УДК 004.912 : 81'33

Anton O. Kontsevoi, Oleh V. Bisikalo

INFORMATION SYSTEM OF DEFINITION FOR INDICATOR CHARACTERISTICS OF PROFILES OF SOCIAL NETWORKS PARTICIPANTS

Vinnytsia National Technical University

Анотація

Розроблена та апробована на реальних прикладах та даних система, яка реалізує процес автоматизованого визначення індикаторних характеристик профілів учасників соціальної мережі, а також визначення реакції учасників соціальної мережі на певні події. Розроблена система показала, що запропонований підхід значно покращив точність та обсяг характеристик профілів користувачів всередині соціальної мережі Twitter, використовуючи підходи, що поєднують різні технології обробки природних мов.

Ключові слова: андроїд, мобільний додаток, синтаксичний аналіз,, класифікація користувачів, соціальна мережа.

Abstract

A system that implements the process of automated determination of indicator characteristics of social network profiles, as well as determining the response of social network participants to certain events has been developed and tested on the real-world examples and data. The developed system showed that proposed approach has substantially improved the accuracy and the amount of the characteristics of user profiles inside the Twitter social network by using the approaches that combine various natural language processing technologies and approaches.

Keywords: android, mobile application, syntax analysis, classification of the users, social network.

Introduction

The relevance of this work is connected with the growth of the Internet and the continuous increase in the amount of data in it, as well as the high popularity of social networks. In this regard, it became necessary to automatically identify and process large amounts of information about participants in social networks. Because existing methods, algorithms and programs do not achieve the desired result, new methods need to be developed to solve these kinds of problems. Examples of such tasks are the task of determining the indicator characteristics of a social network participant profile. Effectively addressing these challenges can help you reach specific groups of people, their interests and hobbies, which will be useful in areas such as SEO and marketing to improve the quality of targeting, trending in the current market, and more.

Social networks are a phenomenon today. The benefits of using social networks are that you can connect with your friends quickly and easily, usually in the form of objects such as posts, pictures, videos and texts. Another feature is networking: friends, colleagues, and family.

Analysis of social networks (related to network theory) has become the main method of research in modern sociology, anthropology, geography, social psychology, informatics and research organizations, and a common topic for research and discussion. Studies in several academic fields have shown that social networks operate at many levels, ranging from families to entire nations, and play an important role in how problems are solved, organizations operate, and succeed in achieving individuals' own goals [1, 2].

Aggregating information from publicly available profiles is very useful for specific purposes, such as building a marketing strategy and identifying groups of individuals associated with banned organizations [3-5].

This raises the problem of collecting, analyzing, and processing large amounts of data about social network users.

The purpose of the work is to increase the efficiency of the system of automated determination of indicator characteristics of profiles of participants of social networks and the process of determining the reaction of users to real-time events.

The object of study is the process of automated determination of indicator characteristics of profiles of participants in social networks, as well as the process of determining the response of users to certain events in real time.

The subject of the study are methods and tools for analyzing the profiles of social network participants, as well as instrumental methods for determining the reaction of social network participants to real-time events on mobile devices.

Scientific novelty:

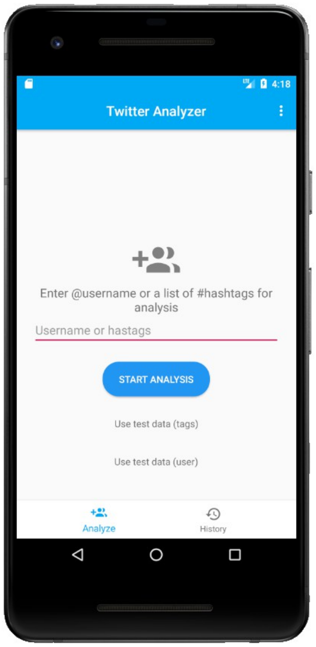
1. For the first time, a new approach to determining the indicator characteristics of profiles of social network participants by analyzing their profiles and their reactions to network events using syntactic methods is proposed and it allows to increase the number of determined characteristics in comparison with existing applications that only show characteristics provided by the social network.

2. The proposed approach allows, unlike existing ones, to analyze user profiles for information about groups of people on the network, as well as their reaction to an event or news on the network on a mobile device by using specialized libraries and combining natural language processing methods in order to increase the accuracy of the results compared to already existing applications.

The practical significance of the results obtained. The research performed in this paper solves the problem of analyzing and determining the indicator characteristics of social network participants' profiles (sample profiles, or specific groups of people) and their reactions to network events.

