
Email: *daniel.suarez.souto@alumnos.upm.es, †{o.araque,carlosangel.iglesias}@upm.es,

***Abstract*—Twitter is a social network that allows its users to exchange messages of 280 characters with the possibility of accompanying them with a photo, video and/or link. This social network has been used as a source of data for numerous research studies on the human being. This study aims to analyse and characterize the messages coming from Spanish-speaking users related to the most common sleep disorder in our society, insomnia. For this purpose, this study provides two machine learning classifiers that enable the classification of users with insomnia together with the self-reported cause. In this context, this paper proposes a novel feature extraction method that exploits the similarity measure that can be computed in word embeddings models. For training these classifiers, a dataset of tweets in Spanish containing the word insomnia has been manually annotated to draw conclusions about the geographical distribution, symptoms and the different topics that users with insomnia treat. In addition, a second dataset has been collected formed by two groups of users from Spain with insomnia and without insomnia. Analysing the timeline of both groups we have been able to extract the differences in the patterns of activity on Twitter of each of these groups.**

***Index Terms*—insomnia, sleep disorder, Twitter, social mining, natural language processing, machine learning, feature extraction**

I. Introduction

The catalogue of different sleep disorders is part of the main problems that must be faced by medicine today. The percentage of people suffering from some of these disorders is 31% in Western Europe, 56% in the USA and 23% in Japan [1].

A study by DOPPS (Dialysis Outcomes and Practice Patterns) concluded that there are a number of common clinical features associated with people who sleep a few hours of sleep such as increased body mass index, pain, coronary artery disease, congestive heart failure, diabetes, lung disease, psychiatric disorders, peripheral arterial disease, depression, and pruritus [2].

However, it is estimated that only some of these people follow some kind of medical treatment for these disorders. This fact makes necessary to create techniques for medical professionals to predict a sleep disorders in an early way and to be able to advise healthy sleep routines and therapies.

This article aims to analyse sleep disorders based on Twitter mining. This social network allows its users to publish messages of a maximum of 280 characters with the possibility of accompanying them with a photo, video, and/or link.

In our case, we are interested in determining if it is possible to characterize potential people to suffer insomnia through the seĺf-reported messages posted on a social network. Specifically, the research will be based on Spanish-speaking messages and users in the social network Twitter.

We are also interested in determining the proportion of users who suffer from each of the different symptoms of insomnia, allowing us to carry out a more extensive study of the phenomenon of insomnia. According to the third edition of the International Criteria for Sleep Disorders (ICSD-3) [3], the main symptoms are: difficulty beginning sleep, short sleep duration and poor energy during the day. Also, another objective consists in measuring the proportion of users that undergo medical treatments of insomnia.

To achieve these goals we will develop two classifiers through machine learning techniques responsible for automating the tasks of characterizing users and tweets on the topic of insomnia. Also, with the aim of fostering research in this field, we make public the generated datasets [1].

The reminder of this paper is organized as follows. Sect. II reviews related work. Sect. III elaborates the methodology for collecting and annotating data. Sect. IV describes the techniques uses for tweet classification. Sect. V presents the analysis of the results of the data. Finally Sect. VI discusses the conclusions of this work.

ORCID info:
D. Suárez: https://orcid.org/0000-0002-6225-3324
O. Araque: https://orcid.org/0000-0003-3224-0001
C. A. Iglesias: https://orcid.org/0000-0002-1755-2712
Corresponding author: Oscar Araque. Tlf: +34 910672166.

[1] https://github.com/gsi-upm/dreampy-resources

978-1-5386-9588-3/18/$31.00 ©2018 IEEE

II. Related Work

Nowadays, social networks have become spaces where people communicate their daily activities, concerns and problems to others [4]. Recently, many studies have emerged in which social networks were used as a data source produced by users to attempt to draw conclusions on health issues [5].

One of the social networks that has experienced this growth in recent years has been Twitter. Because of this, recently a large body of research has been devoted to use Twitter to analyse users behaviour and personality, including possible suicidal tendencies [6], people on diet [7], influence of political candidates on voters [8], users characteristics for engaging customers [9], and many others. Many of these works make use of common text-based features [10] such as keyword extraction, topic classification, Part-of-Speech and Name Entity Recognition.

In addition to the different fields we have mentioned in the previous section, research on sleep disorders has also been carried out using social networks prior to this study. These researches are based on comparisons between a group of users who suffer from insomnia and another group of users who do not suffer from it.

