Vilnius University Press

Vilnius University Press

Politologija
2019/2, vol. 94, pp. 56–80

ISSN 1392-1681 eISSN 2424-6034
DOI: https://doi.org/10.15388/Polit.2019.94.2

The Opportunities and Limitations of Using Artificial Neural Networks in Social Science Research

Lukas Pukelis
Senior Researcher at the Public Policy and Management Institute (PPMI)
email: lukas.pukelis@ppmi.lt

Vilius Stančiauskas
Research Director at the Public Policy and Management Institute (PPMI)
email: vilius@ppmi.lt

Summary. Artificial Neural Networks (ANNs) are being increasingly used in various disciplines outside computer science, such as bibliometrics, linguistics, and medicine. However, their uptake in the social science community has been relatively slow, because these highly non-linear models are difficult to interpret and cannot be used for hypothesis testing. Despite the existing limitations, this paper argues that the social science community can benefit from using ANNs in a number of ways, especially by outsourcing laborious data coding and pre-processing tasks to machines in the early stages of analysis. Using ANNs would enable small teams of researchers to process larger quantities of data and undertake more ambitious projects. In fact, the complexity of the pre-processing tasks that ANNs are able to perform mean that researchers could obtain rich and complex data typically associated with qualitative research at a large scale, allowing to combine the best from both qualitative and quantitative approaches.[1]
Keywords: Deep Learning, Artificial Neural Networks, Natural Language Processing, Text Analysis.

[1] This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 770531. Additionally, it was supported by the Ministry of Science and Education of Lithuania under agreement No. S-424.

Received: 06/03/2019. **Accepted:** 10/07/2019

Neuroninių tinklų modelių taikymo socialiniuose moksluose galimybės bei ribos

Santrauka. Pastaruoju metu neuroninių tinklų modeliai yra vis dažniau taikomi įvairiose mokslo srityse (mokslometrijoje, lingvistikoje, medicinoje). Tačiau socialiniuose moksluose šie modeliai ir giliojo mokymosi metodai apskritai kelią skinasi sunkiai. Pagrindinės menko jų populiarumo priežastys yra tai, kad netiesinių modelių rezultatus yra sunku interpretuoti ir jie nėra tinkami hipotezėms tikrinti. Nepaisant to, šie metodai gali būti labai naudingi socialiniuose moksluose, ypač automatizuojant darbui labai imlias tyrimo proceso dalis. Straipsnyje pristatomas pavyzdys, kaip šie modeliai yra naudojami atliekant didelio masto teksto kodavimą. Pateikiamas pavyzdys rodo, kad kai kuriais atvejais šie metodai leidžia išvengti kompromiso tarp analizės gylio ir imties dydžio, taip pat leidžia kombinuoti geriausias kokybinių ir kiekybinių metodų praktikas.
Reikšminiai žodžiai: gilusis mokymasis, neuroninių tinklų modeliai, teksto analizė, natūraliosios kalbos analizė.

Introduction

It is customary to teach undergraduate social science students the distinctions between "qualitative" and "quantitative" approaches in the very first social science methodology classes. It has long been accepted that social science research involves a certain trade-off between the depth and the width of enquiry. On the one hand, you have quantitative methods, which allow one to work with large sample sizes and draw inferences that can be generalized across vast populations, but suffer from a lack of depth and level of detail. On the other hand, you have qualitative methods, which produce deep and rich insights about a small number of cases but do not allow generalizations.[2]

The distinction between these two approaches results from a simple truth: a typical research team does not have the capacity to collect and process enough data to produce rich and detailed insights for the number of cases sufficiently large to be generalizable. As a result, research teams have to make the choice to either produce shallower insights for a large number of cases (quantitative approach) or rich insights for a small number of cases (qualitative approach).

[2] King G., Keohane R. O., and Verba S., *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton: Princeton University Press, 1994, p. 3.

However, given the recent advances in technology, especially in the fields of artificial intelligence and machine-learning, this limitation can be overcome by outsourcing some of the menial labor to machines. In this paper, we present a broad introduction to how machines can be used for such tasks as data collection and pre-processing, thus allowing to draw comparatively rich insights for a large number of cases. We argue that with these tools, small research teams could be empowered to undertake more ambitious research projects and produce results that would combine the strengths of both qualitative and quantitative approaches.