1 Using the application

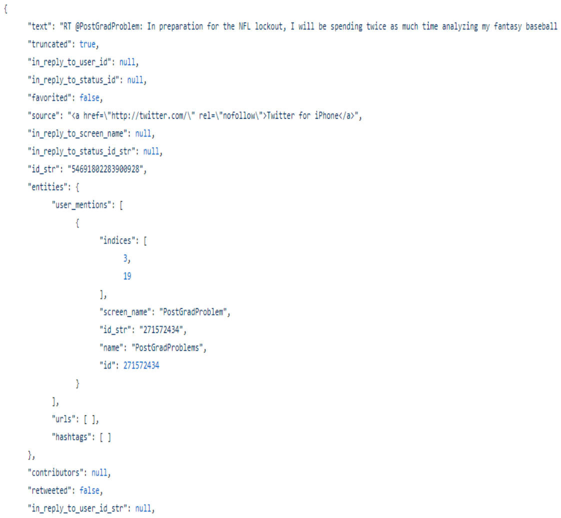
The main screen of the application is shown in Figure 1.



Enter @username or a list of #hashtags for analysis

Figure 1 - The main screen of the application

When entering a username or hashtag, the system first calls the Twitter API. The corresponding module of the program sends the locks to the server using the previously specified access parameters, as well as the user's request. The server returns data in JSON format (Figure 2).



```
{
    "text": "RT @PostGradProblem: In preparation for the NFL lockout, I will be spending twice as much time analyzing my fantasy baseball
    "truncated": true,
    "in_reply_to_user_id": null,
    "in_reply_to_status_id": null,
    "favorited": false,
    "source": "<a href=\"http://twitter.com/\" rel=\"nofollow\">Twitter for iPhone</a>",
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id_str": null,
    "id_str": "54691802283900928",
    "entities": {
        "user_mentions": [
            {
                "indices": [
                    3,
                    19
                ],
                "screen_name": "PostGradProblem",
                "id_str": "271572434",
                "name": "PostGradProblems",
                "id": 271572434
            }
        ],
        "urls": [ ],
        "hashtags": [ ]
    },
    "contributors": null,
    "retweeted": false,
    "in_reply_to_user_id_str": null,
```

Figure 2 - Example JSON response from the Twitter API

At the stage of parsing each reaction of the user, the text of the reaction itself is extracted and divided into tokens, after which these tokens are added to the list of tokens. The list of tokens is necessary for further text processing, since many other text processing methods are needed in preliminary tokenization.

Next, the program determines the names of people, locations and dates that are in the text of tweets using the Named Entities Recognition API library Apache OpenNLP.

Then, using the Chunker API, the program selects keywords and phrases (sets of words combined in meaning and grammatically) in tweet texts. For this function to work, the program needs a list of tokens, as well as the result of the POS Tagger API, whose task is to determine the parts of speech for each word in the incoming text. This will allow the Chunker API to identify words and phrases related in meaning in the text, which in the end result will help determine the reaction of a particular user to an event on the network, as well as to the general reaction of users.

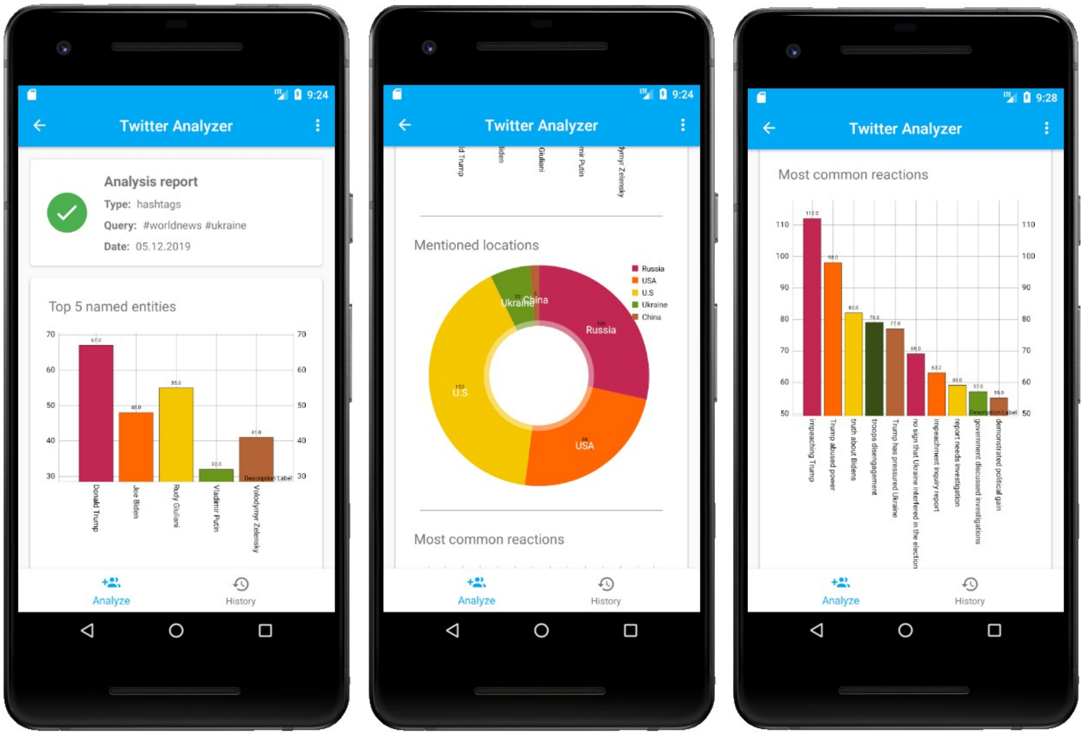
In the case of analyzing the profile of a Twitter participant, the Language Detector API is additionally applied, which, based on the model embedded in it, recognizes the language of the text and presents a list of languages that it was possible to recognize together with the probability coefficient that this language is the language of the text.

