MIMIC: a Multi Input Micro-Influencers Classifier

MIMIC: a Multi Input Micro-Influencers Classifier

Simone Leonardi, *Politecnico di Torino*, Luca Ardito, *Politecnico di Torino*

***Abstract*—Micro-influencers are effective elements in the marketing strategies of companies and institutions because of their capability to create an hyper-engaged audience around a specific topic of interest. In recent years, many scientific approaches and commercial tools have handled the task of detecting this type of social media users. These strategies adopt solutions ranging from rule based machine learning models to deep neural networks and graph analysis on text, images and account information. This work compares the existing solutions and proposes an ensemble method to generalize them with different input data and social media platforms. The deployed solution combines deep learning models on unstructured data with statistical machine learning models on structured data. We retrieve both social media accounts information and multimedia posts on Twitter and Instagram. These data are mapped into feature vectors for an eXtreme Gradient Boosting (XGBoost) classifier. Sixty different topics have been analyzed to build a rule based gold standard dataset and to compare the performance of our approach against baseline classifiers. We prove the effectiveness of our work by comparing the accuracy, precision, recall, and f1 score of our model with different configurations and architectures. We obtained an accuracy of 0.98 with our best performing model.**

***Index Terms*—Deep learning, gradient boosting, image processing, micro-influencers, nlp, social media.**

I. Introduction

The recent COVID-19 pandemic highlighted how companies advertising products through digital platforms and social media influencers were able to thrive even during the lockdown period [4]–[6]. This event led people to spend even more of their time on social media platforms seeking information and advice. The presence of content creators able to sponsor effectively products and messages both from private companies and public institutions influenced the social and economic behavior of a large slice of the worldwide population [7]. In literature, social media users with a high influence power are defined as influencers [8]. Micro-influencers are a specific category of influencers, they tend to be more engaging and they specialize their content over a few topics of interest. They have smaller audiences than famous influencers (5k-100k followers), at the same time they are able to persuade a larger percentage of their community [9]. These characteristics are translated into a higher return on investment for those who employ them. Existing studies seek both influencers and micro-influencers exploiting social graph information, text and images retrievable from social media posts. This heterogeneous input guides researchers towards preferring specialized solutions to maximize the accuracy in the classification of specific characteristics of micro-influencers such us their ability to reach the greatest number of followers rather than empathize with followers with similar tastes about the emotions aroused by images or words [10]–[12]. This work proposes a solution that combines image captioning, text processing and social media graph features to classify micro-influencers. Once selected the topics of interest, our process builds a list of users writing about them. We filter the social media accounts not matching a set of metrics that describes them as potential micro-influencer and we label them creating a balanced dataset for each topic. Eventually we collect and analyze their writings to check the evidence from the initial filters.

S. Leonardi and L. Ardito are with the Software Engineering group in the DAUIN Department of Control and Computer Engineering, Politecnico di Torino, Turin, TO, 10129, IT e-mail: (see https://softeng.polito.it/leonardi/).

Manuscript received - ; revised -.

We have defined three research questions:

- RQ1: Which gold standard can be used for classifying micro-influencers in the context of online social media?
- RQ2: How to classify micro-influencers with comprehensive multi input data?
- RQ3: Is an ensemble method based on gradient boosting more effective than deep neural network models in the micro-influencers classification task?

The content of the next sections is the following. In Section II, we present an overview of the related work published in the research field of Online Social Media with a specific regard to the area of influencers and micro-influencers detection and classification. In Section III, we explain how we create a rule based Gold Standard, a ranking strategy and a XGBoost model to classify micro-influencers. In Section IV, we show the results obtained during the evaluation of the accuracy, precision and recall of our model compared with different baseline classifiers. In Section V, we highlight the difficulties surfaced during the experiments and some possible improvements to deal with them. In the last Section VI, we summarize the outputs and the outcomes of our work and we give some suggestions to keep working on this research topic for future works.

