UDC 004.855.5
https://doi.org/10.31891/csit-2022-4-4

TARAS RUDNYK, OLEG CHERTOV
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

FORECASTING THE RESULTS OF THE PRESIDENTIAL ELECTIONS IN FRANCE BASED ON TWITTER DATA

This paper presents the study to collect, store and analyze data from Twitter to forecast French presidential election results, compared to sociological polls. The first and probably the most important step of the research is to collect, store and clean data, the whole result depends on the amount and quality of data. In the next step of research, datasets are analyzed. Lastly, complete report and visualizations are provided. In the study, we propose modern technics, mathematical algorithms, and machine learning approaches to analyze big amounts of data from the Twitter social network in order to forecast the 2022 French presidential election results. The determined outcome is compared with sociological polls and the real results of elections.

In the conducted research modern types of media are compared to select the best one for election prediction. Selected Twitter social network as the one with the most appropriate data and availability to download big amounts of useful information. The approach based on the usage of Python programming language, Selenium browser emulation and MongoDB database was used to collect, store and clean data about the main French election candidates – Emmanuel Macron and Marine Le Pen. The research was made from August 2021 until the election itself in April 2022. The determined outcome is compared with sociological polls and the results of elections and showed that analysis of social network data could be a good alternative to traditional sociological polls as it shows the same trends month by month and well predicted the win of Emmanuel Macron in elections. Moreover, the proposed approach has its benefits compared to sociological polls such as always being fresh, and close to real-time information, the price of research is much lower and could be reused for the next parliamentary or presidential elections with a small modification.

The research could be extended and adapted for other countries. Currently, the proposed algorithms and mathematical models showed good results in the French and Ukraine elections. It works well with English, French, Ukrainian and Russian languages. This allows us to claim that it will also work fine with other Latin or Cyrillic alphabets but for Asian or Arabic languages more research would be needed. Twitter is a good choice for European and American countries. In the future, other social networks should be considered for the countries in which it is not so popular.

Keywords: political rating, sociological poll, Twitter, Python, Selenium, data collection, machine learning, natural language processing.

ТАРАС РУДНИК, ОЛЕГ ЧЕРТОВ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ВИБОРІВ У ФРАНЦІЇ НА ОСНОВІ ДАНИХ З TWITTER

У цій статті представлено дослідження збору, зберігання та аналізу даних із Twitter для прогнозування результатів президентських виборів у Франції у порівнянні з соціологічними опитуваннями. Першим і, мабуть, найважливішим кроком дослідження є збір, зберігання та очищення даних, оскільки весь результат залежить від кількості та якості даних. На наступному етапі дослідження проводиться аналіз наборів даних. В кінці надається повний звіт і візуалізація отриманих результатів. У дослідженні ми пропонуємо сучасну техніку, математичні алгоритми та підходи машинного навчання для аналізу великих обсягів даних із соціальної мережі Twitter, щоб спрогнозувати результати президентських виборів у Франції 2022 року. Отриманий результат порівнюється із даними соціологічних опитувань та фактичними результатами виборів.

У проведеному дослідженні порівнюються сучасні види медіа, щоб вибрати найкраще для прогнозування виборів. Вибрана соціальна мережа Twitter як така, що має найбільш відповідні дані та доступність для завантаження великої кількості корисної інформації. Підхід, заснований на використанні мови програмування Python, емуляції браузера Selenium і бази даних MongoDB, використовувався для збору, зберігання і очищення даних про головних кандидатів на виборах у Франції – Еммануеля Макрона і Марін Ле Пен. Дослідження проводилося з серпня 2021 року до самих виборів у квітні 2022 року. Визначений результат порівнюється з соціологічними опитуваннями та результатами виборів і показує, що аналіз даних соціальних мереж може бути хорошою альтернативою традиційним соціологічним опитуванням, оскільки він показує ті самі тенденції місяць за місяцем і добре передбачив перемогу Еммануеля Макрона на виборах. Більше того, запропонований підхід має свої переваги порівняно з соціологічними опитуваннями, такі як: завжди свіжа та наближена до реального часу інформація, ціна дослідження значно нижча та може бути повторно використана для наступних парламентських чи президентських виборів із невеликою модифікацією.