More specifically, we provide in this paper an overview of one family of algorithms called *artificial neural networks* (ANNs), which have recently become widely used in business and research spheres due to their capacity to identify and "learn" highly complex and non-linear data patterns. In the first part of the paper, we present a brief historical overview of how these algorithms were developed and how they are used today for a wide variety of purposes. We also present the case why ANNs constitute a significant improvement over the previous generation of algorithms and discuss their features that are of particular interest to social scientists. In the second part of the paper, we describe the weaknesses and limitations of these algorithms as well as the reasons why their adoption in social sciences has been slow. Various recent criticism toward these algorithms is also presented in this part. In the third part of the paper, we make the case of how ANNs could benefit the social science community even in light of their current limitations and drawbacks. Meanwhile, in the fourth part, we present an account of how we applied ANNs to a practical research problem of automated text classification.

1. History and Overview

Artificial Neural Networks (ANNs) belong to the family of machine-learning algorithms. The field of machine learning aims to develop solutions or algorithm-systems that would allow computers to recog-

nize or "learn" the underlying structure in data and make predictions based on those patterns. As such, machine learning algorithms aim to generate predictions, and they are evaluated based on the quality of these predictions. Often the "learning" and prediction-making is powered by the algorithms that are well-familiar to social scientists, such as OLS and logistic regression, k-means clustering, or multi-dimensional scaling. However, the machine learning versions of these algorithms often differ from their cousins in the "classical" statistics in some minor details, such as the way they are optimized or what kind of parameter turning is available for the user in their most popular implementations.

Machine learning models are usually divided into two groups: unsupervised learning, aiming to "learn" and make judgements based only on the structure of data (e.g., k-means clustering), and supervised learning, aiming to approximate *ex-ante* known labels given the input features (e.g., OLS). ANNs can be used in both settings due to their versatility, but for the sake of simplicity from here on out, the paper will only consider the supervised machine learning applications of ANNs.[3] Supervised machine learning problems are usually divided into two groups: regression problems, where the outcome variable/label is continuous, and classification problems, with categorical output variables. ANNs are used to address both these problems, though in this paper the majority of the examples henceforth belong to the classification problem category.

Early versions of ANNs can be traced back to the 1950s, when in the aftermath of WW2, researchers sought to capitalize on the recent advances in computing to create an "artificial brain." In 1958, Frank Rosenblatt presented an idea that the function of a biological neuron can be simulated mathematically using relatively simple equations

[3] For more information on the applications of ANNs for unsupervised machine learning tasks, refer to Bansal S., *How Autoencoders Work – Understanding the Math and Implementation,* <https://www.kaggle.com/shivamb/how-autoencoders-work-intro-and-usecases>, 10 10 2018.

comparable to those used in logistic regression models.[4] Each biological neuron has several dendrites with which it receives signals from the other neurons or receptors; these signals are then processed in the cell body, and a signal (or lack thereof) is then passed to other connected neurons via the axon. Similarly, in the artificial neuron/logistic regression model, the input features are summed, weighted, and passed through the activation function, which outputs either a zero or a one (see Figure 1). In turn, such neurons can be connected into more complex systems or networks that could tackle difficult problems. For instance, some of the early applications of such artificial neurons included recognizing handwritten digits and distinguishing the photos of men from those of women.[5]

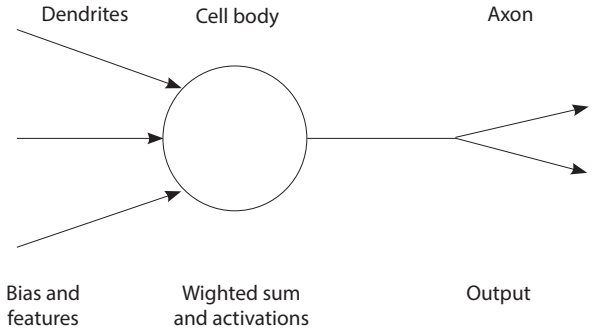

Figure 1. ***Natural and artificial neurons.***
Source: created by the authors.

Despite early success, the interest in ANNs eventually faded away because the research has failed to deliver significant results despite large amounts of funding allocated to the field. This occurred due to

4 Rosenblatt F., "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* 65 (6), 1958, p. 386.

5 Michigan Institute of Technology, "The Thinking Machine" (1961) – MIT Centennial Film," <https://techtv.mit.edu/videos/10268-the-thinking-machine-1961---mit-centennial-film>, 10 10 2018.

two major reasons: first, technology at the time could not cope with the complexity of the tasks; second, the initial versions of the ANNs had a major flaw – they could not tackle non-linear problems.[6]

Consider the examples in Figure 2: on the left, the two classes are linearly separable – it is possible to draw a line in the distribution plane that would cleanly divide the two classes. Meanwhile, on the right side of the figure, it is not possible – the two classes, though clearly distinct, cannot be separated by any straight line. The example on the left is an instance of a linear problem, while the example on the right is a non-linear problem.