II. Related Work

The detection or classification of influencing people online is a research field in the context of Social Network Analysis (SNA). It has produced many publications [13]–[19] and it is still gaining importance due to its socio-economic impact. In [20], Rabiger et al. describes the dynamics of influence and interaction between users as properties of complex graphs made of nodes and connections. The social graph varies based on the problem analyzed and the model describing friendships. In [21], Lü et al. divide the existing approaches to detect influencers into two categories. The first one contains methods of influence maximization by the identification of influential users in accordance with a diffusion model. The second category collects solutions investigating influence measurements to inspect the social media graphs searching locally influential nodes by the scansion of the network. In [22], Kwak et al. present a centrality based approach. They find influencers with the analysis of structural information. The definition of centrality applied in their work is described in the graph theory as a measure of importance of a node within a graph. The further improvements of this work rely on the possibility to adapt to different social network topologies in real world applications. In parallel, in [23], Chen measures how extensively a network collapses or reduces its functionality when a node is removed and its connections are broken. In the last decade, machine learning approaches emerged in this research area thanks to their capability of adapting to unstructured network topologies. In [24], Fan et al. develop a deep reinforcement learning framework to identify key nodes inside a social media graph. They obtain better performance when the quality of the learning set and available data is optimal, while the effectiveness of this approach decreases with noisy data. In [11], Roelens et al. proposes a mixed approach exploiting network analysis and machine learning algorithms. They apply an influence cascade through the network exploiting links between nodes. Even if this method relies on network parameters, it also takes in consideration accounts information and users' behavior to improve the accuracy on large scale networks. In [25], Gan et al. develop a fully data driven approach to detect micro-influencers with a specific ranking system to match them with the companies showing similar characteristics. They exploit both visual and textual information in a bilinear pooling method as described in [26] by Kim et al. The model applies a linear transformation and a non-linear activation function on each feature to balance their weights due to their different sizes in input. In their work, Gan et al. measures micro-influencer engagement power and his similarity with respect to the associated brand in a K-buckets system. In [10], Bashari et al. focus their method on the sole text analysis with a natural language processing pipeline. They collect User Generated Content (UGC) and they do not consider user interactions on the social media platform. After a data cleaning and preprocessing phase, they apply a term frequency–inverse document frequency (TF-IDF) on the UGC weighting each word and then they map these features with captions and hashtags. In the second phase, Bashari et al. input these features into two Support Vector Machines to classify users in a supervised scenario. In a similar way, in [27], Zheng et al. analyze keywords with their on-Demand Influencer Discovery (DID) framework. Even in this scenario, they do not consider the popularity nor the links between users. Their model adopts a Language Attention Network o select social posts related to the given keyword and an Influence Convolution Network to mold the influence propagation on social media with neighborhood aggregation techniques. Each word is mapped to an one-hot encoding representation, then the matrix of transformed words is given as input to a bidirectional Recurrent Neural Network (RNN) to learn a hidden state of each word. A final attention layer retrieves the matrix hidden state and it produces a classification output mixing it with an external topic seed to capture subject related information in the original post. A recent work by Zhuang et al. [12] identifies influencers with a multidimensional social influence (MSI) measurement approach. They propose a framework that is comprehensive in different aspects of influencing mechanisms on social media. They describe information influence measurements, action influence measurements and structure influence measurements. Even if we are inspired by all the presented research works, we develop a model that is able to ensemble existing solutions into a more general framework creating a pipeline that deals with textual, visual and account based information from social media. Our approach is described in detail in Section III.

III. Approach and Contribution

The following section describes how we collected data from Twitter and Instagram to build a gold standard dataset. Then it illustrates our pipeline to create a micro-influencer classification model. We also detail how we exploit visual and textual features of social media posts. The last part of this section presents the application of XGBoost (eXtreme Gradient Boosting) to enhance the performance of our model.

A. Dataset creation and gold standard

We defined rules to create a gold standard in the form of an annotated dataset in the research field of micro-influencers. We performed the entire pipeline: data retrieval from social media, definition of rules and thresholds to label users as micro-influencers and final dataset creation. The rules defined are described below. They follow the definition of micro-influencer as described by Brewster and Lyu [9].

Age counts how many days have passed from the user account creation.

Followers count numbers how many users follow the potential micro influencer. The following concept is unilateral, so the followed user does not need to follow back the agent of the action. The user needs to have between 5k and 100k followers to pass the filter. Outside of this range, a user is considered a nano-influencer, macro-influencer or influencing at all.

Followers growth rate measures on average how many new followers a user acquires each day. It scores the ability of the user to constantly enlarge his community. This marker defines the future potential of the user to reach more people interested in the topics he produces posts about. The threshold for this score is above 4.

$$Followers\_growth\_rate_i = \frac{Followers\_count_i}{Age_i} \tag{1}$$

In equation 1, i stands for the i^{th} user in our dataset.