Дослідження можна розширити та адаптувати для інших країн. Наразі запропоновані алгоритми та математичні моделі показали хороші результати на виборах у Франції та Україні. Добре працюють з англійською, французькою, українською та російською мовами. Це дозволяє нам стверджувати, що вони також добре працюватимуть з іншими латинськими чи кириличними алфавітами, але для азіатських чи арабських мов потрібні додаткові дослідження. Twitter є хорошим вибором для країн Європи та Америки.

Ключові слова: політичний рейтинг, соціологічне опитування, Twitter, Python, Selenium, збір даних, машинне навчання, обробка природної мови.

Introduction

Nowadays there is a lot of data on the Internet. The modern world allows people to exchange opinions around the world about the different variety of topics. Some sources, like tv channels or newspapers, carefully prepare well-structured information and share it on social media. It is a job that is done by much fewer people than by Internet users. Almost everyone nowadays has an account on one or even all social networks such as Twitter, Facebook,

Instagram, or TikTok. Compared to social media in own pages, there are much more users. The information mostly is not well structured and could contain mistakes.

Everyone who analyzes social networks faces one or all listed below challenges:

- How to collect data?
- How to store big amounts of data?
- How to clean and transform data into a dataset?
- How to analyze data?
- How to visualize results?
- How to make a conclusion and create a report?

This research presents an approach that answers the listed above challenges that we faced while we were trying to forecast the results of the 2022 presidential elections in France based on data collected from Twitter. The elections in France were one of the hottest political topics at the beginning of the year 2022. Two candidates – Emmanuel Macron and Maria Le Pen were close to winning the election accordingly to social polls. Moreover, at some period time closer to the election date the rating of Maria Le Pen was growing, and at the same time rating of Emmanuel Macron was falling which was a pretty interesting situation to consider if we can achieve the same result as social polls using data collected from Twitter social network and predict the result of elections?

Related works

Data collection is a popular task in modern research. The lack of publicly available datasets motivates researchers to collect their own data which is often not an easy task. In [1] Yuji Roh, Geon Heo and Steven Euijong Whang conducted a survey on how data could be collected, cleaned, and labeled for machine learning. Accordingly, to the research, each step could have a modification, for example, labeling could be done not only by people but also using data programming or fact extraction. The gaps in data could be covered by generating synthetic data. From a machine learning perspective, results could be enhanced in different ways for example by improving the model or improving data. Before collection, it is important to understand which types of data there are on the chosen platform and which of them, we need. In [2] Hai Liang et al divide data into three types: content, behavior and network structure. For each type of data, authors suggest different approaches to harvest it, for example, content and behavior data – random selection, network structure – probability (or uniform) sampling, breadth-first search (BFS) sampling, and random walk (RW) sampling. For web, harvesting authors suggest using APIs or web scraping. The first type is also used in our research when it is possible but for scraping social networks traditional approach via HTML parsing is not working, therefore, browser emulation via Selenium web driver is used.

Nowadays the amount of textual data across the Internet is extremely large. The task of structuring and analyzing such data is impossible through manual human work. Much more practical to develop and use text analyzing and mining technics to automate these processes. Perhaps Noah Chomsky was the first linguist that started syntactic theories by introducing in 1957 syntactic structures. He defined a set of rules based on universal grammar. In 1965 [3] Noah Chomsky categorized syntactic theories into speech recognition (Higher Level) and natural language (Lower Level). Later in 1967 [4], Charles Hockett found some drawbacks in Noah Chomsky's study, the most important part was that in his study language is a well-defined, stable structure without mistakes which was possible only in rare, idealized conditions. Nowadays in an analysis of Internet sources such as social networks we can see that messages contain a lot of mistakes, aren't well structured and can be in different languages even with a mix of words from 2 or more languages. Text mining and natural language processing technics found their implication in many fields over the years. For example, in [5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao showed that for decade from 2000 to 2011 years text mining techniques were applied to a variety of fields such as academics, industry, web applications and others.

There is a lot of research on social networks text analysis. For example, in [6] authors use modern natural language processing technics to identify fake news in social networks. The research described that usually text analysis starts from tokenization, punctuation, special characters removal, stopwords removal, spell checking, named entity recognition and stemming. In the next stage authors used vector representation of words because, for the calculation of machine learning models, data that could be operated mathematically was needed. TF-IDF [7] algorithm was used to analyze the texts further. Terms frequency (TF) – how many times this word appears in a document (sentence), Inverse Document Frequency (IDF) is the natural logarithm of the total number of documents divided by the total number of documents that contain this certain word + 1. All listed above technics are used in our research. For our dataset, we tried TF-IDF, Word2Vec [8] and Doc2Vec [9] technics and chose the best-performing one for us – Word2Vec.

In our [10] most recent study we conducted research on Ukrainian presidential elections, proposed several algorithms to determine the rating of politicians, detect dates when news affected ratings the most and identify specific news which influenced grows or fall of the rating. Experimental results conducted on Ukrainian President Volodymyr Zelenskyy page on Twitter show that the proposed approach allows not only to detect of ratings and their changes but detect news that influenced such changes the most. In the proposed research the model was enhanced with a new data collection algorithm and natural language processing technics.

Objectives: This study sets the complex problem of choosing data source, collecting information, storing it and analyzing it to detect the political rating of presidential election candidates. The choice of each of the listed above

steps can significantly improve or decrease the achieved results. It is important not only to calculate the rating at some point in time but also to have all updates as close as possible to real time and represent it in an understandable comfortable report to allow the user of the system to react quickly to each important event that caused these changes.

Data collection, preparation and analysis

Data sources. Media could be conditionally divided into 5 types:

1) printed media (newspapers, magazines);
2) broadcast media (radio, television);
3) outdoor media (billboards, posters);
4) websites (news sites, blogs);3) social networks: Twitter (https://twitter.com/), Facebook (https://www.facebook.com/), Instagram (https://www.instagram.com/), TikTok (https://www.tiktok.com/), Linkedin (https://www.linkedin.com/), Pinterest (https://www.pinterest.com/).

Most of the types are controlled by some company or person. Therefore, content is filtered and may contain paid articles or videos, which promote needed results. For our research, we need a place where people freely discuss their thoughts with coverage among ordinary people who share opinions not for money. For the listed above criteria social networks are the best match.

Nowadays there are a lot of social networks in the world. Each one of them potentially could be used for forecasting the results of the elections. Which social network to choose among Twitter, Facebook, Instagram, TikTok, Linkedin and Pinterest? Instagram, TikTok and Pinterest are mostly for photos or videos which are hard to analyze and hard to create, which means the amount of political content would be less than with text-based social networks. Linkedin doesn't match the research because of job orientation and low level of political posts. Both Twitter and Facebook are good choices for the research with a lot of users who discuss politics. Accordingly, to Statista (https://www.statista.com/statistics/284435/social-network-penetration-france/) Facebook is the most popular social network in France. The Facebook API is very limited, and it is almost impossible to download a big amount of data from this social network as it is very protected from data collection. Twitter API compared to Facebook's API allows to collect of much more data with less effort. With help of the official Twitter API, it is possible to collect tweets, the number of likes, retweets, and personal data such as user biography, age, location, website, and date joined. Taking to account all the conditions we decided to use Twitter as a data source.

Workflow for election results forecast. Figure 1 presents the general workflow for data collection, storing, processing, visualization, and reporting from the Twitter social network. Official Twitter API is limited, therefore, to collect data for the research hybrid approach is used. Twitter API is used as much as possible to collect data but the type of data that isn't accessible from API is collected by mimicking browser behavior using the Selenium web driver. As data storage MongoDB is used because of the following benefits:

- high availability of data with automatic fast data recovery. In our case it is important not to lose Tweets to have a whole historical dataset;
- in-build sharding solution. Conducting research over years in different countries requires a lot of space to store raw data. Sharding allows to separate of large databases into smaller, faster, more easily managed parts;
- unstable schema. In the beginning, we did not know what kind of data we would be able to download, and which parts would be useful for the research. With an unstable schema, it is easy to add or remove fields over time.

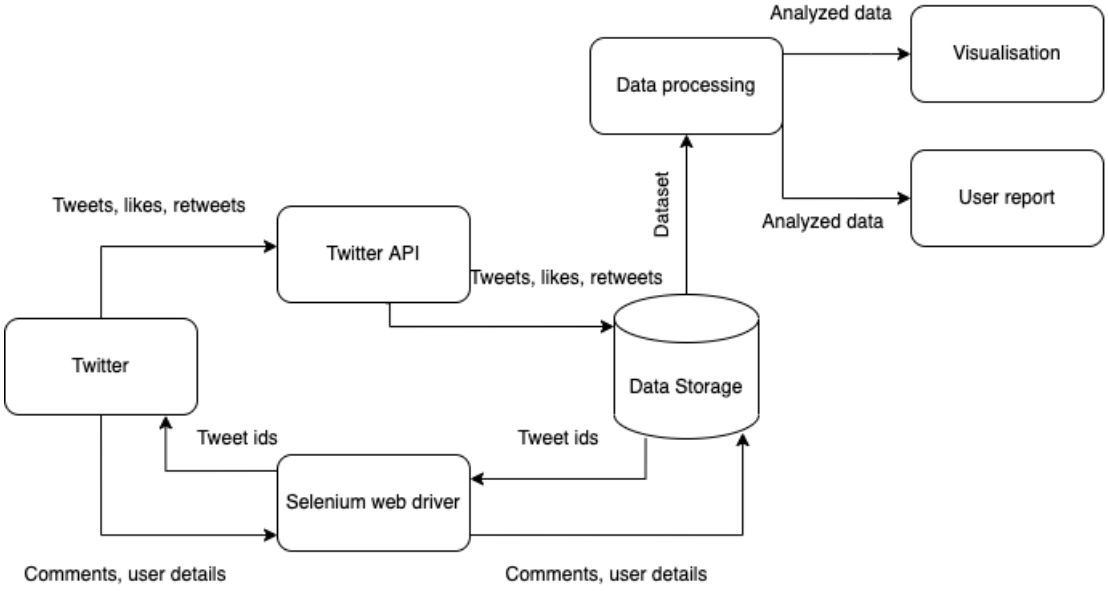

Fig. 1. General workflow for data collection, storing, processing, visualization, and reporting from Twitter social network

From MongoDB data storage CSV datasets are formed for data processing. For each politician, a separate dataset was used. After successful data processing visualization and user report are being built. The report contains whole statistics and comments while visualization – charts which may be used to quickly understand the current situation.

Figure 2 presents a detailed diagram of the data processing of the Twitter CSV dataset. In the research French elections are considered. Tweets about it are mostly in French and English languages. Therefore, other languages are omitted by the algorithm. Twitter text is hard to analyze as a whole. The dataset contains as small one-sentence messages as big messages containing multiple sentences. As the first step for each language text is split into sentences. The next steps perform the following natural language processing techniques:

- remove stop words;
- stemming;
- lemmatization;
- tokenization;
- word sense disambiguation;
- transform words to vectors using Word2Vec;
- sentiment labeling.

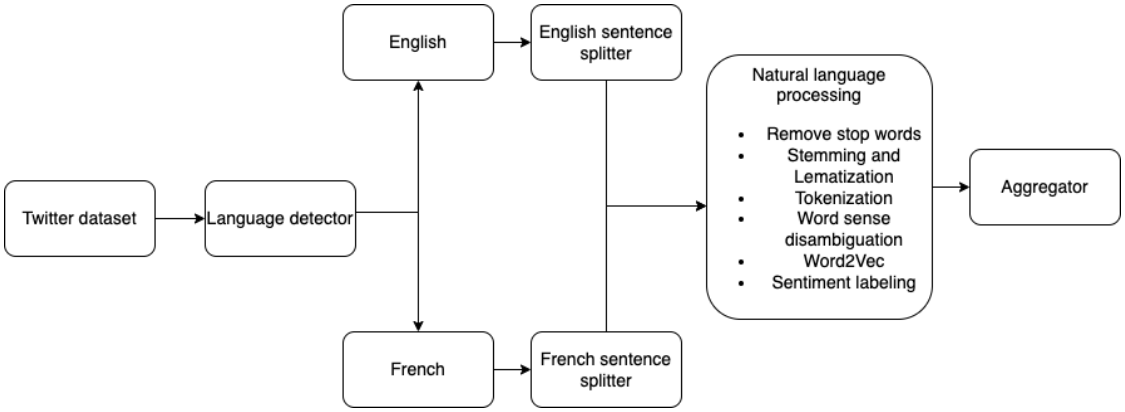

Fig. 2. Detailed diagram of data processing of Twitter CSV dataset

As a result, the algorithm label each tweet from negative to positive represented on a segment from -5 to 5, where -5 is very negative, 0 – neutral, and 5 – very positive. In the final step, tweet scores are aggregated by some period, which could be 1 hour, day, week, month, or custom range inputted by a user.

Data collection. For the research, we collect from Twitter all possible data. The complete list of downloaded data for each tweet is described in Table 1.

Table 1

Downloaded data from each tweet

#	Field	Description
1	Username	Name specified during registration. It is a unique user identifier.
2	Name	No unique name is displayed, which could be edited by a user.
3	Tweet text	The message was written by a user, limited to 280 characters.
4	Hashtags	Proceeds after the "#" symbol. Can be related to an agitation vote for or against any politician.
5	Mentions	Mentioned users. The username proceeds after the "@" symbol.
5	Links	The web address of a website or other tweet.
6	Embedded media	Pictures or videos.
7	Date and time	Date and time when a tweet was published.
8	Replies	Comments to someone's tweet.
9	Favorites	Anyone can highlight tweets that they like.
10	Retweets	Share someone's original tweet.
11	Location	Latitude and Longitude coordinates.
12	Source of the tweet	Android, iPhone or Web.

After collecting tweet data for further research, we need to collect all possible data about the user, who wrote tweets or replies. The complete list of downloaded user data is described in Table 2.

By analyzing user messages and personal information as a result we have a complete portrait of the person and his typical behavior. In most cases, the change in political views is reflected in his social networks acting. For different people it is different. For example, some of them start writing more compliments or hate for political candidates. Some of them just stop writing complement. Such changes are reflected not only in tweets but could also be seen in likes and retweets as well. Some people even change their bio in the profile.

Table 2

Downloaded data for each user

#	Field	Description
1	Username	Name specified during registration. It is a unique user identifier.
2	Name	No unique name is displayed, which could be edited by a user.
3	Location	City, Country.
4	Created date	Date when the Twitter account was created.
5	URL	Link to personal or some other site.
6	Profile Image	User photo or some other image.
7	Language	User-preferred language.
8	Protected	Boolean value (True or False) indicating if a user is protected.
9	Verified	Boolean value (True or False) indicating if a user is verified.
10	Description	Text that the user adds as a profile description.
11	Time zone	Indicates in which time zone the user is.
12	Tweets and replies	All user tweets and replies to other tweets.

Experiment, Results and Discussion

Emmanuel Macron's political rating compared to the sociological poll. The same approach as with the President of Ukraine [10] for political rating detection was tested for the 2021-2022 French presidential campaign. Twitter data was collected weekly from August 2021 until the election itself in April 2022 for two main election candidates – Emmanuel Macron and Marine Le Pen. Overall, the popularity of politicians was growing in social media, candidates had much more new followers compared to those who unfollowed. For example, on August 2021 Emmanuel Macron had 7,219,795 total subscribers but in April 2022 the number became 8,148,825. For each subscriber, the program downloads all tweets from August 2021 and personal data described in Table 1 and Table 2 respectively. By the data, the algorithm found 21,279 bots. Usually, such accounts are created specifically for spreading a lie about politicians, the creation date was in the year 2021 or 2022. Another common part was that such accounts had no real names and surnames in usernames but some text with numbers. By downloading all tweets, we considered which of the accounts have a political position, how often they tweet or like political messages, and whether they changed their opinion over time. After considering and analyzing all possible data political the algorithms forecasted the rating of Emmanuel Macron. Calculated ratings from Twitter data and sociological poll results are presented in Figure 3. Black bars represent the results of the algorithm, and the dashed grey line shows the sociological poll's result (https://www.bbc.com/news/world-europe-59900131). The most important part is that the trend of falling rating right before elections was detected by the proposed algorithm.

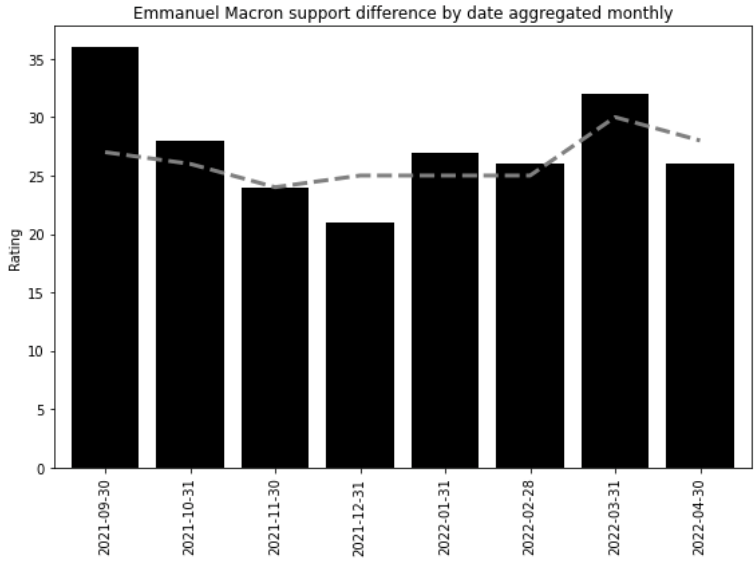

Fig. 3. Emmanuel Macron rating from 30.09.2021 to 30.04.2022

Marine Le Pen's political rating compared to the sociological poll. In the French elections, Marine Le Pen was the main competitor to Emmanuel Macron. Closer to the election date she did a proximity campaign and visited many small towns and villages. Her trips were covered by a lot of local media showing that many people had a chance to see and listen to Marine Le Pen. As with Emmanuel Macron, the popularity in Twitter was growing over time. On August 2021 Marine Le Pen had 2,648,539 total subscribers but in April 2022 the number became 2,818,888. For Marine Le Pen the number of bot accounts was even higher than for Emmanuel Macron – 38 186. Having more bots with a smaller number of subscribers may indicate that more accounts tried to promote Marine Le Pen and

increase the rating of the candidate. Potentially such a strategy may affect rating growth. Calculated Marine Le Pen ratings from Twitter data and sociological poll results are presented in Figure 4.

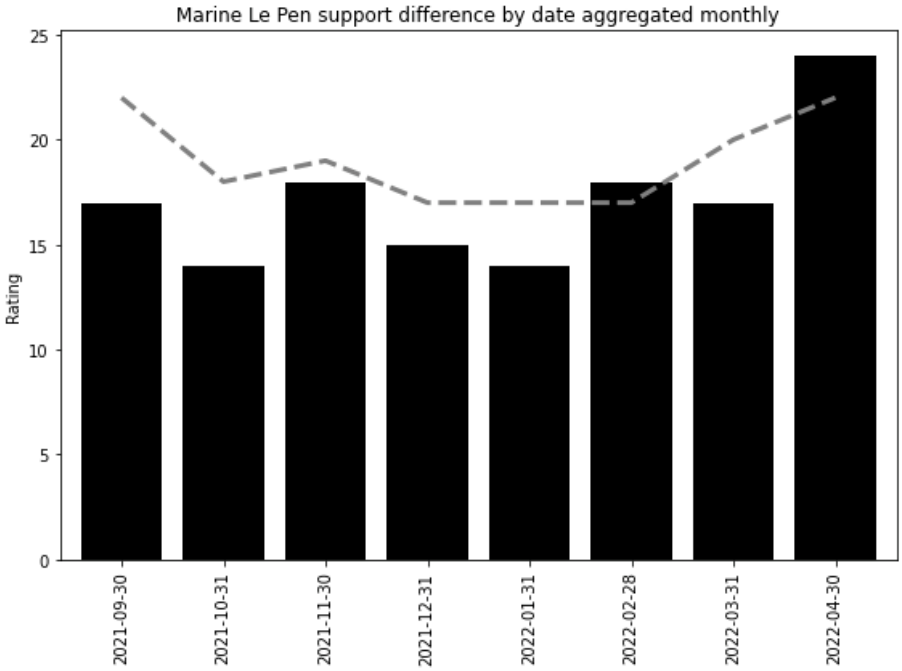

Fig. 4. Marine Le Pen rating from 30.09.2021 to 30.04.2022

Conclusion and Future work

In the conducted research modern types of media are compared to select the best one for election prediction. Selected Twitter social network as the one with the most appropriate data and availability to download big amounts of useful information. The approach based on the usage of Python programming language, Selenium browser emulation and MongoDB database was used to collect, store and clean data about the main French election candidates – Emmanuel Macron and Marine Le Pen. The research was made from August 2021 until the election itself in April 2022. The determined outcome is compared with sociological polls and the results of elections and showed that analysis of social network data could be a good alternative to traditional sociological polls as it shows the same trends month by month and well predicted the win of Emmanuel Macron in elections. Moreover, the proposed approach has its benefits compared to sociological polls such as always being fresh, and close to real-time information, the price of research is much lower and could be reused for the next parliamentary or presidential elections with a small modification.

The research could be extended and adapted for other countries. Currently, the proposed algorithms and mathematical models showed good results in the French and Ukraine elections. It works well with English, French, Ukrainian and Russian languages. This allows us to claim that it will also work fine with other Latin or Cyrillic alphabets but for Asian or Arabic languages more research would be needed. Twitter is a good choice for European and American countries. In the future, other social networks should be considered for the countries in which it is not so popular.

References

1. Y. Roh, G. Heo, S. Euijong Whang. A survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering. 2019. Vol. 33. Issue 4. Pp. 1328-1347.

2. H. Liang, J. J. H. Zhu. Big Data, Collection of (Social Media, Harvesting). The International Encyclopedia of Communication Research Methods. 2017. Pp. 397-416.

3. N. Chomsky. Aspects of the Theory of Syntax. MIT Press (1965).

4. C. F. Hockett. Language, mathematics and linguistics. De Gruyter Mouton (1967).

5. S.-H. Liao, P.-H. Chu, P.-Y. Hsiao. Data mining techniques and applications – a decade review from 2000 to 2011. Expert Systems with Applications. 2012). Vol. 39. Issue 12. Pp. 11303–11311.

6. N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. Medeiros, D. M. F. Mattos. Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information. 2021. Vol. 12. No. 38. P. 1-32.

7. C. Sammut, G.I. Webb. TF–IDF. Encyclopedia of Machine Learning. 2011. Pp. 986-987.

8. T. Mikolov, K. Chen, G. Corrado, Jeff Dean, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Computing Research Repository. 2013. arXiv:1301.3781.

9. Q. V. Le, T. Mikolov. Distributed Representations of Sentences and Documents. Computing Research Repository. 2014. arXiv:1405.4053.

10. T. Rudnyk, O. Chertov. Determining Reasons of Political Rating Changes Based on Twitter Data. XXI International Scientific and Practical Conference "Information Technologies and Security". 2021 (in press).

Taras Rudnyk **Тарас Рудник**	PhD Student of the Applied Mathematics Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" https://orcid.org/0000-0001-9492-0374 e-mail: tarasrudnyk@gmail.com	Аспірант кафедри прикладної математики, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
Oleg Chertov **Олег Чертов**	DrSc (Engineering), Professor, Head of the Applied Mathematics Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" https://orcid.org/0000-0003-0087-1028 e-mail: chertov@i.ua	Доктор технічних наук, професор, завідувач кафедри прикладної математики, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»