2 Analysis of the test results

To study and analyze the reactions of users to events, the responses of users on the Twitter network regarding events in the network united by the tag “#ukraine” were selected as input data for analysis, which means that if a given tag or word is present on a tweet, it will fall into search results.

To conduct the study, 1,500 user responses in English were collected, since the recognition model works with English speech. This figure is small for real analysis, since it covers only a small percentage of network users' reactions, but it is dictated by the limitations of the Twitter API. This limitation can be circumvented by sending requests at fixed intervals, after which the request counter is reset to zero and new requests can be sent. However, this approach will not work in the context of a mobile application, since even if it will collect data on demand in the background for a certain period of time, there is a risk that the system will stop the background processes. For the test needs, this number of tweets is enough, because the goal is not to collect the most accurate statistics, but to demonstrate an approach to solving the problem.

Figure 3 shows the results of an analysis of user reactions to events on the network.



Analysis report

Type: hashtags

Query: #worldnews #ukraine

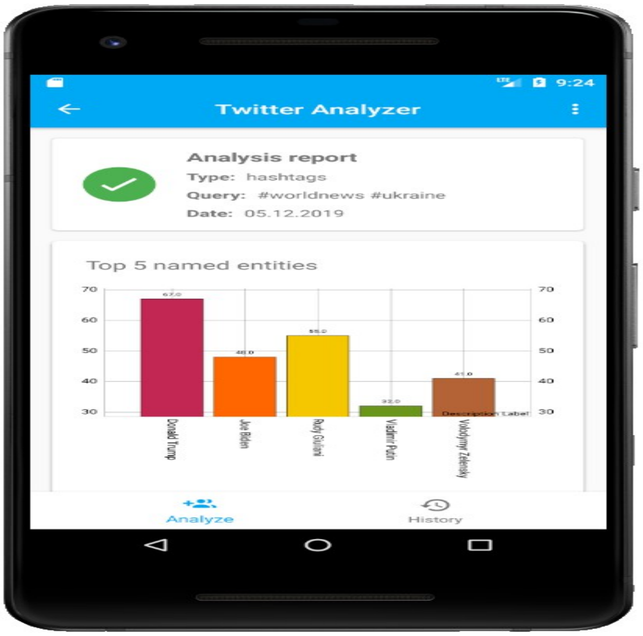
Date: 05.12.2019

Figure 3 - User response to events on the network

The user can interact with the graphs, zooming in and scrolling through them to see all the data.

In the course of processing the results of the analysis of user reactions, the program identified the 5 most common names in the text of tweets, they turned out to be: "Donald Trump", "Joe Biden", "Rudy Giuliani", "Vladimir Putin", "Volodymyr Zelensky" (Figure 4).

With small amounts of data (1-5 thousand tweets), there is no need to display more results, since often the names of other people are rarely found in tweets, more often they are in the @nickname format.



Analysis report

Type: hashtags

Query: #worldnews #ukraine

Date: 05.12.2019

Figure 4 - Diagram of the most frequently encountered names in tweet texts

Figure 5 shows a pie chart with the mentioned locations.