Followers following ratio is a score computed to detect and exclude potential fake accounts called bots that automatically generate social media posts and that randomly follow other social media users to obtain a follow back action in exchange. This definition is in-depth explored in the work of Yang et al. [28]. This approach makes the fake account in the conditions to grow quickly but it leaves trails in the huge number of followed accounts. For these reasons, we set a filter on this score to be above 2. The user needs to have at least double the

TABLE I: Twitter scores and thresholds adopted to select a user as micro-influencer in the initial phase of dataset collection and Gold Standard definition. These thresholds follow the definition of micro-influencer given by Brewster et al. in [9].

Score	Threshold
followers_count	>5k and <100k
followers_growth_rate	>4
followers_following_ratio	>2
verified	false
tweet_frequency	>10
statuses_count	>200

number of followers with respect to the number of accounts followed.

$$Followers\_following\_ratio_i = \frac{Followers_i}{Following_i} \tag{2}$$

In equation 2, i is the i^{th} user in our dataset.

Verified is a boolean value that is a label given by Twitter when an account respects parameters to be defined as authentic, notable and active. Even if the authentic and active rules allows us to exclude fake account, we have already defined a score for that option, instead the notable facet of this label means that the user is *in the top .05% follower or mention count for your geographic location, it may count towards notability evidence for certain categories* [1]. According to this definition, we decide to accept as micro-influencers only those users with this label set as *False* because it is a signal that he is not already really famous as macro-influencers or brand-celebrities.

The last two metrics selected to filter micro-influencers on Twitter are tweet frequency and statuses count.

Tweet frequency scores how much the user is active on the platform. It is computed as the number of tweets posted divided by the days of account existence. The threshold for this score is above 10. If the user is below this threshold he is not able to entertain his audience every day.

$$Tweet\_frequency_i = \frac{Statuses\_count_i}{Age_i} \tag{3}$$

In equation 3, i stands for the i^{th} user in our dataset.

Statuses count is the number of tweets posted on the platform by the potential micro-influencer. We filter out users having less than 200 tweets in their timeline because there is too little content to be analyzed by our framework.

In Table I, we list all the thresholds adopted to label a user as a general micro-influencer in the Twitter social media platform. Three of these scores are also exploited to collect non micro-influencers to balance our Gold Standard by maintaining the quality of samples. A non micro-influencer is considered also if has scores in I negated with the exception of *statuses count* and *tweet frequency*. A non micro-influencer in the Gold Standard has more than 200 tweets in his timeline and a tweet frequency above ten, because otherwise it could be a fake account or a too novel user not having enough produced contents to be analyzed. We adopted a similar approach on Instagram by changing the name of the scores according to the definition given by this social media. *Followers* records the number of followers of the selected user. Also in this case, following the definition by Brewster [9], a user needs to have between 5k and 100k followers to be considered a potential micro-influencer. *Media count* counts how many posts have been posted by the user from the creation of his account on the platform. The threshold of this score is again 200 and the user needs to have more than 200 Instagram posts to be processed in our framework. *Followers per media* is the ratio between the number of his followers and the number of social media contents posted. The threshold for this number is 2. *Followers following ratio* is the same as per Twitter, in fact a user having less than 2 in this proportion is a possible spam fake account performing automated following action to obtain a follow back and so an increment in the number of his followers. As per Twitter also in the Instagram section of dataset collection and Gold Standard definition a user to be considered as a non micro-influencer ans so to be involved in the balancing of the dataset needs to have at least 200 media count and less than 5k followers or more than 100k followers while denying the other thresholds. We build a Gold Standard containing 30 heterogeneous topics, with 300 unique users and a total of 60k tweets, collecting 200 tweets per user for the Twitter section. We collect 30 topics, 300 users and a total of 15k posts with 25 posts per user. These information are recalled in Table II. The Gold Standard contains a total of 600 users labeled as micro-influencer or not micro-influencers. This Section This section answers the RQ1 with the adoption of the scores described with their relative thresholds in Table I for defining a gold standard in the micro-influencer field.

TABLE II: Gold Standard dataset numbers. We collected a total of 75k social media posts, a total of 600 different users over 60 unique topics from Twitter and Instagram social media posts.