Jamison-Powell et al. [11] analysed a corpus of English Tweets in order to explore the role of social media in the discussion of mental health issues, and with particular reference to insomnia and sleep disorders. They present a content analysis which revealed that tweets that contained the word “insomnia” contained significantly more negative health information than a random sample, strongly suggesting that individuals were making disclosures about their sleep disorder. In addition, they also reveal the existence of two themes present in the content of this type of tweets: (i) describing the experience, in which they catalogue the way in which users transmit their insomnia and (ii) coping with insomnia, in which they catalogue how users manage their insomnia. They determine a tendency of users with insomnia to release their frustration of not being able to sleep through social networks. Based on this conclusion, we claim that it is possible to carry out studies on insomnia through the information provided by users suffering from this disorder on twitter.

Another study has also been carried out on the Chinese social network Sina Weibo, but based on the same context [12] of insomnia. Their objective was to evaluate the feasibility of using social media to understand whether and how sleep complaints are discussed online and the feasibility of using their data to inform the detection and prevention of insomnia. To do this, they annotated a corpus in terms of themes and symptoms present in the tweets. For the task of this labelling they developed a text classifier machine learning. They also conducted a demographic study of the tweets and a study of activity on the social network of users. Their results have allowed us to know the most common topics among users suffering from insomnia.

Finally, another relevant research has been done to characterize sleep issues on Twitter [13]. Its objective was to determine whether social media can be used as a method for carrying out research focused on sleep-related issues. To do this they built a corpus of tweets in English containing words related to insomnia. Their conclusions were that users with insomnia present a tendency of negative sentiment in their publications. They also carried out a study of the characteristics of these users on Twitter, concluding that they are less active and have fewer followers.

III. Insomnia Corpus Collection and Annotation

A. Dataset and methodology

The capture of tweets was done through the Streaming Twitter API[2] that allows us to capture tweets in real time. For this task we used an open-source Python library called Tweepy [14] that allows us to communicate with the Twitter API in a simple and intuitive way.

The following criteria were considered for collecting the tweets:

- They must have been in Spanish.
- They had to contain the word insomnia, because it is the phenomenon we want to study.
- They could not be re-tweets (a re-post of a tweet originally posted by a different user) because we are only interested in experiences expressed by people who claim to have insomnia.

In addition to the tweet text, we also store metadata such as the date and time the tweet was published, the user ID and his location, and the tweet ID.

The data collection process has been carried out in two different periods. The first data capture (*General Insomnia dataset*) was made between December 14, 2017 and January 4, 2018. It produced a sample of 54432 tweets following the aforementioned criteria. This sample has been used in the geographical study of insomnia explained in the Sect. III-B. The second capture (*Spanish Insomnia dataset*) was made between April 2 and 10, restricted to Spain and produced a sample of 2361 tweets.

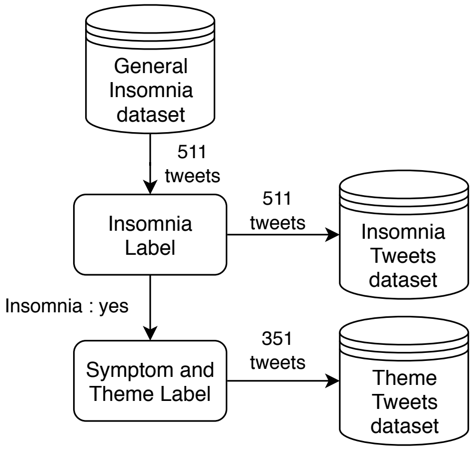


Fig. 1. Methodology for the labelling and creation of datasets.

[2] https://developer.twitter.com/en/docs

From this data, several datasets have been created for our experiments (see Fig. 1).

- *Insomnia Tweets dataset*. This dataset is formed by 511 random tweets of the *General Insomnia dataset*. In it we labelled the *Insomnia* label explained in the Sect. III-C, and it is the one we used to train the *Insomnia Classifier*, explained in the Sect. V-A.
- *Theme Tweets dataset*. This dataset is made up of the 351 tweets of *Insomnia Tweets dataset* dataset tagged with the *Insomnia* label as 'yes' (i.e. tweets in which the author claims to suffer insomnia). In this dataset we labelled the labels *Themes* and *Symptoms* explained in the Sect III-C, and we used it to train the *Theme Classifier*, explained in the Sect V-B.
- *Spanish Insomina Users dataset*. This dataset has been formed with the aim of drawing conclusions about the activity of users with insomnia on a social network such as Twitter. This dataset is formed by 103 users with insomnia based on the *Spanish Insomnia dataset* using the methodology described in Sect. V-C.

B. Geographic Research

The first study we did was a geographical distribution of the 54432 tweets we collected in the first capture. The objective of this study is to understand the importance of insomnia in Spanish-speaking countries.

The location information of a tweet can be extracted in two different ways: (i) the tweet can be geolocated directly by the platform (users can optionally choose to allow their tweets to be geolocated using a GPS system) (ii) using the "Location" field associated with the tweet (Twitter API also provides information about the user who has written the tweet).