While early ANNs could tackle problems where the classes were linearly separable, they were not effective in solving non-linear problems and, as such, were not very useful.

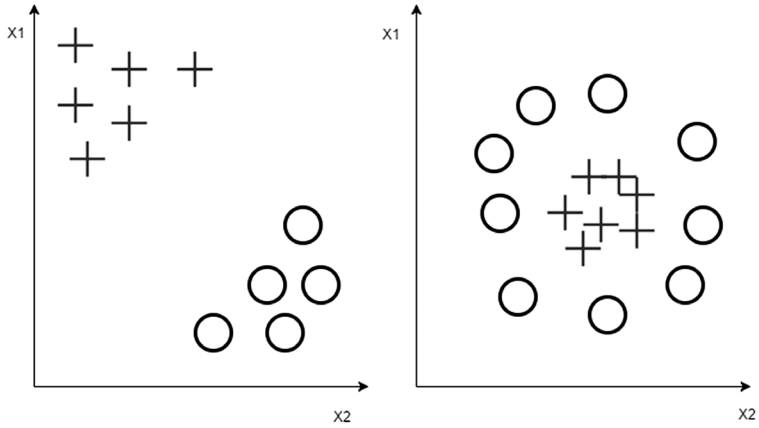

Figure 2. ***Linearly separable and non-separable classes.***
Source: created by the authors.

The problem of non-linearity was solved by developing a method of back-propagation,[7] which allowed to stack artificial neurons not

[6] Nils J. N., *The Quest for Artificial Intelligence: The History of Ideas and Achievements*, Cambridge: Cambridge University Press, 2010, p. 38–100.

[7] Rumelhart D. E., Hinton G. E., and Williams R. J., "Learning Representations by Back-Propagating Errors," *Cognitive Modeling* 5 (3), 1988, p. 1.

only side-by-side but in multiple layers, creating so-called deep neural networks, which have hidden layers of artificial neurons between the input and the output. The hidden layers can identify complex interplay and relationships between the different input features, which allows them to solve nonlinear problems. Figure 3 shows a multilayered perceptron, or an ANN, with at least one hidden layer. As can be seen from the figure, each node in the hidden layer receives inputs from all the nodes in the input layer, aggregates them, and passes them on. It can be said that the hidden layers in such ANNs serve to compute complex interactions between all the input features and then generate intermediate outputs based on these complex interactions rather than the raw input features.

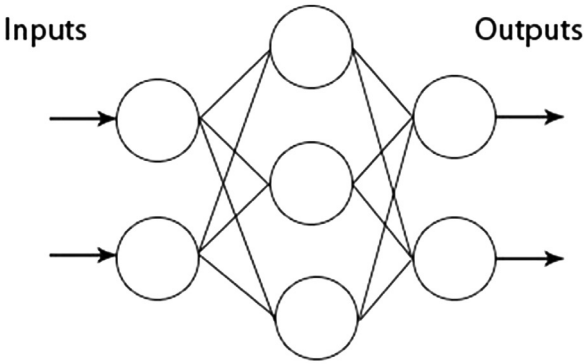

Figure 3. ***Multi-Layer Perceptron.***
Source: created by the authors.

Nonetheless, even with these advances, ANNs did not gain mainstream popularity outside a small, dedicated research community until the 2010s. This was mostly due to the fact that training ANNs with multiple hidden layers is very computationally expensive and out of reach for a typical lay user. Yet with recent advances in computational power, and especially GPU computing, they have been adopted as the current state of the art in multiple fields.

2. Problems of Using ANNs in Social Sciences

The previous section has established that modern ANNs have the capability to solve complex and nonlinear problems due to hidden layers in their structure. This feature of ANNs, while allowing them to generate accurate predictions, is inherently problematic for social scientists. Usually, social scientists seek to *understand* the relationship between two or more variables, i.e., to answer a question like "Are X and Y related?" or "Does X have a significant effect on Y?" Meanwhile, the goal in machine learning is to be able to *predict* what will happen in the future. In machine learning, the most important question is "Can I *approximate* Y if I know X?" Hence, in machine learning, models are developed on sets of data where X (features or independent variables) and Y (outputs or dependent variables) are known. The models are evaluated by supplying them with previously unseen batches of Xs and seeing how accurately they can predict Y. Therefore, the metrics on which models are evaluated are not the statistical significance of individual variables but various metrics used to estimate how well the model fits the data. In classification exercises, they can be accuracy scores (correctly classified cases/total number of cases), and in regression tasks – R^2.