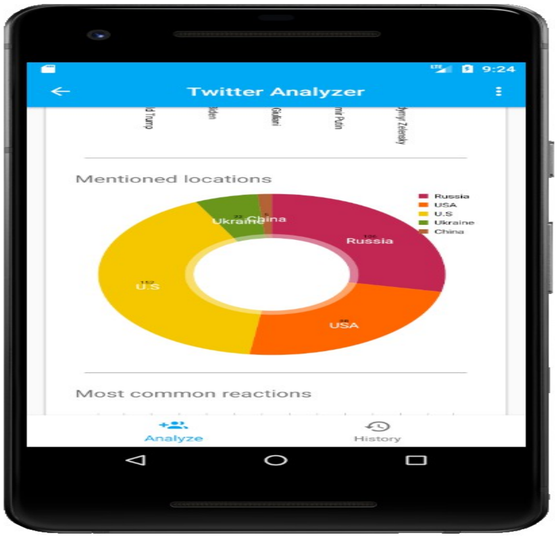


Figure 5 – Most mentioned locations

Named Entity Recognition API was able to highlight the names of locations in the text, in this case, the names of the most mentioned countries. It is worth noting that, despite the high accuracy of the results obtained, the model may not determine the names of little-known cities, which may be a problem when analyzing more “local” events in the network.

Figure 6 shows a diagram with the most common keywords and phrases in the texts of reactions to news on the network.

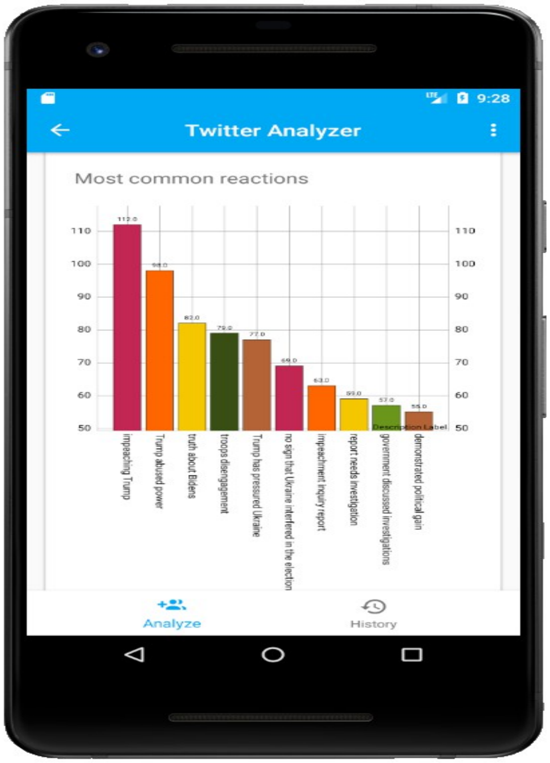
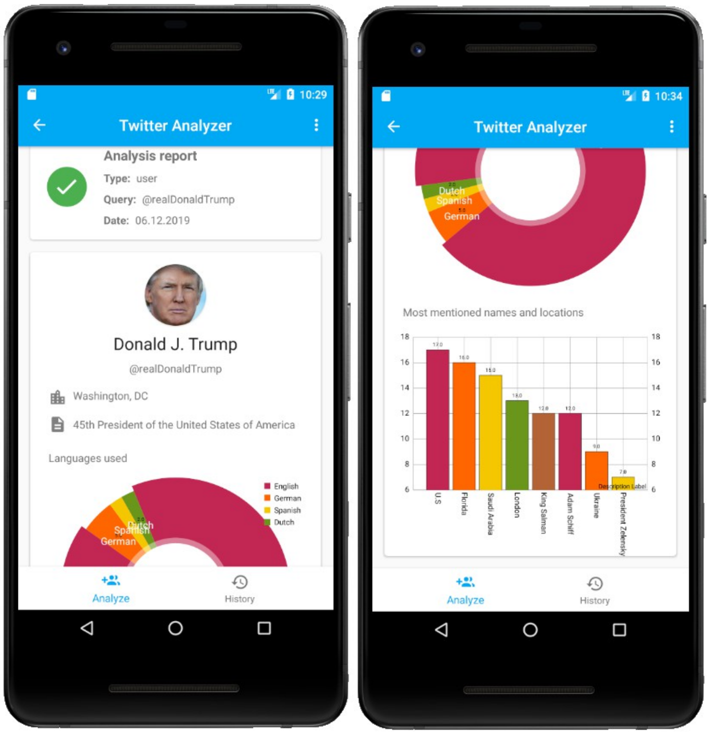


Figure 6 - The most common keywords and phrases

It is worth noting that the approach using Tokenization, POS Tagging and Chunking API gives much greater accuracy and greater semantic load embedded in the final results compared to using Tokenization and Tuples (for generating pairs of keywords in the text). The data approach works well in English, since it has a strict word order in a sentence, therefore there is a high probability of finding matches among the received words and phrases.