When analysing our *General Insomnia dataset*, we determined that only 1% of the tweets contained information about the location from which they were published. For this reason, we carried out the analysis of the location fields of the users who had written the rest of the tweets. To do this we use the Google Maps Geocoding API[3] which allows us to convert this field in text to geographic coordinates, and so be able to work with the locations in a more effective way than with texts [15].

With this procedure, we obtained 32404 geolocated tweets that is 59% of our sample.

For the part of visualizing the geolocation of the tweets we have used the Python GeoPandas [16] library that allows us to represent the geolocalized tweets in their corresponding country, as shown in Fig. 2.

TABLE I
TOP COUNTRIES IN TWEETS OF THE GENERAL INSOMNIA DATASET.

Country	Tweets	Percentage
Argentina	11132	34.53%
Mexico	5349	16.59%
Spain	5112	15.85%
United States	3052	9.46%
Colombia	1884	5.84%

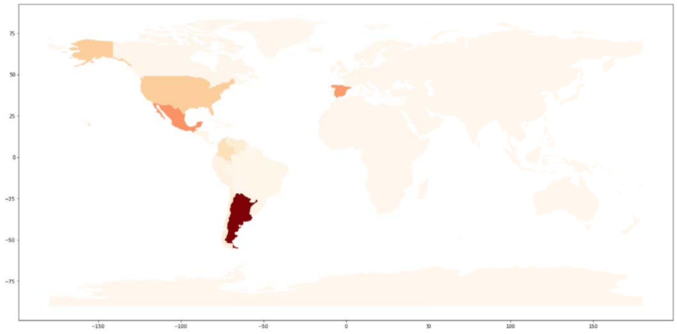


Fig. 2. Origin countries of the General Insomnia dataset.

We have also highlighted the countries with more tweets with the word insomnia as shown in Table I, in which it is important to highlight that Argentina, Mexico and Spain have a much greater presence than the rest of the Spanish-speaking countries.

Specifically, in the case of Spain, we want to determine the importance of insomnia in the Twitter user community.

To do this we have determined the percentage of users who have posted at least one tweet in a given period with the word insomnia. This percentage is given by the number of different users that we have obtained from our dataset from Spain among the total number of users that have written in that country during the time of capture that has been three weeks.

The number of different users who have posted in the three weeks of capture at least one tweet about insomnia is 1982. This number has to be compared with the number of active Spanish users in the period of three weeks who have the location field filled in, as our sample only includes users who have an insomnia tweet and the location field filled in. The number of users in Spain who have a written localization field is 2.9% [17].

With this number and the data that there are 7.52 million active Twitter users (they write at least one tweet) for a month in Spain[4], it can be obtained that there are 164124 active users with the location filled in during the time of capture.

With these three facts, we can determine that approximately 1.21% of users in Spain have ever written about insomnia.

C. Labelling

Our goal is to detect and characterize tweets from users who claim to have insomnia as a supervised classification task and two text classifiers was trained to perform this task.

The *Insomnia Tweets* dataset has been labelled with the *Insomnia* label and the *Theme Tweets* dataset has been labelled with the *Themes* and *Symptoms* labels.

1) Insomnia: The first labelling we did was in *Insomnia Tweets* dataset and it consisted of differentiating between tweets that contain the word insomnia and come from a person who claims to have insomnia and those who do not.

[3] https://cloud.google.com/maps-platform/

[4] https://es.statista.com/

Specifically, this tag is added to tweets that match these characteristics:

- They do not come from a corporate account, that is, the user who has written the tweet is a natural person.
- In the tweet the user informs that it is the user who suffers from insomnia, so there are no valid tweets where it is transmitted that an acquaintance suffers from insomnia for example.

It is important to remember that none of the tweets that form the dataset are re-tweets, a filter that was implemented in the capture process. The Insomnia Tweets dataset distribution is shown in Table. II

TABLE II
DISTRIBUTION OF INSOMNIA TWEETS DATASET AND AVERAGE NUMBER OF WORDS

Insomnia	N (%)	Avg. no. words	Example
yes	351 (69)	12.57	I have insomnia again
no	160 (31)	11.61	Do you have insomnia?

The average number of words has been included in order to provide information on the size of the dataset. For this average we have considered all the words contained in a tweet, including stop words

2) Themes: The second labelling was done in the *Theme Tweets* dataset and it aims to group tweets by type of information, allowing us to better characterize the messages and thus draw better conclusions. We remember that this dataset is the one we will use to train the *Theme Classifier*. The topics we defined were:

- User expresses only his sleep disorder.
- User expresses his sleep disorder and gives the causes of his problem.
- User expresses his or her disorder and requests help.
- User expresses his sleep disorder and transmits what he does at night.
- User expresses his/her sleep disorder and takes some action to try to solve it.

The distribution of this label in *Theme Tweets* dataset is shown in Table III.

TABLE III
DISTRIBUTION OF THEME LABELS IN THEME TWEETS DATASET AND AVERAGE NUMBER OF WORDS

Theme	N (%)	Avg. no. words
Only sleep disorder	214 (61)	10.6
Sleep disorder and cause	34 (9.7)	17.11
Sleep disorder and requests help	28 (8)	11.9
Sleep disorder and current activity	41 (11.6)	16.22
Sleep disorder and current therapy	34 (9.7)	16.6

The most worrying fact is that only 9.7% say they take some action to solve it. This percentage shows that there is a need to develop techniques for the detection of sleep disorders such as insomnia in order to help and diagnose those affected and thus increase the number of people who take some remedy.

3) Symptoms: As mentioned above, we are interested in knowing what are the different symptoms of tweets for users with insomnia. In the third edition of International Criteria for Sleep Disorders (ICSD-3), the main symptoms of insomnia were defined as: difficulty in starting sleep; short sleep duration; difficulty in sleeping and low energy during the day.

As in the case of *Theme* labels, we have labelled *Theme Tweets* dataset in each of the three defined symptoms. The distribution of this label in *Theme Tweets* dataset is shown in Table IV.

TABLE IV
DISTRIBUTION OF SYMPTOMS LABEL AND AVERAGE NUMBER OF WORDS IN THEME TWEETS DATASET

Symptom	N (%)	Avg. no. words
Difficulty in starting sleep	32 (66.7)	15
Short sleep duration	9 (18.7)	13.5
Difficulty in sleeping and low energy during the next day	7 (14.6)	20.7

The first conclusion we have drawn from of this labelling was that most of the tweets did not express any of the defined symptoms. Specifically, only 14% (48) of the posts commented on any of the symptoms. The small amount of tweets with this information suggests that the population does not analyse how they suffer these kinds of disorders.

IV. TEXT MINING TECHNIQUES

With the aim of automating the annotation task, two text classifiers have been developed: (i) one that performs the task of tagging tweets that meet the characteristics of the insomnia tag; and another (ii) that classifies the tweets among the defined themes. Regarding the implementation, Python libraries `scikit-learn` [18] and `NLTK` [19] have been used; for orchestration and providing the system with scalability capabilities, we have used Luigi [5].

As for feature extraction methods, we separate these into two categories: statistical and semantic methods. In this section, the developed steps –preprocessing, feature extraction (both statistical and semantic), and classification– are described.

A. Preprocessing

In this phase the raw text is taken and cleaned using common NLP techniques [20]: elimination of punctuation marks, stop-words, rare characters, URLs, etc. Besides, the tokenization has been done oriented to the domain, since Twitter has specific characteristics [21] that need to be addressed: user mentions are normalized (e.g., *@potus* is transformed to *user*), and appearance of hashtags is flagged. Finally, tokens are stemmed using the Porter method.

B. Statistical feature extraction

Regarding the dataset-specific statistic features, we consider TF-IDF weighting over n-grams. In this way, these features encode information regarding the presence of n-grams, adding

[5] https://github.com/spotify/luigi/

frequency statistics of those same n-grams. After performing a preliminary evaluation, we have found that using n-grams with n ranging from 1 to 3 (1-grams, 2-grams and 3-grams) yields the better results. To a more extent description of the TF-IDF measure, we refer to the reader to [22].

We also consider the use of Latent Dirichlet Allocation (LDA) features as a baseline. This algorithm considers each document as a mixture of several topics, and each word in the document belongs to one of the topics in the document. LDA works by calculating the probability that a document belongs to a topic, based on the probability of each of the words that make up the document [23].

In addition, the Labeled LDA (L-LDA) feature extraction algorithm is used. In contrast to LDA [23], which is a modeling algorithm that discovers textual topics in an unsupervised manner, L-LDA is a variation of the LDA algorithm that models topics in a supervised setting. It does not only indicate the number of topics that are latent in a set of document, but it does so by using the topic labels [24].

The L-LDA method works as follows. Firstly, we used the Stanford Topic Modelling Toolbox (STMT) [25] tool which generates a dictionary for each tagged topic composed of all the words in the dataset and the probability that each of those words belongs to that particular topic. As a result, the algorithm is able of estimating the probabilities that describe if a certain document belongs to each of the topics.

C. Semantic feature extraction

This paper presents a feature extraction method that uses the information contained in a word embedding model and a selection of relevant words to represent a document. This method makes use of the similarity measure that can be done in a word embedding model via cosine similarity.

In this way, we consider a selection of words S that serves as a "template" to wich project input documents. Given a input document (e.g., tweet), the similarity between the input word vectors of that document and each of the words in S is computed. This can be done because the embedding model maps each word to a vector in a n-dimensional vector space. For an input word the method computes a similarity vector; when done for the whole document, a matrix of dimensions $m \times |S|$ is obtained, where m is the number if input words in a certain document. In order to reduce the dimensionality to a feature vector, the maximum is computed column-wise. The final dimensionality of the extracted feature vector is $|S|$, that is, the number of selected words.

The selection of relevant words is independent to the method, and can be implemented in a number of ways. This work tackles two different approaches for this word selection.

- Firstly, the words of a sentiment lexicon are used. The idea is that insomnia users probably show a bias towards negative sentiment when complaining about insomnia, while users that do not suffer insomnia may not complain, and thus do not use negative sentiment. For this, we use the Spanish sentiment lexicon *ElhPolar* [26].
- Secondly, domain-wise words have been extracted. For this, most frequent words have been collected from our dataset. Please note that these words have been extracted after the preprocessing steps (Sect. IV-A).

The proposed feature extraction method can be seen as a projection of the input text to a selection of words. Generated feature vectors are dense, and its dimensionality depends on the number of selected words; this value is set to 100 words with a parameter exploration. This method is, to the best of out knowledge, a novel work.

D. Classification

Finally, as a last step in out data processing pipeline, a machine learning classifier is used. This work tackles two different classification tasks, namely insomnia detection and theme classification. Thus, we use two different machine learning systems to address these tasks.

With respect to the methodology of the evaluation, all experiments have been done with cross validation. More specifically, we have used the k-fold configuration with a $k = 5$. Besides, accuracy and F1-score metrics have been used to measure the performance of the learners.

There are many options among machine learning models that can be used. In this project, as done in [12], we have evaluated the use of the following classification algorithms: Multinomial Naive Bayes, Support Vector Classification (SVC), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest Classifier. The hyperparameters of each of these algorithms have been tuned using a cross-validation grid search. In a preliminary evaluation, we have found that Logistic Regression yields better results in the two classification tasks at hand.

V. Results

A. Insomnia Classifier

For the task of determining whether a tweet has been written by an user that suffers insomnia or not (sect. III-C), a machine learning classifier has been trained. As we have already mentioned, we have used the *Insomnia Tweets* dataset to train this classifier Consequently, to facilitate this training, we have balanced the dataset with under-sampling, resulting in a total of 320 tweets. In this task, the feature extraction methods used are: TF-IDF combined with LDA, and semantic similarity (Sect. IV-C) with the two variations considered: using the sentiment lexicon or a domain-wise word selection.

TABLE V
Evaluation results in insomnia detection.

Method	Accuracy	F1-Score
TF-IDF + LDA	0.84	0.82
Semantic Similarity (lexicon words)	0.85	0.85
Semantic Similarity (domain words)	**0.89**	**0.90**

Table V shows the results for the evaluation in this task. It can be seen that the semantic similarity features perform, by a large margin, better that the ones extracted by the TF-IDF + LDA combination. Also, using a domain-adapted word selection leads to further improvement, reaching a 90% of F1-score. This result indicates that the proposed feature extraction methods yields good performances. Also, the fact that using a domain-adapted selection increases the performance suggests that our method can be easily used for other text categorization tasks just by changing the selection of words.

B. Theme Classifier

Following, the task of determining the *Theme* label (Sect. III-C) of a message is performed similarly as in Sect. V-A. As described in Table III, the messages are distributed along five categories that represent the themes. From this distribution, it can be seen that there is a considerable disproportion of tweets among the aforementioned themes. Therefore, and to avoid problems of unbalanced data, a sample subset of 172 tweets from the *Theme Tweets* dataset has been selected, so that the categories distribution was approximately uniform for each theme.

Regarding the feature extraction, we have evaluated the use of TF-IDF representations and, due to the nature of the problem, the L-LDA feature extraction method (Sect. IV). As described, L-LDA is able to effectively learn topics by means of annotations, characteristic that is exploited in this task.

TABLE VI
EVALUATION RESULTS IN THEME DETECTION.

Method	Accuracy	F1-Score
TF-IDF	0.57	0.58
TF-IDF + LDA	0.51	0.52
TF-IDF + L-LDA	**0.78**	**0.78**

The obtained results are shown in Table VI, where different feature extraction methods are evaluated. It can be seen that TF-IDF and LDA features do not lead to a good performance in this task. This suggests that the information contained in such features is not useful for the task of theme detection. Nevertheless, the evaluation shows that using L-LDA method highly improves the performance of the system, reaching a 0.78 in the F-score. In light of these results, it is safe to assume that such improvement is due to the use of the L-LDA method. Since the task of theme dection involves, at least to some extent, topic detection, we hypothesize that the features extracted by a supervised topic detection method are of use.

C. Social Network activity

As we have commented in the Sect. III-A, we have made a capture of tweets with the word insomnia from Spain. The aim of this capture was to form the *Spanish Insomnia Users* dataset to be able to determine, through their timeline, a pattern of activity on a social network such as Twitter of people suffering from insomnia, and compare it with the activity of a set of normal users.

The process of selecting a user for the Sleep Group (set of users with insomnia) is described in the Fig. 3.

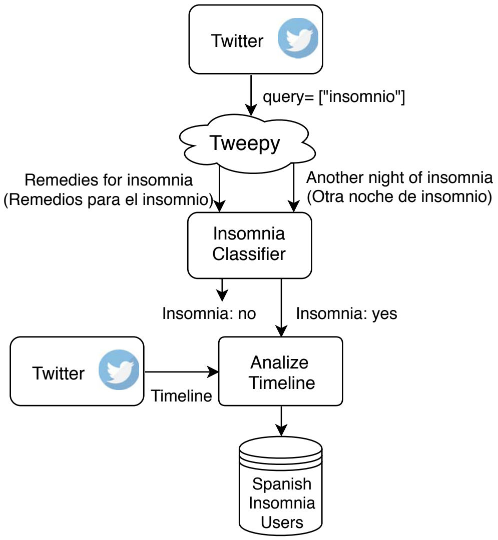


Fig. 3. Process to elaborate the dataset of users.

Basically, this process consisted of evaluating each of the tweets captured with the *Insomnia Classifier*, so that only the tweets of users who claim to have insomnia would follow in the process. After this first filter, we analysed the last 200 tweets of the account associated with each tweet; if in these 200 tweets there were 2 or more tweets in which the user mentioned the word ”insomnia”(insomnio) or ”I can't sleep”(no puedo dormir), that user became part of the *Spanish Insomnia Users* dataset.

As we have already mentioned we want to determine a pattern of activity of users with insomnia and normal users, so we also perform a user dataset without insomnia (Normal Group). To form this group we made a random capture of tweets from Spain, and hand filtered to have only tweets from people and thus not have in the study tweets from accounts from some kind of organisation. We also analysed the timeline of these users to make sure they didn't have any insomnia-related tweets.

The first feature that we focused on was the number of tweets related to insomnia that each of the *Spanish Insomnia Users* had (Fig. 5). We can observe that most of the users have two relatively recent messages related to insomnia (72%).

After that, we analyse the time associated to all the tweets of the timeline of each one of the users of the two groups. In the Fig. 4 we have plotted the percentage of tweets per hour of average that the users of each of the groups post.

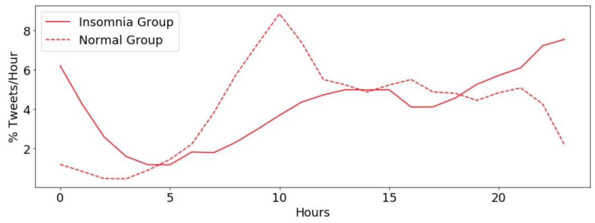


Fig. 4. Activity on Twitter per hour in users with and without insomnia.

It can be seen that users who belong to the *Spanish Insomnia Users* dataset have a greater activity throughout the night due

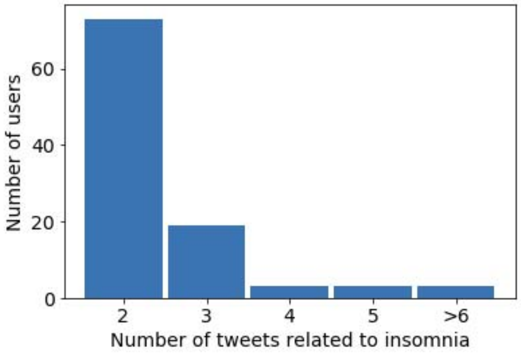


Fig. 5. Distribution of the Spanish Insomnia Users dataset according to number of Insomnia tweets on their timeline.

to the inability to fall asleep. During the morning, the activity of these users may decrease due to nighttime fatigue and in exchange the Normal Group has its highest point of activity around 10 a.m. In the afternoon, the activity is balanced, until the evening, when the users with insomnia again show more activity.