Social	Attribute	Value
Twitter	number of topics	30
Twitter	unique users	300
Twitter	total tweets	60k
Instagram	number of topics	30
Instagram	unique users	300
Instagram	total tweets	15k

B. Micro Topic with Image Captioning and Text Processing

All the general metrics adopted in literature as described in Section II are useful to detect general micro-influencers. We notice a lack of specific topic centric metrics to understand if a user is not only a general micro-influencer but also a micro-influencer for a specific topic. We fill this gap developing four new scores related to topic expertise. We compute these scores developing two topic extraction pipelines for social media posts: visual and textual. *Image captioning* is specific for data collected on Instagram. We generate a textual description per each image retrieved. The caption is then processed to compute scores on micro topics expertise. The architecture is a stacked neural network divided into two modules. A computer vision model extracts image features and a language model translates these features into a meaningful sentence. This pipeline is shown in Figure 1 and detailed in the work of Mokady et al. [29]. The visual encoder exploits the pre-

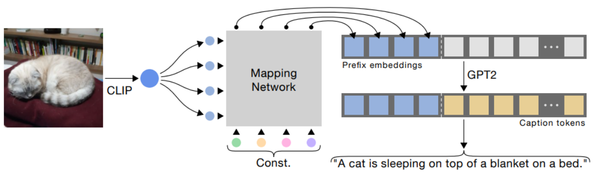

Fig. 1: ClipCap model presented in the [29] by Mokady et al. to translate image features into meaningful image captions.

trained CLIP (Contrastive Language-Image Pre-Training) to extract image features translating them into CLIP embeddings. The embeddings are then processed by a mapping network producing a prefix that is concatenated to the output to textual embedding of the caption from the dataset. The extended embeddings, made of prefix plus captions embeddings, are fed to the Generative Pre-trained Transformer GPT-2 described in details in the research work by Radford et al. [30]. The final layer of this neural network infers the caption combining the CLIP model output features and the GPT-2 mapped captions embedding. For each token produced by the sentence splitting into words, the language model outputs probabilities used to generate the next word in the sequence for future predictions selecting the most probable. In our work, the Common Object in Context COCO dataset is used as Gold Standard for image captioning [2]. This dataset provided by Google has 330k images paired with their captions. Each figure has five different captions. Once trained, the ClipCap model described in Figure 1 is able to produce a textual description of an image retrieved from Instagram social media posts. After image captioning, all input data is textual. We measure six more scores, two for Twitter and four for Instagram, to understand if a micro influencer has specific skills and interest in selected topics.

Topic % in tweets counts the percentage of tweets containing the topic searched with respect to the total number of tweets written in the user account, this ratio is defined in Equation 4 where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_in\_tweets = \frac{Total\_tweets\_with\_topic_i}{Total\_tweets\_written_i} \tag{4}$$

Topic % in words counts the percentage of words being equal to the topic searched with respect to the number of words in the entire user timeline, this ratio is defined in Equation 5 where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_in\_words = \frac{Total\_topic\_words_i}{All\_words\_written_i} \tag{5}$$

Topic_%_in_captions: counts how many captions, that are image description on Instagram, are equal to the searched topic with respect to the total number of captions written by the user in his account. It is described by the Equation 6, where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_in\_captions = \frac{Total\_captions\_topic_i}{Total\_captions\_i} \tag{6}$$

Topic_%_in_caption_words: counts the percentage of words being equal to the topic searched with respect to the number of words in all the captions, this ratio is defined in Equation 7 where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_in\_cap\_words = \frac{Total\_topic\_words_i}{All\_caption\_words_i} \tag{7}$$

Topic_%_in_pictures: counts how many image captions, obtained by the ClipClap processing, contains the searched topic with respect to the total image captions processed. This ratio is defined in Equation 8 where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_in\_pictures = \frac{Total\_img\_cap\_topic_i}{Total\_img\_cap_i} \tag{8}$$

Topic_%_in_picture_words: counts the percentage of words being equal to the topic searched with respect to the number of words in all the image captions processed, this ratio is defined in Equation 9 where *i* represents the i^{th} user in our dataset.

$$Topic\_\%\_img\_cap\_words = \frac{Total\_topic\_words_i}{All\_img\_cap\_words_i} \tag{9}$$

Thanks to the computed scores we setup a ranking strategy for micro-influencers. In Tables III and IV is listed the scoring mechanism. In the General Statistics section, in the Categories column of these tables are listed the metrics computed after data retrieval. While on the other columns headers there is the score assigned to each percentile. The topic statistics part of the tables are paired with scores computed on textual posts and image captions searching for the desired topic. Once every user collected in our dataset receives the score, he is labeled as a micro-influencer for the specific topic only if he has a score greater than the general medium score among all the users.

C. *Sentiment Analysis*

We complete the features assessment with the sentiment analysis on tweet, Instagram post description and Instagram image caption. The outputs are three percentages: positive, neutral and negative sentiment per each post. The scores per user are obtained as a mean per each sentiment over all posts in the user's timeline, obtaining a total of 9 scores three for Twitter and six for Instagram. We decide to compute these scores to detect a topic specific pattern paired with topics. One topic may be more suitable for positive messages, another for neutral or negative ones. We used the model proposed by Barbieri et al. in [31]. The model is named Cardiff Twitter Roberta Base Sentiment. Before feeding the neural network, the text is cleaned following these steps. stop words are removed, emojis are converted into text. We also remove links. Finally, we lemmatize words.

The entire pipeline from data retrieval to topic score computation and sentiment analysis is illustrated in Figure 2. After this phase, we are ready to feed our classifier with the described features. Our Multi Input Micro Influencer Classifier (MIMIC) is described in the next paragraph.

General Statistics					
	Points				
Categories	2	4	6	8	10
Followers count	5k - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	P_{80} - 100k
Followers growth rate	4 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Followers following ratio	2 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Tweet frequency	10 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Topic statistics					
	5	10	15	20	25
Topic % in tweets	0 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Topic % in words	0 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Maximum range score	20	40	60	80	100

TABLE III: Twitter Micro Topic Influencer selection ranking

General Statistics					
	Points				
Categories	2.5	5	7.5	10	12.5
Followers count	5k - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	P_{80} - 100k
Followers growth rate	2 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Followers following ratio	2 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Topic statistics					
	2.5	5	7.5	10	12.5
Topic % in captions	0 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Topic % in cap words	0 - P_{20}	P_{20} - P_{40}	P_{40} - P_{60}	P_{60} -P_{80}	$>P_{80}$
Topic % in pictures	0 - P_{92}	P_{92} - P_{94}	P_{94} - P_{96}	P_{96} -P_{98}	$>P_{98}$
Topic % in pictures words	0 - P_{92}	P_{92} - P_{94}	P_{94} - P_{96}	P_{96} -P_{98}	$>P_{98}$
Maximum range score	20	40	60	80	100

TABLE IV: Instagram Micro Topic Influencer selection ranking

D. *Multi Input Micro-Influencers Classifier*

The classification of micro influencers is a supervised binary classification problem. We test six different models to measure which one was the most suitable given our input scores, as described in Section III and the expected output. We perform the experiment with XGBoost, Random Forest Classifier, Support Vector Classifier (SVC), Multi-layer perceptron (MLP), Logistic Regression and Stochastic Gradient Descent (SGD). They produce one label to define if a user is a micro influencer in the broad definition and another one to define if a user is a micro influencer for a specific topic given as input to the model. The summary of this pipeline is presented in Figure 3. The fine tuning parameters of these algorithms are selected by the Grid Search CV of the Scikit Learn library [3]. The (eXtreme Gradient Boosting) XGBoost model obtains the best results. The Gradient Boosting prediction model is an ensemble of many weak classification models. It is defined as a stage-wise model. It concedes the optimization of an arbitrary differentiable loss function, allowing an improved tuning based on the selected problem to be solved. XGBoost provides a parallel tree boosting system. One of the main drawbacks of XGBoost though is the low interpretability of the generated results. XGBoost trains a huge variety of models on different subsets of the training dataset and eventually selects the best performing one. Some important features of the XGBoost algorithm are the parallelization (training with multiple CPU cores), the regularization (penalties mechanisms to avoid over-fitting), non-linearity, cross-validation and scalability to deal with very large quantities of data without losing performance. Thanks to this model we are able to answer the RQ2 selecting an XGBoost model that handles the multi input given by our scores retrieval on our dataset.

IV. Experimental Results

We use Twitter and Instagram social media platforms to retrieve user writings and visual posts. We select users having recently posted about topics we selected for the experiment. Users are filtered and labeled following the thresholds of Table I. For each user we download his last 200 posts. In the Instagram case, the images are translated into captions and stored as a textual post. The entire written production of an user receives the text cleaning procedure described in Section III. The initial dataset is then divided into 80% training set and 20% test set balancing the number of micro and not micro

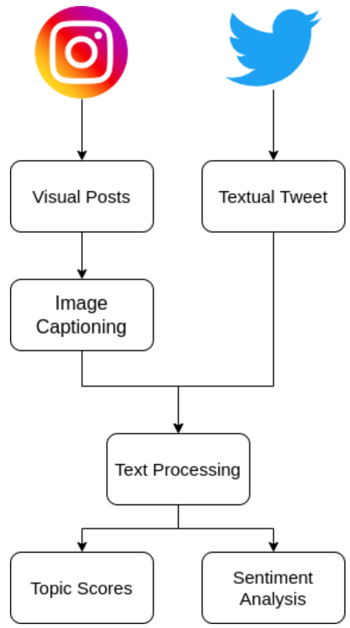

Fig. 2: Data processing pipeline describing the collection of tweets and Instagram posts and the next features extraction. The pipeline for visual posts includes the image captioning step as depicted in Figure 1, while for textual posts this step is not needed.

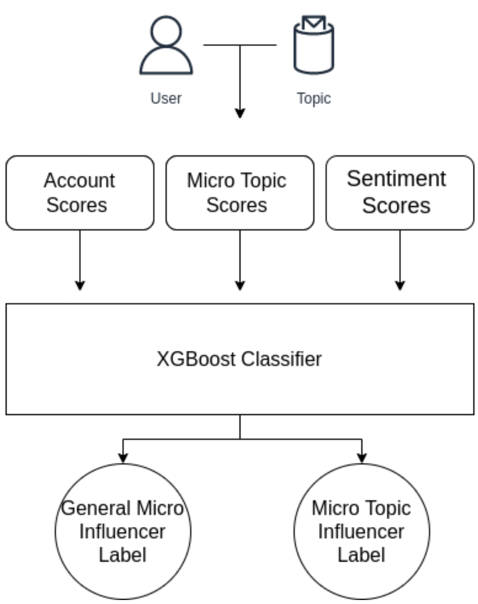

Fig. 3: Multi input micro influencer classifier pipeline. This is the general schema for data retrieval from a social media platform, score computation, classification model selection and final labels definition for both general micro influencers and micro topic influencers.

influencers for the general case and the topic specific one. The lists of features collected for both classifier are the following:

Twitter features: followers count, age, followers growth rate, followers following ratio, tweet frequency, topic % in tweets, topic%in words, positive sentiment, neutral sentiment, negative sentiment.

Instagram features: followers count, followers growth rate, followers following ratio, topic % in captions, topic % in cap words, topic % in pictures, topic % in pictures words, positive sentiment captions, neutral sentiment captions, negative sentiment captions, positive sentiment captions words, neutral sentiment captions words, negative sentiment captions words.

While the metrics adopted to assess the effectiveness of different models are recall, precision and accuracy metrics as described in Equation 10, 11, 12. In this work, we classify a user as a micro-influencer or non-micro-influencer. In the second of our scenarios, we classify a user as a micro topic influencer or as a non-micro-topic influencer. In both cases we make a binary classification. We measure the validity of our model with recall, precision, f1-scores and accuracy that are described in the following. We define true positive, true negative, false positive and false negative in accordance with the following descriptions. *tp* (true positive) is a user correctly classified as a micro-influencer (label 1). *tn* (true negative) a user is labeled as a non-micro-influencer correctly (label 0). *fp* (false positive) is a non-micro-influencer user classified as a micro-influencer. *fn* (false negative) is a micro-influencer classified as non-micro-influencer.

$$recall = \frac{tp}{tp + fn} \tag{10}$$

$$precision = \frac{tp}{tp + fp} \tag{11}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{12}$$

The results obtained in the validation process are always the highest for the XGBoost model. Also the comparison with a deep neural network that applies BERT for sequence classification and then maps all the sequence embeddings into a user embedding finally max pooled to derive a classifier with a final neuron layer to classify micro influencer is less performing than XGBoost. These results are presented in Tables V and VI.

	Twitter General Micro Influencer Classification Metrics			
Model	Accuracy	Precision	Recall	f1-score
XGBoost	0.99	0.99	0.99	0.99
BERT-based	0.80	0.81	0.80	0.80
SVM	0.73	0.74	0.73	0.73
MLP	0.90	0.90	0.90	0.90
LR	0.77	0.77	0.77	0.77
SGD	0.52	0.31	0.52	0.39

TABLE V: Comparison metrics between different models tested. Twitter general influencer classification results.

In the Instagram case, as described by Tables VII and VIII, XGBoost outperforms all the other models in all the four validation metrics computed. It is interesting to notice that it also beats a CNN-based model that performs a convolution over sequence of text and over images to extract features

Twitter Topic Micro Influencer Classification Metrics				
Model	Accuracy	Precision	Recall	f1-score
XGBoost	0.93	0.93	0.93	0.93
BERT-based	0.79	0.81	0.80	0.80
SVM	0.88	0.90	0.88	0.89
MLP	0.63	0.62	0.63	0.62
LR	0.88	0.90	0.88	0.89
SGD	0.65	0.69	0.65	0.66

TABLE VI: Comparison metrics between different models tested. Twitter topic influencer classification results.

embedding then fed into a single layer neural network to compute the final label.

Instagram General Micro Influencer Classification Metrics				
Model	Accuracy	Precision	Recall	f1-score
XGBoost	0.98	0.98	0.98	0.98
CNN-based	0.77	0.75	0.76	0.76
SVM	0.60	0.60	0.60	0.60
MLP	0.65	0.65	0.65	0.65
LR	0.60	0.60	0.60	0.60
SGD	0.55	0.55	0.55	0.55

TABLE VII: Comparison metrics between different models tested. Instagram general influencer classification results.

Instagram Topic Micro Influencer Classification Metrics				
Model	Accuracy	Precision	Recall	f1-score
XGBoost	0.98	0.98	0.98	0.98
CNN-based	0.73	0.74	0.73	0.73
SVM	0.62	0.62	0.62	0.61
MLP	0.63	0.63	0.63	0.63
LR	0.58	0.58	0.58	0.58
SGD	0.50	0.50	0.50	0.50

TABLE VIII: Comparison metrics between different models tested. Instagram topic influencer classification results.

Thanks to these results we are also able to answer positively to the third and last research question RQ3.

V. Discussion

There are different approaches to assess influencers and while many of them are also applicable to the case of micro influencers others are not. We focus on the latter case adopting a multi input approach to consider a wide set of input features retrievable directly from initial post or after a step of image or text processing. We compute metrics related to the user social media account paired with features regarding specific topics. This kind of approach better matches the requirements of private companies and public institutions to find and classify micro influencers in their areas of interest. Even if the pipeline we developed is straight forward with the combination of already existing models stacked, data retrieval is the main difficulty. In fact, Twitter already offers a developer API platform to collect data even if the timeout makes the collection really slow, while Instagram is more restrictive and the use of a library as Instaloader to collect social media posts is not the best performing option anyway it is the only one exploitable. We are also aware that images translated into textual caption may lose important visual features. Anyway it is difficult to understand how to match these features with the final objective of micro influencer classification, this way still needs to be further explored. This automatic procedure is useful both to find micro influencer and to rank him among other users with similar characteristics, at the same time though, the gold standard dataset has been built from scratch according to business rules, it is advisable in future to perform a user supervision on the assigned label after the verification of effective capabilities of micro influencing of these users.

VI. Conclusion

This work presents a new framework for Twitter and Instagram to collect and classify micro influencers in general and in specific topics cases. We proved through validation metrics that XGBoost is the most effective model to perform this task receiving the features collected with an overall accuracy above 0.93. The process can be reproduced to expand the dataset and to explore different topics. This work is mainly focused on text for Twitter and on text and images for Instagram. An evolution to better understand users' communication skills may be based on the visual information study, performing video translations to text for both Instagram Reels and Instagram stories. In addition, some of the metrics adopted in this work counts the presence of the topic word inside the entire obtained text. An extension of this approach can involve the adoption of topic detection algorithms to capture even synonyms or periphrasis. Even if there are many directions still to be explored, we proved the effectiveness of our model and we created a new dataset to be exploited for further analysis.

References

[1] help.twitter.com/en/managing-your-account/about-twitter-verified-accounts
[2] cocodataset.org
[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.h
[4] Rahmanov. F., Mursalov. M., and Rosokhata, A. (2021). Consumer Behavior in Digital Era: Impact of COVID19. Marketing and Management of Innovations, 2, 243-251. http://doi.org/10.21272/mmi.2021.2-20
[5] Purbasari, Ratih, Zaenal Muttaqin, and Deasy Silvya Sari. "Digital entrepreneurship in pandemic Covid 19 Era: The digital entrepreneurial ecosystem framework." Review of integrative business and economics research 10 (2021): 114-135.
[6] Haudi, H., et al. "The role of e-marketing and e-CRM on e-loyalty of Indonesian companies during Covid pandemic and digital era." Uncertain Supply Chain Management 10.1 (2022): 217-224.
[7] Francisco, Evelina, et al. "Impact of the COVID-19 Pandemic on Instagram and Influencer Marketing." International Journal of Marketing Studies 13.2 (2021): 20-35.
[8] Karen Freberg, Kristin Graham, Karen McGaughey, and Laura A. Freberg. 2011. Who are the social media influencers? A study of public perceptions of personality. Public Relations Review 37, 1 (2011), 90 – 92.

[9] Maureen Brewster and Jewon Lyu. "Exploring the parasocial impact of nano, micro and macro influencers." In: International Textile and Apparel Associa- tion Annual Conference Proceedings. Vol. 77. No. 1. Iowa State University Digital Press, July 2020.
[10] Benyamin Bashari and Ehsan Fazl-Ersi. "Influential post identification on Instagram through caption and hashtag analysis". In: Measurement and Control 53.3-4 (2020), pp. 409–41
[11] Iris Roelens, Philippe Baecke, and Dries F Benoit. "Identifying influencers in a social network: The value of real referral data". In: Decision Support Systems 91 (2016), pp. 25–36
[12] Yun-Bei Zhuang, Zhi-Hong Li, and Yun-Jing Zhuang. "Identification of influencers in online social networks: measuring influence considering multi- dimensional factors exploration". In: Heliyon 7.4 (2021), e06472
[13] Tabassum, Shazia, et al. "Social network analysis: An overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.5 (2018): e1256.
[14] Albery, Gregory F., et al. "Unifying spatial and social network analysis in disease ecology." Journal of Animal Ecology 90.1 (2021): 45-61.
[15] Dahesh, Mehran Badin, et al. "Reviewing the intellectual structure and evolution of the innovation systems approach: A social network analysis." Technology in Society 63 (2020): 101399.
[16] Hung, Man, et al. "Social network analysis of COVID-19 sentiments: Application of artificial intelligence." Journal of medical Internet research 22.8 (2020): e22590.
[17] Kim, Jooho, and Makarand Hastak. "Social network analysis: Characteristics of online social networks after a disaster." International journal of information management 38.1 (2018): 86-96.
[18] Yum, Seungil. "Social network analysis for coronavirus (COVID-19) in the United States." Social Science Quarterly 101.4 (2020): 1642-1647.
[19] Fabián Riquelme and Pablo González-Cantergiani. "Measuring user influence on Twitter: A survey". In: Information processing & management 52.5 (2016), pp. 949–975
[20] Stefan Räbiger and Myra Spiliopoulou. "A framework for validating the merit of properties that predict the influence of a twitter user". In: Expert Systems with Applications 42.5 (2015), pp. 2824–283
[21] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. "Vital nodes identification in complex networks". In: Physics Reports 650 (2016), pp. 1–63
[22] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?" In: Proceedings of the 19th international conference on World wide web. 2010, pp. 591–600
[23] Xin Chen. "Critical nodes identification in complex systems." In: Complex & Intelligent Systems 1.1 (2015), pp. 37–56
[24] Changjun Fan, Li Zeng, Yizhou Sun, and Yang-Yu Liu. "Finding key players in complex networks through deep reinforcement learning." In: Nature machine intelligence 2.6 (2020), pp. 317–32
[25] Tian Gan, Shaokun Wang, Meng Liu, Xuemeng Song, Yiyang Yao, and Liqiang Nie. "Seeking micro-influencers for brand promotion." In: Proceedings of the 27th ACM International Conference on Multimedia. 2019, pp. 1933–1941
[26] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. "Hadamard product for low-rank bilinear pooling." In: arXiv preprint arXiv:1610.04325
[27] Cheng Zheng, Qin Zhang, Sean Young, and Wei Wang. "On-demand Influencer Discovery on Social Media." In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020, pp. 2337–2340
[28] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. "Scalable and generalizable social bot detection through data selection". In: Proceedings of the AAAI conference on artificial intelligence. Vol. 34. 01. 2020, pp. 1096– 1103
[29] Ron Mokady, Amir Hertz, and Amit H Bermano. "Clipcap: Clip prefix for image captioning". In: arXiv preprint arXiv:2111.09734 (2021)
[30] Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya. "Learning Transferable Visual Models From Natural Language Supervision". In: International Conference on Machine Learning. PMLR. 2021, pp. 8748–876
[31] Barbieri, Francesco and Camacho-Collados, Jose and Espinosa Anke, Luis and Neves, Leonardo. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In Findings of the Association for Computational Linguistics: EMNLP 2020.

Simone Leonardi received the B.Sc. and M.Sc. degrees in computer engineering from the Poltecnico di Torino, where he is currently pursuing the Ph.D. degree in computer and control engineering. He is also a member with the Software Engineering Group. His current research interests include natural language processing and personality analysis applied at social media analysis and software testing.

Luca Ardito (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from the Politecnico di Torino. He is currently an Assistant Professor with the Department of Control and Computer Engineering, Politecnico di Torino, where he is also working with the Software Engineering research group. His current research interests include mobile development and testing, green software, new programming language analysis, and empirical software engineering methodologies.