Cognitive, Psychological and Social Influence on Spread of COVID-19

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute, Jožef Stefan IPS
Jamova cesta 39
Ljubljana, Slovenia

Vito Janko
vito.janko@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Tine Kolenik
tine.kolenik@ijs.si
Jožef Stefan Institute, Jožef Stefan IPS
Jamova cesta 39
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We investigated and confirmed the hypothesis that cognitive, psychological and social features of citizens in each country influence the spread of COVID-19 more than any other semantic feature group. Additionally, we investigated five sub-hypotheses in regards to socio-psychological traits of people and the spread of COVID-19, confirming two and rejecting three. Finally, we attempted to obtain deeper understanding of our results by finding which individual features within the social psychology group are most important.

KEYWORDS

psychology, sociology, covid-19, machine learning, feature analysis

1 INTRODUCTION

Since the spring of 2020, Coronavirus disease 2019 (COVID-19) has increasingly influenced our daily lives. The first wave of infections started to manifest globally around March, and different countries reacted differently and with different amounts of success in order to stop the early exponential growth. Countries differ from one another in many aspects, such as weather, demographics, development, economic strength, etc. Another important but often overlooked difference between countries is in the cognitive, psychological and social features of their citizens. We argue that these are some of the most important factors that might influence the spread of COVID-19, as they in turn influence how much people spend time with each other, how often they attend social and cultural events, etc. Thus, we focused on analysing these features in terms of their influence on spread of COVID-19 and their importance compared to other groups of features. Additionally, we investigated the importance of individual features that comprise the category of cultural features in an attempt to investigate if there is a single defining trait that dominates others.

The rest of this paper is structured as follows: we first investigate the related work in Section 2, then we list hypotheses in section 3 and describe the data in Section 4. We continue with the methodology and experimental setup in Section 5, and conclude with results and discussion in Section 6.

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

2 RELATED WORK

We focused on COVID-19 related work that deals with some properties of different world regions (typically countries) and compares them to a target variable related to the spread of COVID-19 in that region – with the goal of establishing the relationship between the two.

Many authors defined the spread of the disease in different ways. Most commonly researchers simply used the number of daily infections as the metric, which has the weakness of being biased towards countries with higher population, but can be normalized per capita [1]. Some other options are also possible, such as computing the reproductive rate of the virus, as proposed by Gupta et al. [6].

The country properties used to investigate the influence on virus spread were also varied. Most commonly, weather attributes were investigated [6], as well as indicators of development [1] and demographics [8].

In terms of machine learning (ML) methods, classical regression (e.g., linear regression) was used predominantly [6], while others used traditional statistical approaches [8], testing for statistically significant correlation between features and target variables.

Despite the large amount of research conducted in regards to COVID-19, the aspect of cognitive, psychological and social influence on the potential spread of COVID-19 has been poorly researched thus far, to the best of our knowledge. We aim to investigate and highlight the importance of the aforementioned influences and hopefully motivate more researchers to consider this important area.

3 RESEARCH HYPOTHESES

Unlike the various different influences on COVID-19 spread that related works focused on, the aim of this study was to concentrate on human behavior in terms of their social psychology, or interaction between their cognitive and psychological features and their social behavior. Generally, we believe that these significantly affect COVID-19 spread and should therefore be investigated to further understand not only this particular pandemic, but the influence of human behavior on pandemic in general.

Our primary research hypothesis is that human behavior statistically significantly affects COVID-19 spread. Furthermore, we have five secondary hypotheses:

(1) People with higher tendencies for social activities (higher extraversion) correlate with higher COVID-19 spread.
(2) People with higher tendencies for social compliance (higher agreeableness) correlate with lower COVID-19 spread.
(3) People with higher tendencies for being careful (higher conscientiousness) correlate with lower COVID-19 spread.
(4) People with higher tendencies for group consideration (lower individualism) correlate with lower COVID-19 spread.
(5) People with higher tendencies for desire gratification (higher indulgence) correlate with higher COVID-19 spread.

To investigate our research hypotheses, we turned to data repositories with psychological, cognitive and social features across countries. Since our final dataset will contain other features as well, those will be also investigated. The next section describes this data.

4 DATA

As our aim was to use ML algorithms to investigate the potential relationship between cognitive, psychological and social features of citizens and the spread of COVID-19 on per-country basis, we had to obtain and structure suitable data. The cognitive, psychological and social features were used as input features and were obtained for as many countries as possible. The spread of the virus itself was modelled using several binary classes, which were the targets of our classification.

4.1 Features on social psychology

To research our hypotheses, we did a limited literature review to find data spread between features that describe individual behavior and features that describes group behavior of societies as a whole. We selected three frameworks with which to work in this research. To account for individual behavior, the Big Five personality traits model [5] was selected, along with a feature on preferred interpersonal distances [11]. To account for group behavior, Hofstede's cultural dimensions theory [7] was selected.

The Big five personality traits model (B5) identifies five orthogonal dimensions which supposedly reflect an individual's personality and psyche. B5 is measured with a questionnaire. Extensive research has found significant statistical connections with a number of human behaviors (decision-making, crime, voting, health behavior, educational outcomes, etc.) [2]. B5 includes the following dimensions:

(1) Openness: describes how inventive or curious someone is.
(2) Conscientiousness: describes how careful, efficient or organized someone is.
(3) Extraversion: describes how outgoing or energetic someone is.
(4) Agreeableness: describes how friendly or compassionate someone is.
(5) Neuroticism: describes how sensitive or nervous someone is.

Data on preferred interpersonal distances comes from human spatial behavior research [11] and describes how comfortable people are in regards to different distance boundaries when in contact with other people.

Hofstede's cultural dimensions theory (HCDT) identifies six orthogonal dimensions that describe a country's values that drive their group behavior. They have been found to correlate with a number of social phenomena (security, progress, environmental outcomes, etc.) [7]. HCDT includes the following dimensions (we did not include *Power distance* as it did not relate to our goal of finding data that describes phenomena that lie between individual and group behavior):

(1) Individualism-collectivism: describes how citizens of a country prefer and care for their in-group.
(2) Uncertainty avoidance: describes how averse citizens of a country are to uncertainty.
(3) Long-term orientation: describes how traditional citizens of a country are in terms of solving society's questions and their proclivity for change and adaptation (higher score means more long-term thinking, more adaptation and change).
(4) Indulgence: describes the degree to which citizens of a country seek desire fulfilling behavior.
(5) Task- vs. person-orientation: describes preference of citizens of a country towards tasks versus towards people.

Data on B5 questionnaire answers, which was collected from Open-Source Psychometrics Project's public database [9] (under "Answers to the IPIP Big Five Factor Markers"), had to be additionally pre-processed for this research. We processed the answers to the questionnaire to get individual personality profiles with the five dimensions for every person. Then we filtered the data by only keeping the countries where we had 100 individuals answering the questionnaire. Afterwards, we averaged the scores by countries to get group personality profile, each country having five dimensions.

Finally, we also considered data on levels of a nations' strength of social norms – referred to as (cultural) tightness-looseness. We used the tightness measure from Gelfand and colleagues [3]. The measure captures the strength of norms in a nation and the tolerance for people who violate norms. The final dataset we constructed contains 59 countries (meaning 59 instances) with 11 features.

The dataset can now be related to the hypotheses: 1) for secondary hypothesis 1, extraversion will be used for correlation; 2) for secondary hypothesis 2, agreeableness will be used for correlation; 3) for secondary hypothesis 3, conscientiousness will be used for correlation; 4) for secondary hypothesis 4, individualism-collectivism will be used for correlation; 5) for secondary hypothesis 5, indulgence will be used for correlation.

4.2 Virus spread classes

We chose three distinct binary classes, each having two possible values: a country is considered positive if its infection rate, given the chosen metric, is faster than half the countries analyzed. The class value was always computed in country-specific time frame, starting when the testing was adequate in a country according to the recommendation given by the World Health Organization (WHO), and ending when at least 3 countermeasures of sufficient intensity were applied. This intensity was marked with an integer in the range from 0 to 4 in the Oxford Covid-19 Government Response Tracker [12], and we took value 2 as the threshold.

4.2.1 Daily number of infections (daily average). The first calculated metric was the daily number of infections, averaged over the appropriate time interval and normalized based on the country population. This metric is the most intuitive and commonly reported.

4.2.2 Reproductive rate. The reproductive rate R_0 is a metric commonly used by virologists to determine the severity of an infection. Simply put, it estimates how many new infected are generated by each currently infected.

To estimate the reproductive rate we used the SIR model [10]. For details on the computation of the values, we refer the reader to the original paper.

4.2.3 Exponential shape. The last metric we calculated was the shape of the infection time series. An exponential shape indicates that the number of infections is raising fast, and is likely to continue. To determine if the growth is exponential, we fitted both a linear and an exponential curve to the data. After both were fitted, the one with the lower error was chosen as the better fit. If the exponential fit was better, the class value for this metric was positive.

Once the class was determined, we could split the countries into infected, non-infected and those for which we do not have enough data, based on each of the three classes. An intuitive display of the split is shown in Figure 1, where countries are colored based on the number of positive virus spread classes.

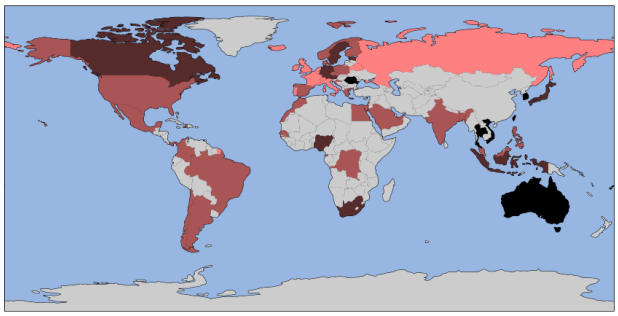

Figure 1: All countries, colored based on how many infection classes are positive. If all of them are positive, the color is light red, and conversely gets darker for every negative one. Countries without sufficient data are gray. Note that the data is from spring 2020, showing only the early spread.

5 METHODOLOGY

We first focused on testing our hypothesis of social psychology feature group being among the most important in the spread of COVID-19 compared to other feature groups describing a country. After confirming our initial hypothesis, we then investigated importance of individual features within this group.

5.1 Feature group importance

We obtained over 100 different country-describing features in order to compare them against the social psychology group, and to investigate our primary hypothesis, which was that the social psychology group is highly important. To do this, we first grouped all other individual features into the following semantic groups: weather, travel, health, economy, development, geography, countermeasures. We then evaluated the importance of each feature group using a Random Forest (RF) classifier. The model was trained using all the features and individual feature importances were obtained out of the box via the *feature_importance* property of the model, which is available in the scikit-learn implementation. In summary, this metric trains an RF classifier consisting of a number of different trees. When training a tree, it computes how much each feature decreases the weighted impurity in this tree. This impurity decrease is then summed up over all the trees in the forest to form the feature importance. We then summed feature importances within each previously defined group to compare the aggregate importance of groups. This was done for each of the three virus spread classes.

5.2 Individual feature importance

Once we estimated feature groups importance, we turned our focus to analysis of individual features within social psychology group. We investigated whether an individual or small set of features dominate a group in regards to importance, or is the importance rather evenly spread. We did this for each of the three classes using three different methods. Additionally, this gives us information about specific best features within the group, which allows for potential expert interpretation.

(1) **RF feature importances**: First, we again used the out-of-the-box feature importances of RF to compare the importance of individual features.
(2) **Statistical testing**: Second, we used statistical tests depending on the type of feature (continuous, categorical, binary, normally distributed, non-normally distributed). The feature values of countries positive with respect to a class were compared to those negative with respect to a class. We used the T-test, Mann-Whitney U-test and Fisher-exact test, respectively, for continuous normal features, continuous non-normal features, and binary features.
(3) **Wrapper method**: Third, we developed a custom feature selection wrapper method similar to the one used in our related work [4], which did the following: the features were first sorted using RF feature importance (as before). Then, if two features were correlated (Pearson coefficient > 0.7) we discarded the lower ranking one. We started by using only the best feature for the classification. Then, we iteratively added the next best one, but only kept it if it did not decrease the classification accuracy by more than two percentage points. This method improves upon the first one by considering internal correlations between features.

The five secondary hypotheses were investigated using correlation analysis, by computing the correlation between the values of the selected individual feature relevant for the hypothesis, and the daily average class. We did this to get a deeper understanding and potentially new knowledge of exactly which features influence acceptance or rejection of our hypotheses.

6 EXPERIMENTS AND RESULTS

Aggregate RF feature importances for each group and each class are given in Table 1. Looking at the average importance, we see that the social psychology group of features proved the most important, alongside development, confirming our initial hypothesis.

The importances of top 5 individual features within the social psychology group for all three classes is given in Table 2. The importances were evaluated using the three different feature importance methods described previously.

Finally, the evaluation of our initial secondary hypotheses using correlation analysis is given in Table 3.

7 CONCLUSION

We investigated the cognitive, psychological and social influence on spread of COVID-19. Comparing against other semantic

Table 1: Aggregate feature ranking using RF feature score. Values are normalized (sum to 1).

	Repr. rate	Exp.	Daily avg.	Average
Weather	0.09	0.08	0.09	0.09
Social psychology	0.18	0.21	0.14	**0.18**
Travel	0.12	0.08	0.18	0.13
Economy	0.15	0.13	0.09	0.12
Development	0.16	0.18	0.12	0.18
Geography	0.12	0.06	0.11	0.10
Health	0.11	0.19	0.11	0.14
Countermeasures	0.04	0.02	0.06	0.04

Table 2: Individual feature ranking using RF feature score, statistical testing and wrapper method. Top 5 features and corresponding scores are shown.

RF feature importance (higher is better)		
Repr. rate	*Exp.*	*Daily avg.*
Tightness (0.071)	EST_perc (0.053)	AGR_perc (0.032)
EST_perc (0.014)	Masculinity (0.017)	Individual. (0.024)
OPN_perc (0.013)	Individual. (0.015)	OPN_perc (0.014)
Future ori. (0.013)	CSN_perc (0.015)	Future ori. (0.012)
Masculinity (0.07)	Tightness (0.014)	Masculinity (0.011)
Statistical significance (lower is better)		
Tightness (0.010)	EST_perc (0.076)	Individual. (0.030)
Future ori. (0.064)	Tightness (0.148)	AGR_perc (0.045)
OPN_perc (0.171)	CSN_perc (0.148)	Indulgence (0.112)
EXT_perc (0.259)	Uncert. avoid. (0.2)	OPN_perc (0.134)
AGR_perc (0.259)	Masculinity (0.241)	EXT_perc (0.147)
Wrapper method (higher is better)		
Tightness (0.071)	EST_perc (0.053)	AGR_perc (0.032)
CSN_perc (0.005)	Masculinity (0.017)	Individual. (0.024)
/	Individual. (0.015)	OPN_perc (0.014)
/	CSN_perc (0.015)	Future ori. (0.012)
/	Tightness (0.014)	CSN_perc (0.005)

Table 3: Correlation analysis of our secondary hypotheses in respect to the daily average virus spread class.

Hypothesis	Correlation	Accept/Reject
Higher extraversion correlates with higher virus spread	0.33	ACCEPT
Higher agreeableness correlates with lower virus spread	0.40	REJECT
Higher conscientiousness correlates with lower virus spread	0.04	REJECT
Higher individualism correlates with higher virus spread	0.46	ACCEPT
Higher indulgence correlates with higher virus spread	0.09	REJECT

groups of features describing countries, we showed that the social psychology group has the highest feature importance alongside development. Additionally, we found that there is no single dominant feature in our set of 11 in the social psychology group, but instead the importance is spread among several. We also used correlation analysis to confirm two out of our five hypotheses, showing high correlation between extroversion and individualism and higher virus spread. This shows that the cognitive, psychological and social features are among the most important in relation to spread of COVID-19 and should be investigated more thoroughly.

ACKNOWLEDGMENTS

This work is part of the ongoing research at the Department of Intelligent Systems, Jožef Stefan Institute. It is a subset of a larger COVID-19-related research, which is subject to potential future publications. The authors also acknowledge the financial support from the Slovenian Research Agency (ARRS).

REFERENCES

[1] Rodrigo M Carrillo-Larco and Manuel Castillo-Cara. 2020. Using country-level variables to classify countries according to the number of confirmed covid-19 cases: an unsupervised machine learning approach. *Wellcome Open Research*, 5, 56, 56.

[2] P.T. Costa and R.R. McCrae. 2013. *Personality in Adulthood: A Five-Factor Theory Perspective*. Taylor & Francis. ISBN: 9781135459703.

[3] Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, et al. 2011. Differences between tight and loose cultures: a 33-nation study. *science*, 332, 6033, 1100–1104.

[4] Martin Gjoreski, Vito Janko, Gašper Slapničar, Miha Mlakar, Nina Reščič, Jani Bizjak, Vid Drobnič, Matej Marinko, Nejc Mlakar, Mitja Luštrek, et al. 2020. Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors. *Information Fusion*.

[5] L. R. Goldberg. 1993. The structure of phenotypic personality traits. *The American psychologist*, 48, 1, 26–34.

[6] Akash Gupta and Amir Gharehgozli. 2020. Developing a machine learning framework to determine the spread of covid-19. *Available at SSRN 3635211*.

[7] Geert Hofstede. 2011. Dimensionalizing cultures: the hofstede model in context. *Online Readings in Psychology and Culture*, 2, 1.

[8] Yothin Jinjarak, Rashad Ahmed, Sameer Nair-Desai, Weining Xin, and Joshua Aizenman. 2020. Accounting for Global COVID-19 Diffusion Patterns, January-April 2020. Technical report. National Bureau of Economic Research.

[9] Open-Source Psychometrics Project. 2011. Accessed: 2020-04-10.

[10] David Smith, Lang Moore, et al. 2004. The sir model for spread of disease: the differential equation model. *Loci.(originally Convergence.)*

[11] Agnieszka Sorokowska and Piotr Sorokowski et al. 2017. Preferred interpersonal distances: a global comparison. *Journal of Cross-Cultural Psychology*, 48, 4, 577–592. DOI: 10.1177/0022022117698039.

[12] Anna Petherick Toby Phillips Thomas Hale Sam Webster and Beatriz Kira. 2020. Oxford covid-19 government response tracker. https://github.com/OxCGRT/covid-policy-tracker.