To study and analyze the profile of a member of a social network, the profile of the current US president was chosen, as well as his tweets in the amount of 2500 pieces. This figure is small for real analysis, since it covers only a small number of recent user tweets and is dictated by the limitations of the Twitter API.

Figure 7 shows the results of the analysis of the profile of a member of the Twitter network.



10:29

Twitter Analyzer

Analysis report

Type: user

Query: @realDonaldTrump

Date: 06.12.2019

Donald J. Trump

@realDonaldTrump

Washington, DC

45th President of the United States of America

Languages used

Figure 7 - Results of the analysis of the profile of a member of the social network Twitter

The analysis results show that the system recognized 4 languages that were supposedly used when writing tweets by a user. The result with naiballs was the accuracy of the English language, which is true. The remaining languages were recognized, since the text could contain names, names or locations that are written similar to these languages, as well as due to errors in the recognition model (Figure 8).

Figure 8 shows the pie diagram with the languages detected in the user tweets.

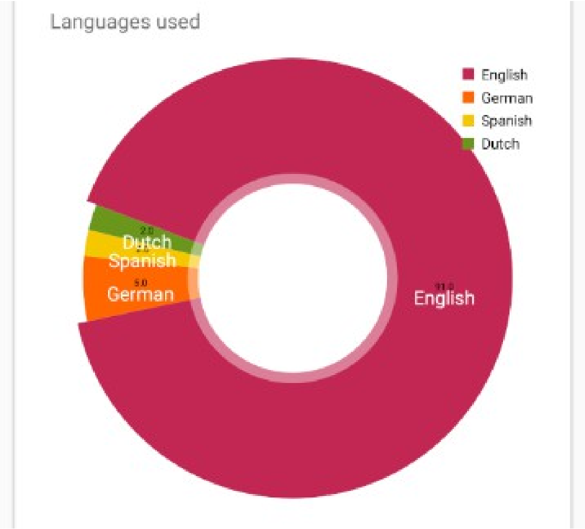


Figure 8 – The languages detected in the user tweets

We also analyzed the most mentioned places and personalities on the same principle as when analyzing user reactions to events on the network. These results were put on one chart in order to determine the dependence between the name and location. With this approach, between related pairs of names and locations, the difference in the frequency of references will be insignificant. However, in the case of this user and data set, this dependence was not detected.

Conclusions

Summing up, it can be argued that the created program successfully fulfills the tasks of analyzing the profiles of members of a social network, as well as determining the response of users to events in it.

In the course of the work, a new co-ordinated approach to solving the above problems was proposed, implemented using the Apache OpenNLP library, which made it possible to increase the accuracy of the received results, which was confirmed by the testing results of the developed program.

REFERENCES

1. Концевой А. О. Автоматизоване визначення індикаторних хаарктеристик профілів учасників соціальних мереж / А. О. Концевой // Матеріали доповідей XLVII науково-технічної конференції підрозділів Вінницького національного технічного університету, 21–23 березня. – Вінниця : ВНТУ, 2018.
2. Bisikalo O. Modeling the phenomenological concepts for figurative processing of natural-language constructions / Oleg Bisikalo, Yuriy Ivanov, Vladyslava Sholota // Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference. – Kharkiv, Ukraine, April 18-19, 2019. – Pp. 1-11.
3. Wasserman S. Social Network Analysis: Methods and Applications / S. Wasserman, K. Faust. — Cambridge: Cambridge University Press, 1994 — 857 p.
4. Martino F. Social Network Analysis: A brief theoretical review and further perspectives in the study of Information Technology / F. Martino, A. Spoto // Psychology Journal. — 2006. — Vol. 4, No 1. — P. 53—86.
5. Butts C. T. Social network analysis. A methodological introduction // C. T. Butts // Asian Journal of Social Psychology. — 2008. — Vol. 1

Концевой Антон Олександрович – студент групи 3АКІТ-18м, Вінницький національний технічний університет, м. Вінниця, e-mail: anton.96k@gmail.com

Науковий керівник: ***Бісікало Олег Володимирович*** – д-р техн. наук, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Kontsevoi Anton Oleksandrovych – student of the Faculty of Automation, Electronics and Computer Control Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: anton.96k@gmail.com

Supervisor: ***Bisikalo Oleh V.*** – Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua