IMPACT OF ONLINE SOCIAL NETWORKS AND PUBLIC DEBATES ON THE BEHAVIOUR OF LEGISLATORS

by

Olha Pokhvalenna

A thesis submitted in partial fulfillment of the
requirements for the degree of

MA in Business and Financial Economics

Kyiv School of Economics

2022

Thesis Supervisor: Dr. Tymofii Brik

Approved by _____
Head of the KSE Defense Committee, Professor [Type surname, name]

_____

_____

_____

Date _____

ACKNOWLEDGMENTS

The author is very gratefull to Thesis Advisor Dr. Tymofii Brik and all KSE admission for advices and audience. The author is also deeply thankful to all my groupmates for motivation and belief.

The author is eternally grateful to her parents a for his wisdom, experience, and constant moral support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Number *Page*

LIST OF ABBREVIATIONS

CNN Convolutional Neural Network

LSTM Long-Short Time Memory Neural Network

SVM Support Vector Machines

RF Random Forest

CHAPTER 1. INTRODUCTION

Online social networks have been actively developing for the past 20 years, since the appearance of the first social networks, such as Friendster, Hastag in 2002, and Facebook in 2004. At that time, social networks were created as simple networks for communicating with friends or acquaintances, conducting small talk, and describing one's life. No one could have guessed that social networks are so firmly integrated into people's lives. However, social networks have come a long way and experienced many transformations.

In the 20s of the 21st century, it is impossible to imagine a business that is not promoted on social networks. It is pretty challenging to imagine a person who was able to become famous without social networks. Opinion leaders who are not in social networks are a relatively rare phenomenon - they can be experts in specific narrowly focused scientific or artistic fields, who, in principle, do not have high competition in the market of their activity, or experts who gained popularity many decades ago and are still specialists of the highest equal.

Modern man is, one way or another, integrated into social networks. People either deliberately register in one or more social networks or regularly falls under the influence of social networks through intermediaries: mass media that publish interviews with influencers from social networks or write articles that use publications from social networks, their dynamics, mood, and expert weight.

Politicians and the most influential people on the planet communicate with their audience through social networks. For example, Joe Biden and Donald Trump, as presidents (ex-presidents) of one of the most developed countries in the world, the USA, regularly use Twitter and Instagram for official communications with the population, for a public demonstration of their position on one or another issue, or simply for expressing personal opinions as it was described in article Stolee, Galen, and Steve Caton. "Twitter, Trump, and the base: A shift to a new form of presidential talk?".

Similarly, European and Ukrainian politicians use social networks. For example, in the conditions of war, the President of Ukraine communicates with the population of Ukraine and people from all over the world who are interested in the Russian war on the territory of Ukraine, primarily through social networks - Instagram, his own YouTube channel, and public Telegram channels. Similarly, representatives of the president's office use Twitter and Facebook to publish news, express their opinions, and publicly communicate with foreign colleagues and partners.

In addition to the political sphere, the influence of influencers in social networks penetrates all spheres of life. With one tweet, Elon Musk affects the course of cryptocurrencies as it was described in article Ante, Lennart. "How Elon Musk's twitter activity moves cryptocurrency markets.". A considerable part of marketing campaigns falls precisely on advertising on social networks since people want to use the same products as influencers, visit the same places as influencers and generally get as close as possible to the image of influencers. Moreover, ordinary residents build their worldviews based on information received on social networks.

Currently, the influence of influencers on ordinary people is quite apparent. But do ordinary people have the same impact on influencers and their country? For example, ordinary people run small lifestyle blogs on Instagram, Twitter, and Facebook, where they post their photos, report on significant life events, joke, and express their thoughts. In addition, however, a certain percentage of people express their opinions on current economic and political issues on social networks. This phenomenon is especially relevant for democratic countries with a high level of integration of the population into economic and political life.

For Ukrainians, this is a particularly relevant issue since Ukrainians are at the same time very digitalized people, active users of social networks, and a very politically active nation, especially in the conditions of the Russian-Ukrainian war. Therefore, for Ukrainians, the question constantly arises: is there a point in these constant discussions on Facebook, hundreds of comments under posts on Instagram, and thousands of Ukrainian hashtags

on Twitter? Will the familiar people be heard through these communication channels by our leadership?

That is why it was decided to investigate this question: whether the influence of the dynamics of discussion of political topics in social networks impacts decision-making at the state level. Since the process of communications and meetings on the Ukrainian Internet is quite abnormal in the conditions of war, it was decided to study the data of another country with a high level of involvement in politics and economy and increased activity in social networks - the United States. The tax system and taxes have always been one of the most pressing issues for US residents, and discussions are constantly held on this topic. Therefore, this work will investigate how the dynamics of conversations in the social network Twitter on the subject of taxes affect the intensity of discussions on the tax system and tariffs in the US Congress.

CHAPTER 2. INDUSTRY OVERVIEW AND RELATED STUDIES

2.1 Influence of public opinion in web to governmental decisions

Nowadays, social networks have become one of the most effective tools for self-impression and communication. The government uses Social Networks to publish news and the most important ideas and projects. Ordinary users use it for discussions and replies. The topic of policy is one of the most popular worldwide because of freedom of speech. In democratic countries, people use Twitter and Facebook to discuss and criticize governmental projects and decisions.

For an example of Ukraine, we can see that Facebook posts and twits have a substantial informational sense. Social media use Twitter and Facebook as data sources for articles, the government informs citizens about the essential news and war digests, volunteers spread information about army needs and programs of volunteers organizations, and bloggers call to help refugees and charity programs, ordinary users make a lot of reposts of important hashtags that make the Russian invasion and Russians' war crimes available information for people all over the world. Moreover, relevant modern topics such as human rights, sexism, racism, or LGBTQA+ rights are discussed in social networks much more than in real life. But at the same time, social network users are often criticized. For example, some people say that most posts on social networks on politics and economics are meaningless because ordinary users of Twitter or Facebook have neither education nor experience in those fields, and their posts are white noise in informational space.

Actually, there are a lot of studies that political research discussions on Twitter and Facebook. It's an exciting topic for sociological research like the article "Social Networks and American Politics: Introduction to the Special Issue." The authors reviewed and summarized many theoretical articles and made the structural view of trends in social networks and political studies. In the last 20 years, social network analysis has been one of the most discussed scientific fields, so various approaches have been applied like archive

analysis, sociological surveys on different levels, and laboratory experiments. So network analysis is topical question in politics.

2.2 Studies of Twitter

In article Martin Rehm, Frank Cornelissenb, Ad Nottenc, Alan Dalyd, Jonathan Supovitz "Power to the People?! Twitter Discussions on (Educational) Policy Processes" described how discussions among teachers and educational managers affect policy process using social network analysis, bibliometric analysis, and qualitative interview analysis. Based on this research result, Twitter discussions positively affect academic and social processes because of sharing information and real-time learning. Furthermore, research shows how social network analysis can be used to study the impact of communication on the quality of education.

As in research will be used time-series data article "Detecting trends in Twitter time series" must be beneficial as a reference for time-series analysis. This article described the detection of twitter trends that can be used in market analysis and social network monitoring. The article described different statistical methods for analysis like Poisson process modeling, cross-correlation, and mutual information analysis. Data collection and model tuning was also described in the report. In conclusion, a trend detection is an efficient approach for finding tendencies in tweets' time series. As a result of this research, was created probability model for the detection and prediction of data trends.

How the popularity of social networks and the political activity of the user correlate was investigated in the article Park, Chang Sup. "Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement". This article describes that active users of the Twitter social network tend to research the topic they write about in a short tweet. And since popular users influence ordinary users-readers, it would be logical to suggest that when a small number of opinion leaders are interested in an issue, later the

broad masses also will be interested in this issue. And the active position of society can encourage politicans to discussions.

2.3 Tweets Mood Classification

Analysis of social network posts is a prevalent task in machine learning. For those purposes, NLP approaches are used. Those tools are described in several articles like Balabantaray, Rakesh C., Mudasir Mohammad, and Nibha Sharma "Multi-Class Twitter Emotion Classification: A New Approach." In this article, the authors build a classifier of the emotional class of the writer.

As individuals' emotions play an essential role in all social processes, it's imperative to understand social moods, and Twitter can be a representative sample of it. As people on the internet (mainly anonymous accounts) do not have any reasons to be afraid to show emotions and express their opinion, for research correlation between activity on social networks and parliament activity, Twitter users' average mood must be very representative and valid predictor.

So, in research will be used classification methods described in the article: multi-class SVM kernels model and tweets preprocessing with one-hot encoding. In the article, the model was built with the language of programming Java. Still, the same packages are available in R. We can use the modeling process described in the article as a baseline but use not only SVN but also other classifiers based on LSTM and CNN Neural networks like in article "An Optimized Deep Learning Model for Emotion Classification in Tweets" and random forest like in article "Classification of tweets based on emotions using word embedding and random forest classifiers".

CHAPTER 3. METHODOLOGY

The study aims to quantify how the public activity of bloggers and ordinary Twitter users' posts influence and decision-making process. Another goal is to study the dependencies between the moods of twitter-posts and governmental action. The research's main aim is to understand the "power" activity in social networks and how social trends affect government. It isn't significant if we should discuss political questions on Twitter or Facebook.

Hypothesis 1:

Public opinion about tax policy influences parliament activity in specific time bounds. There will be tested several possible time bounds: 1 day, one week, two weeks, and one month. So, it will be tested four sub-hypotheses: governmental activity in U.S. Congress correlates with activity in social networks with 1-day/1week/2 weeks/4 weeks lag.

This hypothesis is based on the assumption that the government uses the social network as the data source of public opinion and Twitter trends are potent tools for self-opinion impressions which may draw attention not only to friends and like-minded people but also to the government.

Data will consist of:

1. Time Series of twitter-activity - the daily amount of posts with hashtags on topics of taxes and fiscal policy (data extracted by Twitter API).
2. Time Series of Congress activity – daily indicators of taxes and fiscal policy were discussed in Congress, and if some changes were accepted (data scrapped U.S. Congress Daily Digests).

As data has a time-series nature, there must be applied time-series approaches. As almost all real-time-series data has seasonality and periodicity, each time series must be studied as auto-correlated series. For such purposes must be used autocorrelation tests (ACF plot, the Durbin-Watson test and the Breusch-Godfrey test) and time-series

decomposition. Seasonality, periodicity, and trend can be extracted, so it's possible to work only with the trend – increasing or decreasing tendencies in series without periodical fluctuations.

To study the correlation between trends must be applied cross-correlation – a quantitative measure of the correlation between series with the same maximum and minimum values. It is used to compare multiple time series and objectively determine how well they match up with each other and, in particular, at what point the best match occurs.

The sample cross-correlation function between series X and Y:

$$r_{x,y}(k) = \frac{c_{x,y}(k)}{\sqrt{c_{x,x}(0) * c_{y,y}(0)}} \quad (1)$$

Where covariance between series X and Y:

$$c_{x,y} = \frac{1}{N}\sum_{t=1}^{N-k}(x_i - x_{avg}) * (y_i - y_{avg}) \quad (2)$$

And $c_{x,x}(0)$ and $c_{y,y}(0)$ that are the series X and Y variances.

Also, for correlation study can be applied mutual information and partial mutual information measures - information-theoretic measures of time-series cross-correlation.

Mutual information is a nonlinear quantitative metric of mutual dependence between two-time series. This metric is frequently used in signal analysis and machine learning. It can be defined as:

$$\mathrm{MI}(\mathrm{X},\mathrm{Y}) = -\sum_{x_i,y_i} \mathrm{p}(\mathrm{x}_i,\mathrm{y}_i)\log_2\frac{\mathrm{p}(\mathrm{x}_i,\mathrm{y}_i)}{\mathrm{p}(\mathrm{x}_i)\mathrm{p}(\mathrm{y}_i)} \quad (3)$$

Partial mutual information is a generalization of partial correlations, which is sensitive to nonlinear dependencies, to which Pearson's correlation and partial correlation are strictly not sensitive.

All these metrics are appropriate for correlational analysis, so by those values, we can check how strict the correlation is described in Table 1.

Table 1. Correlation measures

Correlation abs Value	**Dependency between variables**
0.90 - 1.00	Very high correlation
0.70 - 0.90	High correlation
0.50 - 0.70	Moderate correlation
0.30 - 0.50	Low correlation
0 - 0.30	Negligible or weak correlation

Source: Author's calculations

Hypothesis 2: Average emotional moods of twitter-influencers influence the decision-making process in specific time bounds.

As additional parameters for research, the correlation between social network activity and governmental decision-making process can be used moods of twitter-posts. With Natural Language Process tools, it's possible to define a post's emotion: is it angry, sad, fearful, neutral, or satisfied. There is an assumption that angry posts have a mode impact on government. Therefore, they may avoid real-life protests and pay attention to the negative moods of society on Twitter.

To define the level of unsatisfaction in posts will be applied several NLP models based on neural networks LSTM-CNN and classical Machine Learning models (logistic regression, SVM, tree-based algorithms).

It's common practice to classify texts by emotions, so there is already an existing solutions based on the combination of convolutional neural networks and long-short

term memory neural networks. This networks are tuned on the massive amount of data and called Simple Transformer.

As already existing datasets with labeled tweets emotions (SMILE Tweeter emotion dataset)- it's possible to fit transformer with this dataset and test its accuracy. Tweets on the topic "taxes and fiscal policy" must be classified with Simple Transformer, SVM, logistic regression, Decision trees, RF and other models. After that the best model must be selected. After that data can be aggregated, there will be several new variables:

- Percent of angry posts on the date;
- Percent of negative posts on the date;
- Percent of fear posts on the date.

So, several other time series must be tested in correlation with the time series of Congress activity. Approaches of correlational analysis will be the same as in Hypothesis 1 (Cross-Correlation, Mutual Information, Partial mutual information).

And of course, as the last step, it is very appropriate to study the influence of combinations of various indicators of the dynamics of activity in social networks (the total number of likes and reposts, the number of posts, indicators of the emotional coloring of posts, the metric of the importance of posts) on the activity of discussing taxation issues in the US Congress. For this purpose, it is possible to use several models that can be interpreted to explain the interrelationships of the VAR model.

A linear regression model using panel data is used as a baseline model. In addition, panel data are increasingly used in econometric analysis tasks: a model built on panel data allows you to study a sample of objects over time and identify and take into account the peculiarities of each sampling unit.

Panel data is a two-dimensional array in which one of the dimensions is "spatial" (the axis of observations, usually the dates of the data frame), and the other is temporal (the axis of time slices, usually the columns of the data frame). Such arrays arise when data is collected on a given set of objects over a certain period. In the case of research for testing the above hypothesis, the data should be presented in the form of:

- Lines - observations responsible for the analytical instrument. Each observation is characterized by the end date of the week as an index by which the values of the columns can be determined;
- Columns - characteristics responsible for the description of the observation. In the case of the baseline, they are the target variable, the percentage of reports in which taxation was mentioned, and independent variables - the number of tweets on the topic of "taxation" for the previous week, the week before the previous one, etc., a total of 4 time slices.

The resulting baseline model, built on panel data, will allow the specified test of hypothesis 1. In addition, by interpreting the results, it will be possible to understand the influence of the number of tweets with 1-day/1week/2 weeks/4 weeks lag.

As a development of the model, linear regression will be used, including sections on the number of tweets and areas on the emotional coloring of tweets (the percentage of tweets whose text is negatively colored, expressing anger, fear, anxiety, or sadness). This model will help with testing the second hypothesis mentioned above.

In addition, we will study the influence of some other characteristics of tweets - the number of reposts, the number of likes, and the general estimation of the tweet (estimate of the power of the tweet). For this purpose, variables acting as model predictors for the previous two weeks will be selected for a more explicit interpretation.

VAR or VEC models (depending on the nature of the target variable) will be used for a more qualitative and in-depth analysis of the time series and the search for hidden effects.

VAR models are used to forecast interrelated time series systems and analyze the dynamic influence of disturbances (shocks) on the design of selected indicators. VAR models are used for forecasting systems of interconnected time series and for analyzing the active influence of disturbances (shocks) on the method of selected indicators.

The VAR model is a system of equations in which the value of each endogenous variable is determined by the previous values of this and the other endogenous variables of the system. The model describes the mathematical expectation of the future importance of variables as a linear function of several variables' current and past values.

A VAR model can include current and lag values of exogenous variables and logical variables responsible for regime changes in economic policy or individual shocks in the economy.

The VAR-model or VAR(p) for endogenous variables Y1, Y2 with the selected order of the model p = 2 (two time periods) is a system of equations:

$$\begin{cases} Y_{1t} = Y_{1t-1}\alpha_{11} + Y_{1t-2}\alpha_{12} + Y_{2t-1}\beta_{11} + Y_{2t-2}\beta_{12} + \alpha_1 + u_{1t} \\ Y_{2t} = Y_{1t-1}\alpha_{21} + Y_{1t-2}\alpha_{22} + Y_{2t-1}\beta_{21} + Y_{2t-2}\beta_{22} + \alpha_2 + u_{2t} \end{cases} \quad (4)$$

Similarly, when the order of the model is increased, the equation will become more complicated.

Conclusions about the reaction of target variables to shocks in predictors are made on the basis of analysis of impulse response functions "Shock" or "innovation" – a

one-moment change in an endogenous (exogenous) variable equal to its one standard deviation fluctuates over the entire observed period The impulse response function characterizes time the return of an endogenous variable to an equilibrium trajectory with a single shock of an exogenous variable.

Thus, with the help of the VAR model, we will be able to detect the impact of shocks in independent variables (the influence of the number of tweets, emotional coloring of tweets - the percentage of tweets whose text is negatively colored, expressing anger, fear, anxiety, or sadness, number of reposts, the number of likes, and the general estimation of the tweet) on the dynamics of the target- the percentage of reports in which taxation was mentioned.

CHAPTER 4. DATA

Data from several sources were used to study the influence of the activity of social network users on the activity of members of Congress.

4.1 Target variable

The first stage was the design of the dataset. It is necessary to push away from the target variable. There were two target options to study the influence of social networks on the activities of the U.S. Congress:

1. Binary variable, an indicator of mention of the tax system and taxes at congressional meetings
2. The number of mentions of taxes at meetings.

The website of the U.S. Congress was used as a data source. This site publishes records of congress speeches, congress digests, committee meetings, official publications, and other official information.

At work, the members of Congress' speeches were used as the data source for the most complete and representative information.

In the process of preliminary analysis of digests of congressional meetings from the official website of the U.S. Congress, it was found that the discussion of taxes and the tax system is a consistently relevant topic for conferences, as an example of mentions of taxes – “BUILD BACK BETTER ACT; Congressional Record Vol. 168, No. 139” or “PROVIDING FOR CONSIDERATION OF SENATE AMENDMENT TO H.R. 5376, BUILD BACK BETTER ACT; Congressional Record Vol. 168, No. 135”. Moreover, economic issues, budgeting, and taxes are

discussed at almost every meeting. Therefore, the binary variable is not representative enough. Moreover, its distribution is unbalanced since a meeting without mention of taxes is a relatively rare event.

Therefore, it was decided to count the number of mentions of the keywords "taxes," "tax system," "income tax," and "taxes for the population" in the records of congressmen's speeches. Data were collected over one calendar year: from 01.08.2021 to 30.08.2022. At the same time, the fact that there may be a different number of speeches at meetings was taken into account, so a correct target would be the percentage of reports in which taxation was mentioned.

The peculiarity of the data is that the congress meetings are not held every day, so it was possible to collect about 200 observations.

In the future, more detailed targets can be explored, such as mentioning taxation of the population in public statements, adoption of bills related to tax, etc.

4.2 Predictors mining

Currently, most social networks provide access interfaces for information in them. G giants such as META (Facebook, Instagram), Google, Twitter, and many other companies offer APIs.

However, the user privacy and confidentiality policy should be considered. There are closed profiles on Facebook and Instagram - i.e., profiles whose publications cannot be accessed legally. According to a 2018 study, 45% of Americans run all social networks privately, and 20% run social networks partially privately.

At the same time, the difficulty of working with Facebook data is that Facebook is a desirable platform for marketing and a lot of advertising. Unfortunately, some advertisements "trigger" users with clickbait titles and descriptions, and the topic of taxes is of interest to many. Therefore, such advertisements can introduce quite

significant Bayes bias into the data under study. Thus, a relatively limited amount of data can be obtained from Facebook, although Facebook users often discuss politics and economics.

Another product is Instagram meta, which provides primarily visual information with posts. Features of privacy settings on Instagram are the same as on Facebook. That is, you can get a relatively limited number of posts. At the same time, Instagram is primarily a social network for visual content, whether private or public. Users publish mostly visually engaging content for social recognition, finding friends, business development, gaining popularity, etc. Popular topics include humor, travel, brands, and personal blogs. Content about news, politics, and economics is very narrowly focused and impacts a small percentage of the audience.

Google can provide quite exciting information - for example, you can get the number of search queries in a specific area by keywords. However, this search engine monitoring option is only available for marketing campaigns. To collect enough data, you need to create and configure a marketing campaign that will track queries to the search engine over a long period. The marketing campaign must work for at least a year to collect data for a year.

Similar to Instagram, data from social networks such as TikTok and Snapchat are not intended to spread opinions about politics and the economy. In addition, they mostly do not use descriptions, so their analysis is inappropriate.

Twitter is optimal as a data source. Tweets are short messages in which users mainly express their thoughts. Although Twitter is not a popular advertising platform, tweets are usually user messages. Furthermore, all Twitter accounts are public, so it is an optimal data source.