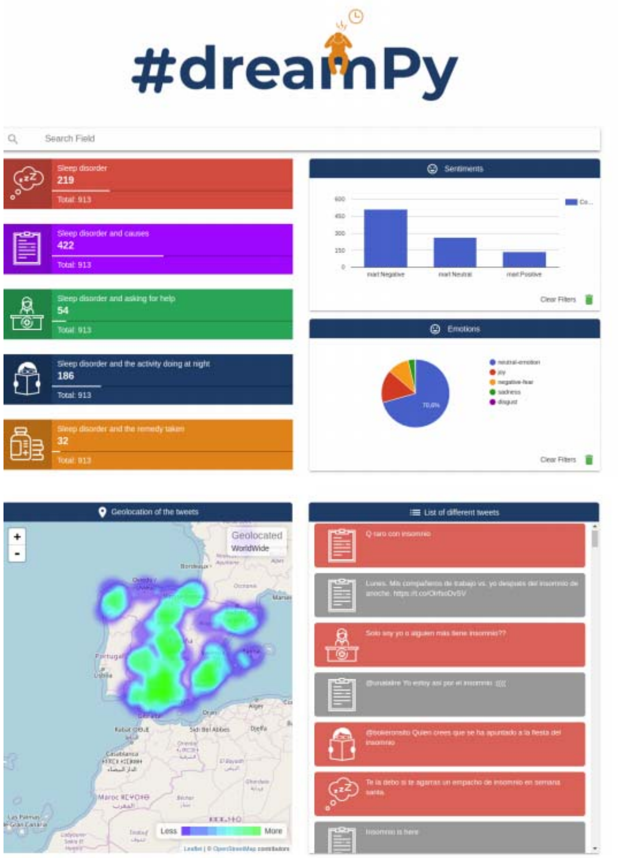
An online service [6] has been developed with the aim of being able to analyse Twitter messages about insomnia. The interface of this service is shown in Fig. 6. In addition to the analysis we have carried out in this study, this system also allows us to analyse both the sentiments and the emotions transmitted by users with insomnia through the Senpy [27] tool.

VI. Conclusions and Outlook

This research has been carried out through messages related to insomnia published on the social network Twitter with the aim of obtaining information on how this disorder was represented in the users of this social network. Specifically, it has been done through Spanish-speaking posts and users that, as far as we know, is something that has not been done previously and that, due to the large number of people who speak this language, in our opinion was very necessary.

In relation to the proposed objectives, we have characterized the presence of insomnia in the main Spanish-speaking countries and, specifically in Spain, we have been able to quantify the presence of insomnia in its community of users in which we have determined that approximately 1.21% (corresponding to 54450 users) write at least one tweet per month related to insomnia. This conclusion determines that insomnia is a disease that is very present in our society today.

Through the analysis of the posts related to insomnia we have been able to characterize the way in which users transmit this disorder, however the conclusions we have been able to draw are limited by the size of our dataset and by the number of tweets we have from some of the themes. As we have mentioned, 61% of tweets only give information that the user suffers from insomnia, this makes it difficult for us to determine a characteristic that is common to the context in which all tweets about insomnia are found.



#dreamPy

Search Field

Sleep disorder
219
Total: 913

Sleep disorder and causes
422
Total: 913

Sleep disorder and asking for help
54
Total: 913

Sleep disorder and the activity doing at night
186
Total: 913

Sleep disorder and the remedy taken
32
Total: 913

Sentiments

Clear Filters

Emotions

Clear Filters

Geolocation of the tweets

List of different tweets

Q raro con insomnio

Lunes. Mis compañeros de trabajo vs. yo despues del insomnio de anoche. https://t.co/OlrfsoDvSV

Solo soy yo o alguien más tiene insomnio??

@urudalire Yo estoy así por el insomnio :(((

@bokeronsito Quien crees que se ha apuntado a la fiesta del insomnio

Te la debo si te agarras un empacho de insomnio en semana santa.

Insomnio is here

Fig. 6. Dashboard of the monitoring service

However, the most worrying thing is that only 9.6% of the tweets analysed contained information about some kind of measure or treatment that users perform to try to solve their problem with insomnia. The conclusion we draw from this fact is that the majority of users do not give the necessary importance to insomnia, which supports our idea that studies and tools are needed to understand and diagnose this disorder more effectively.

Regarding the users that form the *Spanish Insomnia Users* dataset, we have been able to determine that these users have a significantly higher activity posting tweets during the night hours regarding the activity of a normal user group. In addition, this dataset has been captured using the *Insomnia Classifier* developed in the project, so seeing the results we have just commented on the night activity, we can determine the correct functioning of this classifier when determining tweets about insomnia.

In addition to this, this paper explores the possibilities of applying machine learning and NLP techniques to the task of determining whether a Twitter message has been written

[6] http://dreampy.cluster.gsi.dit.upm.es/

by an user that suffers insomnia or not. In this sense, this work proposes a feature extraction mechanism that exploits the similarity measure that can be computed in a word embedding model. Moreover, this approach can be adapted to a certain domain by means of tuning word selection processes. This feature extraction method has been evaluated using the captured dataset. Obtained results highly indicate that this method can yield high performance in this task, and that domain adaptation can further improve these results. This suggest that the proposed method can be generalized to different domains or even NLP tasks.