Hidden layers in ANNs make it impossible to tell exactly how the inputs and outputs are related. ANNs can determine whether it is possible to correctly predict/estimate the output given by a particular input, but what exactly happens "under the hood," or how exactly a particular input feature influences the output, remains unknown and impossible to determine. Only in recent years there has been some success in demonstrating what happens in the hidden layers of the neural network. For instance, "TensorFlow Playground"[8] does a wonderful job visualizing MLP performance, but it is limited to a small number of data sets and few model parameters.

[8] A website for visualizing ANNs "TensorFlow Playground", <https://playground.tensorflow.org/>, 10 10 2018.

Furthermore, such an inability to pinpoint exactly how the input affects the output and whether this effect is significant makes ANNs not very useful in answering questions typically posed in social science research. Let us consider a question whether a person's physical height affects workplace success and income.[9] To answer this question, we would like to not only explain variations in workplace success and income using a model, which, among other things, uses a variable for a person's height, but we would also be interested in whether height exercises a significant individual effect even when controlling for various background characteristics (such as age, gender, and education). To answer this question, we would have to turn to familiar statistical modelling, most likely a form of linear regression, and ANNs would be of little help. However, it is very likely that given the same inputs, an ANN model would be able to predict workplace success or income and that the ANN model would have better explanatory power (higher R^2) than the linear regression.

Recently, additional criticisms have been levied against the use of ANNs in various spheres due to the ethical concerns they pose.[10] As machine learning algorithms learn the underlying structure of the data and then make decisions based on these learned features, their decisions are of the same quality as the data on which they are trained. In other words, given data that are biased in some ways, the ANNs will make decisions containing the same bias. Consider a now famous example of how the city of Boston had attempted to crowdsource the scheduling of road repairs. The idea was simple: citizens would download an app to their phones and drive around town, and if their car hits a pothole, the app will report the location of the pothole to the authorities. With enough datapoints, the city would have a real-time map of pothole locations and could schedule their repairs

9 Judge T. A., and Cable D. M., "The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model," *Journal of Applied Psychology*, 89 (3), 2004, p. 428.

10 Mittelstadt B. D. et al., "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3 (2), 2016.

accordingly. The problem is that socially vulnerable residents in poor neighborhoods are less likely to have a smartphone and participate in the experiment, which means that the potholes near their homes are not mapped and repaired.[11]

Since the functioning of ANNs is relatively opaque, it is very hard to detect whether the input data contain any significant biases and whether the model outputs are affected by such biases in any way. However, several initiatives were recently started to increase "algorithmic fairness" and to address such concerns.[12]

3. Opportunities of Using ANNs in Social Sciences

The abovementioned problems are sufficient to question whether it is appropriate to use ANNs in social sciences at all. We maintain that these algorithms, despite their shortcomings, are extremely powerful tools that could benefit many social scientists. Their opaqueness and general orientation toward predicting rather than explaining make them of limited utility as analytical instruments, though recently in some disciplines (like psychology) there have been several attempts to model highly complex interactions with ANNs rather than more established techniques, such as structural equation models.[13]

However, we propose that ANNs could be most beneficially used in a different manner – as tools to pre-process the data and prepare it for modelling with more traditional methods. In fact, it can even be stipulated that with the help of ANNs, the existing gap between qualitative and quantitative methodologies can be bridged. In essence, qualitative analysis entails taking raw unstructured data, which can

---

11 Crawford K., "The Hidden Biases in Big Data," <https://hbr.org/2013/04/the-hidden-biases-in-big-data>, 10 10 2018.

12 Algorithmic Fairness and Opacity Working Group in UC Berkeley, <https://afog.berkeley.edu/>, 10 10 2018.

13 Janciauskas M., and Franklin C., "Input and Age-Dependent Variation in Second Language Learning: A Connectionist Account," *Cognitive Science* 42 (S2), 2018, p. 519–554.

come in a variety of formats (text, images, audio, etc.), and then, using one's analytical prowess, recasting these data in a more structured way. Using ANNs such tasks, normally assigned to human coders, could be "outsourced" to computers; this would enable even small research teams to process extremely vast quantities of data.

It is really beyond the scope of this paper to summarize the current advances in analyzing all types of raw data, so instead the remainder of the paper focuses on a type of raw/unstructured data mostly analyzed in social sciences – text (in a narrow sense of the word). Given the depth and richness of detail that the current analysis tools can produce, the distinction between quantitative and qualitative text analysis has been abandoned in many fields and instead the analytical endeavor is referred to as natural language processing or NLP.