Twitter provides a unique interface for software developers to access real-time and historical data (streaming data). With the help of the API, a developer who has

received confirmation of access rights from Twitter receives requests to read recorded data.

With the help of the API, you can get data about such instances as posts, users, user statistics (number of followers, number of followings, estimate of user influence metric), and post statistics (number of likes, reposts, comments, and estimation of tweet influence). You can also get the post's text and work with it.

Keywords associated with the topic of taxes and the tax system were selected. Such as "tax", "taxation", "tax system", "income-tax" and variations of their spelling. After that, a sample of tweets containing the required keywords and the word USA/US/United States was downloaded using the API and the python programming language. In total, over 120,000 tweets were uploaded over the past year.

The data from the posts were further processed and aggregated for use in various statistical and machine learning models.

The following were identified as predictors that can be further used in models in various combinations:

1. Number of tweets;
2. Percentage of tweets expressing fear or apprehension;
3. Total number of likes on posts;
4. Total number of retweets on posts;
5. Number of likes on the most popular tweet;
6. Number of reposts on the most popular tweet;
7. The percentage of tweets with a negative emotional color (determined using machine learning classification models), separately for the emotion of fear, the emotion of anger, and the emotion of sadness;

All data were aggregated daily. The number of tweets and other aggregation metrics was calculated for each day. However, since the target data is not daily (the Congress sits with breaks), the data was aggregated further. Yes, the year was not

divided by days, but by weeks, since every week there was at least one session of the Congress. Yes, new variables were created:

1. Total number of tweets per week;
2. Average daily percentage of tweets expressing fear or apprehension;
3. The total number of likes on posts per week;
4. Total number of retweets on posts per week;
5. The number of likes on the most popular tweet per week;
6. Number of reposts on the most popular tweet per week;
7. The average percentage of tweets with a negative emotional color (determined using classification models of machine learning), separately for the emotion of fear, the emotion of anger, and the emotion of sadness;

Example of result dataframe is described in table 2 below.

Table 2. Example of final dataset:

Date	Congress mentions	Congress records	Retweet count sum	Retweet count max	Likes count max	Likes count sum	Tweet value max	Tweet value mean	Emotion anger share	Emotion fear share	Emotion sadness	Tweet count sum
26.09.2021	0	19	0	0	0	0	11,9	3,18	0,08	0	0,39	13
27.09.2021	13	121	0	0	0	0	2,31	0,70	0	0	0,2	5
28.09.2021	23	193	47	26	95	170	88,6	5,26	0,03	0,02	0,13	72

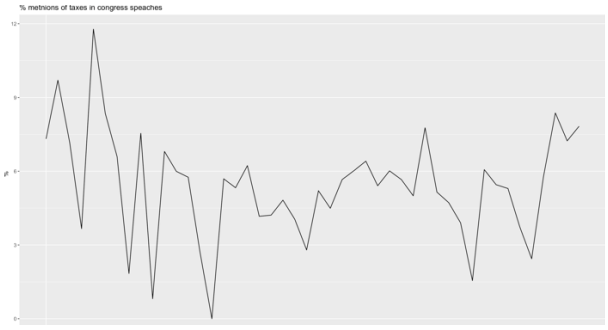
Source: Author's calculations

The main idea is that the target (the number of speeches by members of Congress mentioning taxes) is influenced by the dynamics of changes like tweets (the predictors discussed above).

4.3 Analysis of variables

First, the target variable's dynamics were investigated (figure 1). Visually, the percentage of mentions of the tax system at congressional meetings does not have a clearly defined dynamic, trend or seasonality.

Figure 1. Target time Series

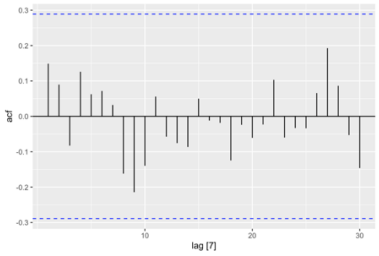


Source: Author's calculations

According to the ACF and PCF graphs (Figure 2 and Figure 3), the data of the target variable are stationary. In the future, it will be possible to use the VAR model for more in-depth modeling of the nature of the dynamics of the time series and its relationships with the time series of independent variables.

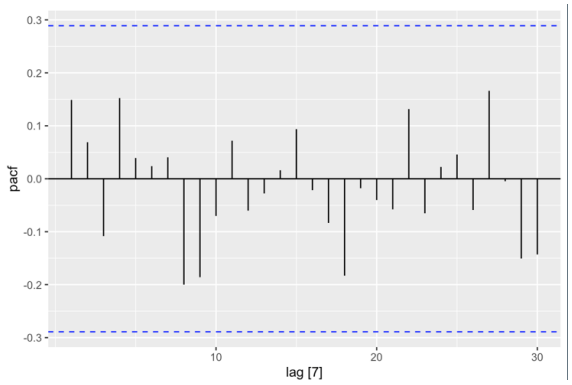
Dependent variables are described in the table 3 below.

Figure 2. Target ACF



Source: Author's calculation

Figure 3. Target PCF



Source: Author's calculations

Table 3. Descriptive statistics of independent variables:

	min	**max**	**range**	**sum**	**median**	**mean**
Retweet count sum	0,00	5193	5193	49901	677,50	1084,80
Likes count sum	0,00	19959	19959	232740	3458,00	5059,57
Tweet count sum	16,00	3287	3271	66247	1351,50	1440,15
Metnions_share	0,00	0,12	0,12	17564	0,06	0,05
Tweet value mean	2,17	18,48	16,31	266,09	5,08	5,78
Emotion anger share	0,00	0,36	0,36	4,33	0,08	0,09
Emotion fear share	0,00	0,05	0,05	0,59	0,01	0,01

Source: Author's calculations

For the interpretation of statistical models, there must be no multicollinearity. Multicollinearity occurs when more than two factors are linearly related. That is, there is an influence of factors on each other. The multicollinearity means that some elements will always act in unison. In other words, the correlation coefficient between these two factors has a value close to or equal to 1.

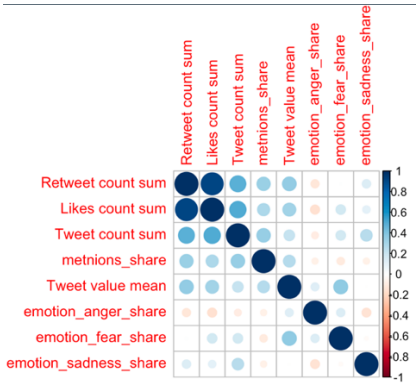
Having checked the dataset, we can establish whether there is a correlation between variables and thus find out which combinations of columns can be used as predictors for models. As a result of the check, a cross-correlation table of independent variables was kept in the table 4 below.

Table 4. Results of cross-correlation testing of predictors:

	Retweets count	**Likes count**	**Tweet count**	**Metnions share**	**Avg tweet value**	**Emotion anger share**	**Emotion fear share**	**Emotion sadness share**
Retweets count	1,00	0,91	0,48	0,33	0,35	-0,13	0,01	0,14
Likes count	0,91	1,00	0,50	0,28	0,32	-0,16	0,17	0,11
Tweet count	0,48	0,50	1,00	0,35	0,22	-0,09	0,18	0,25
Metnions, %	0,33	0,28	0,35	1,00	0,26	-0,08	-0,11	-0,08
Avg tweet value	0,35	0,32	0,22	0,26	1,00	0,12	0,37	0,00
Emotion anger, %	-0,13	-0,16	-0,09	-0,08	0,12	1,00	0,13	-0,15
Emotion fear, %	0,01	0,17	0,18	-0,11	0,37	0,13	1,00	-0,03
Emotion sadness, %	0,14	0,11	0,25	-0,08	0,00	-0,15	-0,03	1,00

Source: Author's calculations

Figure 4. Correlation matrix



Source: Author's calculations

According to the results of the cross-correlation test, it was found that Retweets count and Likes count are highly correlated. Therefore, we will use only one of these two predictors in the models. In other cases, the variables correlate with each other either weakly or ultimately insignificantly.

CHAPTER 5. RESULTS

5.1. Determining the emotional color of a tweet

The process of determining the sentiment rating of a tweet should be highlighted separately. This is a fairly common task that many researchers and developers have previously dealt with. For example, Balabantaray, Rakesh C., Mudasir Mohammad, and Nibha Sharma study "Multi-class twitter emotion classification: A new approach." There are already marked datasets in the format of pairs of correspondences between the text of a tweet and its emotional coloring. Since targeting 120,000 tweets manually is a rather time-consuming task, machine learning models were used to classify tweets, which were built based on public datasets of marked tweets. For simulation, tweets were converted into numerical multi-hot encoding using the CountVectorizer transformer approach - the tweet's text was transformed into a matrix of token counts.

Several statistical models of classical machine learning (SVM, decision tree, random forest, multivariate logit regression) and deep learning LSTM/CNN were used in the python programming language. The results of the efficiency of the models are shown in Table 5.

Table 5. Metrics results of multiclass classification of tweets:

Model	**Score**
Logistic Regression	0.8113
Decision Tree	0.8357
Random Forest (500 estimators)	0.8676
SVM	0.8835
LSTM/CNN	0.8187

Source: Author’s calculations

The models are implemented at the basic level, without the selection of hyperparameters. In further researches, the models could be fine-tuned using a cross-validation approach and, theoretically, the metrics should improve. The model with the best value of the multi-class classification accuracy metric - SVM - was selected. Tweets were categorized as follows in table 6.

Table 6. Classification of tweets predictions:

Type	**Tweets Count**
neutral/positive	76595
sadness	38783
anger	11543
fear	1848

Source: Author's calculations

As a result of all aggregations, the final dataset was obtained. Before the modeling stage, a preliminary analysis of the dataset was carried out. It will help to understand the nature of the data and to choose the models that will be used in the future.

5.2. Dependencies research

Several models were built for the study to clarify the nature of the dependencies.

As a baseline, the percentage of reports in which taxation was mentioned depended on the number of tweets on taxes in the previous week, two weeks ago, three weeks ago, and four weeks ago. But, first, let's find out whether the predictors in the panel form correlate. The results of the collinearity test are presented in Table 7.

Table 7. Panel Predictors Correlation:

	Prev week tweets count	**Prev 2 week tweet count**	**Prev 3 week tweets count**	**Prev 4 week tweets count**
Prev week tweets count	1,00	0,03	0,00	0,27
Prev 2 week tweets count	0,03	1,00	0,21	-0,03
Prev 3 week tweets count	0,00	0,21	1,00	0,19
Prev 4 week tweets count	0,27	-0,03	0,19	1,00

Source: Author's calculations

Based on this table, we can see that the data correlate very weakly, so the interpretation of the linear model will be pretty simple and correct. Finally, the results of the model are shown in Table 8.

Table 8. Baseline linear model coefficient:

	Estimate	**Std.**	**Error**	**t value**	**significance**
(Intercept)	4.773e	0.862	5.535	2.67e-06	***
prev_week_tweets_cnt	0.0012	0.0003	3.407	0.0016	**
prev_2week_tweets_cnt	-0.0008	0.0003	-2.218	0.0328	*
prev_3week_retweet_cnt	0.00001	0.0003	0.034	0.9730	
prev_4week_retweet_cnt	-0.00019	0.0003	-0.626	0.5354	

Source: Author's calculations

That is, the baseline model can be interpreted as follows: if the number of tweets on the topic of taxation increases by 1000 in the previous week (10E3), then the percentage of reports in which taxation was mentioned will increase by 1.7%. At the same time, if there were 1,000 more posts two weeks ago, the percentage of reported mentions of taxation would decrease by 0.75%. This can be explained as follows: the delta of tweets is not so noticeable. The dynamics are pretty stable, and there is no flash of user interest, so the impact is insignificant.

The model explains 30.78% of the variance of the target variable. However, there is a hypothesis that the emotional coloring of tweets can also influence the target variable. Let's take the data for the last two weeks since the previous model showed that this data is both economically and statistically significant.

Table 9. Linear model built on twitter mood predictors coefficient

	Estimate	**Std.**	**Error**	**t value**	**significance**
(Intercept)	3.6799118	1.0752506	3.422	0.0016	**
prev_week_tweets_cnt	0.0011044	0.0003454	3.198	0.0029	**
prev_2week_tweets_cnt	-0.000754	0.0003576	-2.11	0.0420	*
prev_week_emotion_anger_share	0.82768	0.47073935	1.705	0.0969	.
prev_2week_emotion_anger_share	0.1434608	0.4794647	0.299	0.7665	
prev_week_emotion_fear_share	-7.869377	23.9118738	-0.329	0.74404	
prev_2week_emotion_fear_share	9.0373915	23.9719479	0.377	0.70845	

Source: Author's calculations

The model explains 37.55% of the variance of the target variable. As seen in the model output, the percentage of tweets expressing anger and displeasure is economically and statistically significant. For example, with a 1% increase in dissatisfied tweets over the previous week, the target percentage of reported mentions of taxation will increase by 0.82%.

Similarly, variables were added with the number of likes (both maximum and average), number of retweets, and tweet rating. These variables did not have a statistically significant effect.

A VAR model was also constructed. Automatically, it was advised to build a model for lag 1 (that is, only data from the previous week acted as predictors). With such settings, there are no statistically significant predictors for the target.

Specific patterns already appear if you increase the lag to at least 2. The variables that affect the target are the percentage of tweets expressing anger in the current week and the number of tweets written two weeks ago. This is quite counterintuitive, as the number of tweets in the last week would be expected to have the biggest impact. The results of the VAR model can be found in Appendix A.

In general, the results of the VAR model confirm the correlations found in previous models: the highest impact had activity on social networks last week and angry sentimental tweets.

However, it is important to understand that based on the results of the models, we cannot unequivocally assert a causal relationship. There may be external factors that influence both the activity of users on social networks and the discussions in the US Congress. For example, articles in economic publications or lectures by prominent scientists. Theoretically, users may actively discuss and quote significant genders on Twitter in no time. At a time when this article will have an impact on politicians, however, they will study the materials in more detail and discussions may come later. Therefore, the search for correlations between the frequency of mentioning the topic of taxation in leading offline and online publications (number of articles,

circulation of the publication, number of likes), the dynamics of the discussion of this topic on Twitter, and the dynamics of the discussion of the topic of taxation in the US Congress can be a possible way of developing the research.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

Often, social networks are mistakenly perceived as platforms for frivolous conversations, jokes, and the promotion of small businesses. In addition, political discussions on social networks are usually condemned and criticized. The argument for such statements is that the average user is not an expert in economics or politics.

There was investigated question whether the influence of the dynamics of discussion of taxational topics in social networks impacts on public debates on the behaviour of legislators. The data of activity in social networks in the United States was studied: a country with a high level of involvement in politics and economy. Therefore, this work investigates how the dynamics of conversations in the social network Twitter on the subject of tax affects public debates on the behaviour of legislators in the US Congress.

As a result of the study, the hypothesis was confirmed that users' activity on Twitter correlates quite significantly with the action of discussions in the congress. This is shown by both simple linear models built on panel data and the VAR model. Data from one week ago have the most significant impact: the more tweets on the topic of "taxes" were published by users in the previous week, the more attention is paid to the issue of the tax system in political and economic discussions at the state level.

The following hypotheses were confirmed:

1. Public opinion about tax policy influences parliament activity in specific time bounds. Were tested several possible time bounds: one week, two weeks, 3 weeks and one month. So, there were tested four sub-hypotheses: governmental activity in U.S. Congress correlates with activity in social networks with 1week/2 weeks/3 weeks/4 weeks lag.

   The highest impact had activity on social networks last week. That means that Congress can react quickly to changes in society's moods on social networks.

2. Average emotional moods of twitter-influencers influence the decision-making process in specific time bounds.

   The highest impact had angry posts. That means Congress can react to angry society's moods on social networks, not fearful or sad.

Such correlations may indicate that social networks are a powerful tool in the hands of the public. Popularization of hashtags on Twitter, active discussions, mass appeals to politicians on Twitter - all this makes sense.

Public opinion and activism undoubtedly have an impact on state authorities. One way or another, activity in social networks is a kind of marker of the population's mood. In addition, writing a post on the web is much easier than being a member of a social movement or an activist in real life.

In the future, this research can be significantly improved and expanded.

First, the system for determining the mood of a tweet could be improved. At this stage of research, the most basic models have been implemented. It is possible to build a more complex model based on an ensemble of fast models with pre-selected hyperparameters using cross-validation methods.

In the future, a slightly different target can be investigated - for example, indicators of the adoption of draft laws. This will study the impact not so much on the discussion process but on the final decision-making process. This will be a somewhat more complex study, as many more factors affect the operation of passing bills, and the process of passing bills is much more complex and lengthy.

Another important aspect is the study of the influence of discussions on other social networks, particularly on Facebook. It would be very appropriate to investigate this specific social network for several reasons:

1. Posts are mostly text or contain a text description
2. The average age of a Facebook user is 40 years old. That is, the audience is mature and perhaps more inclined to conduct serious, reasoned discussions

3. There are thematic publics on Facebook, in particular political ones. Data from such publics can be very informative.
4. To conduct such a study, research the legal limitations of data scraping from Facebook and get access to the META API.

At the same time, the analysis of search queries on Google is very informative. These are data of a different nature - they are anonymized. That is, no one from the network can directly see search queries. That is, Google search statistics can show the accurate level of interest of the people and not the level of interest that they are willing to demonstrate.

REFERENCES

Ante, Lennart. "How Elon Musk's twitter activity moves cryptocurrency markets." *Available at SSRN 3778844* (2021).

Balabantaray, Rakesh C., Mudasir Mohammad, and Nibha Sharma. "Multi-class twitter emotion classification: A new approach." *International Journal of Applied Information Systems* 4.1 (2012): 48-53.

Gaikwad, Govin, and Deepali J. Joshi. "Multiclass mood classification on Twitter using lexicon dictionary and machine learning algorithms." *2016 international conference on inventive computation technologies (icict)*. Vol. 1. IEEE, 2016.

Fiedor, Paweł. "Partial mutual information analysis of financial networks." *arXiv preprint arXiv:1403.2050* (2014).

Park, Chang Sup. "Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement." *Computers in human behavior* 29.4 (2013): 1641-1648.

Papana, Angeliki, and Dimitris Kugiumtzis. "Evaluation of mutual information estimators for time series." *International Journal of Bifurcation and Chaos* 19.12 (2009): 4197-4215.

Rehm, Martin, et al. "Power to the people?! Twitter discussions on (educational) policy processes." *Mixed Methods Social Network Analysis*. Routledge, 2019. 231-244.

Singla, Chinu, et al. "An Optimized Deep Learning Model for Emotion Classification in Tweets." *Comput. Mater. Contin* 70 (2022): 6365-6380.

Stolee, Galen, and Steve Caton. "Twitter, Trump, and the base: A shift to a new form of presidential talk?." *Signs and society* 6.1 (2018): 147-165.

Vora, Parth, Mansi Khara, and Kavita Kelkar. "Classification of tweets based on emotions using word embedding and random forest classifiers." *International Journal of Computer Applications* 178.3 (2017): 1-7.

Congressional records https://www.congress.gov/

Twitter developer API https://developer.twitter.com/en/docs/twitter-api

SMILE Tweeter emotion dataset

https://figshare.com/articles/dataset/smile_annotations_final_csv/3187909

APPENDIX A

Results of VAR MODEL

	Dependent variable:				
	y				
	(1)	(2)	(3)	(4)	(5)
metnions_share.l1	0.008	3,956.728	-0.081	0.023	-1.131**
	(0.176)	(7,543.064)	(0.540)	(0.111)	(0.545)
Tweet.count.sum.l1	0.00000	0.202	0.00000	-0.00000	0.00002
	(0.00000)	(0.205)	(0.00001)	(0.00000)	(0.00001)
emotion_anger_share.l1	0.094*	2,513.601	-0.232	0.001	0.076
	(0.055)	(2,367.086)	(0.170)	(0.035)	(0.171)
emotion_fear_share.l1	-0.099	372.812	0.039	-0.024	-0.603
	(0.289)	(12,374.970)	(0.886)	(0.181)	(0.894)
emotion_sadness_share.l1	0.012	1,859.807	-0.258	0.007	0.156
	(0.056)	(2,401.102)	(0.172)	(0.035)	(0.173)
metnions_share.l2	0.213	-808.060	-0.347	0.007	-0.047
	(0.182)	(7,804.863)	(0.559)	(0.114)	(0.564)
Tweet.count.sum.l2	-0.00001*	0.020	0.00001	0.00000	0.00001
	(0.00000)	(0.206)	(0.00001)	(0.00000)	(0.00001)
emotion_anger_share.l2	0.046	3,273.113	0.055	-0.008	0.086
	(0.056)	(2,389.081)	(0.171)	(0.035)	(0.173)
emotion_fear_share.l2	0.173	775.464	-0.708	-0.127	-0.628
	(0.284)	(12,193.410)	(0.873)	(0.179)	(0.881)
emotion_sadness_share.l2	-0.004	1,076.044	-0.217	-0.016	-0.050
	(0.052)	(2,218.354)	(0.159)	(0.033)	(0.160)
const	0.035	-403.871	0.257***	0.016	0.270***
	(0.029)	(1,227.045)	(0.088)	(0.018)	(0.089)
Observations	44	44	44	44	44
R2	0.221	0.185	0.162	0.074	0.257