As we have already mentioned, from the analysis of the tweets on insomnia we have been able to conclude that the number of tweets that contain relevant information for the study of this phenomenon is small. Therefore, we consider that the Insomnia Monitoring System developed is another great contribution to the research and development of systems for the analysis of insomnia as it allows the capture and analysis of only the data that interest us in an automated way.

Our future work focuses on a more detailed study of the users that form the *Spanish Insomnia Users* dataset. So far, we have focused the entire study on posts from users who already claim to have insomnia. Therefore, our goal will be to conduct a more detailed study on the timeline of all Spanish insomnia users. Specifically, we want to study the change of emotions and feelings that these users have about time and try to determine patterns that allow us to investigate for the prediction of users who may suffer insomnia in the future.

Acknowledgements

This research work is supported by the Spanish Ministry of Economy and Competitiveness under the R&D projects SEMOLA (TEC2015-68284-R) and ITEA3 Emo-Spaces (ITEA3-14012/RTC-2016-5053-7) and for the project ITEA3 SoMeDi (ITEA3-15011).

References

[1] D Leger, B Poursain, D Neubauer, and M Uchiyama. An international survey of sleeping problems in the general population. *Current medical research and opinion*, 24(1):307–317, 2008.

[2] Stacey J Elder, Ronald L Pisoni, Tadao Akizawa, Rachel Fissell, Vittorio E Andreucci, Shunichi Fukuhara, Kiyoshi Kurokawa, Hugh C Rayner, Anna L Furniss, Friedrich K Port, et al. Sleep quality predicts quality of life and mortality risk in haemodialysis patients: results from the dialysis outcomes and practice patterns study (DOPPS). *Nephrology Dialysis Transplantation*, 23(3):998–1004, 2007.

[3] Michael J Sateia. International classification of sleep disorders. *Chest*, 146(5):1387–1394, 2014.

[4] Lucia Falzon, Caitlin McCurrie, and John Dunn. Representation and analysis of twitter activity: A dynamic network perspective. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1183–1190. ACM, 2017.

[5] Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates. *Journal of medical Internet research*, 20(5):e168, 2018.

[6] Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer, 2014.

[7] Sherry Pagoto, Kristin L Schneider, Martinus Evans, Molly E Waring, Brad Appelhans, Andrew M Busch, Matthew C Whited, Herpreet Thind, and Michelle Ziedonis. Tweeting it off: characteristics of adults who tweet about a weight loss attempt. *Journal of the American Medical Informatics Association*, 21(6):1032–1037, 2014.

[8] Sanne Kruikemeier. How political candidates use twitter and the impact on votes. *Computers in Human Behavior*, 34:131–139, 2014.

[9] Shintaro Okazaki, Ana M Díaz-Martín, Mercedes Rozano, and Héctor David Menéndez-Benito. Using twitter to engage with customers: A data mining approach. *Internet Research*, 25(3):416–434, 2015.

[10] Seyed-Mehdi-Reza Beheshti, Alireza Tabebordbar, Boualem Benatallah, and Reza Nouri. On automating basic data curation tasks. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 165–169. International World Wide Web Conferences Steering Committee, 2017.

[11] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1501–1510. ACM, 2012.

[12] Xianyun Tian, Guang Yu, and Fang He. An analysis of sleep complaints on Sina Weibo. *Computers in Human Behavior*, 62:230–235, 2016.

[13] David J McIver, Jared B Hawkins, Rumi Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P Fitzgerald, Sachin H Jain, and John S Brownstein. Characterizing sleep issues using twitter. *Journal of medical Internet research*, 17(6), 2015.

[14] Joshua Roesslein. tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5, 2009.

[15] Gabriel Svennerberg. *Beginning Google Maps API 3*. Apress, 2010.

[16] K Jordahl. Geopandas: Python tools for geographic data. *URL: https://github. com/geopandas/geopandas*, 2014.

[17] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

[18] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[19] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

[20] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[21] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[22] Ray R Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.

[23] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[24] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[25] D Ramage and E Rosen. Stanford topic modeling toolbox (stmt), 2009.

[26] X Saralegi Urizar and I San Vicente Roncal. Elhuyar at tass 2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013)*, pages 143–150, 2013.

[27] J Fernando Sánchez-Rada, Carlos A Iglesias, Ignacio Corcuera, and Oscar Araque. Senpy: A pragmatic linked sentiment analysis framework. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 735–742. IEEE, 2016.