Even without ANNs, current NLP techniques can already yield impressive results.[14] In an extremely brief summary, NLP allows to perform text summarization tasks, sentiment analysis, and simple linear text classification. In text summarization tasks, various algorithms (TFIDF,[15] LDA,[16] etc.) are used to reduce either a single text or a corpus of texts to a small number of keywords that would best describe the content of the text/corpus. In sentiment analysis, a text is assigned a numeric value based on the strength of a sentiment expressed within. These can range from a simple assessment of whether a text is positive or negative to a more nuanced detection of various emotions in the text corpus.[17, 18]

---

14 To read more on the subject, see: Manning Christopher D., and Schütze H., *Foundations of Statistical Natural Language Processing*, Cambridge: The MIT Press, 1999.

15 "What Does tf-idf Mean?" <http://www.tfidf.com/>, 10 10 2018.

16 Blei D. M., Andrew Y. Ng, and Jordan M. I., "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, 2003, p. 993–1022.

17 Kouloumpis E., Wilson T., and Moore J., "Twitter Sentiment Analysis: The Good the Bad and the OMG!" *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

18 Nithyanand R., Schaffner B., and Gill P., "Online Political Discourse in the Trump Era," *arXiv preprint arXiv:1711.05303*, 2017.

Finally, text classification allows to divide a corpus of texts into a number of categories. This can be done in an either supervised or unsupervised manner – i.e., classification can either be performed based on patterns "learned" from the labelled training set or simply induced from the underlying data structure.

Such supervised classification has been applied in a multitude of projects involving both binary or multiclass classification. For instance, a binary classification was applied to track positive and negative mentions of politicians in Twitter[19] or in determining commenters' political views from their comments on the internet.[20]

Finally, the power of text classification was demonstrated in a number of experiments where an algorithmic classifier was used to analyze and structure unstructured text data. For instance, text classification was applied to convert unstructured interview transcripts into structured, survey-style datasets.[21] In other projects, text classifiers were used to analyze open-ended questions in population survey data.[22, 23]

Though the range of potential applications is quite impressive, more traditional NLP techniques for text classification still have certain limitations, especially when working with highly non-linear problems. Since roughly 2015, a lot of emphasis has been placed on

---

19 Saleiro P. et al., "Popmine: Tracking Political Opinion on the Web," *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, IEEE, 2015, p. 1521–1526.

20 Park S. et al., "The Politics of Comments: predicting Political Orientation of News Stories with Commenters' Sentiment Patterns," *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ACM, 2011, p. 113–122.

21 Giorgetti D., and Sebastiani F., "Automating Survey Coding by Multiclass Text Categorization Techniques," *Journal of the American Society for Information Science and Technology* 54 (14), 2003, p. 1269–1277.

22 Esuli A., and Sebastiani F., "Machines that learn How to code Open-Ended Survey Data," *International Journal of Market Research* 52 (6), 2010, p. 775–800.

23 Giorgetti D., Prodanof I., and Sebastiani F., "Automatic Coding of Open-Ended Questions Using Text Categorization Techniques," *Proceedings of the 4th International Conference of the Association for Survey Computing,* ASCIC 2003, p. 173–184.

detecting and countering deceptive content ("fake news") online, and many research efforts have demonstrated that it is possible to build algorithmic solutions to detect such "fake news" items. However, as the differences between "real" and "fake" news are often delicate and nuanced, the traditional NLP approaches did not deliver good results based on the text data alone.[24, 25] Yet it was demonstrated that ANN-based text classifiers can perform very well at detecting fake news. Due to their capacity to detect non-linear patterns in the data, in most cases, they significantly outperform more traditional alternatives in "fake news" detection tasks.[26, 27, 28]

Based on the overview above, we argue that ANN-based NLP techniques can be used to aid researchers in automating tasks that are very labor-intensive. ANNs could be used to automate the laborious pre-processing and coding of text data in preparing it for further analysis and thus bring significant value to social scientists. Furthermore, we would go as far as to argue that the capacity of ANNs to extract rich and nuanced features from the unstructured data make it possible to perform tasks typically associated with qualitative analysis at a sufficient scale, to be able to make generalizations for larger-groups or populations. In other words, using ANNs for certain questions would allow to bridge the gap between the quantitative and qualitative approaches and combine the best features from both worlds: richness of detail and a large-scope of analysis.

---

24 Conroy N. J., Rubin V. L., and Chen Y., "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology* 52 (1), 2015, p.1–4.

25 Tacchini E. et al., "Some Like It Hoax: Automated Fake News Detection in Social Networks," *arXiv preprint arXiv:1704.07506*, 2017.

26 Wang W. Y., "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2017.

27 Riedel B. et al., "A Simple but Tough-to-Beat Baseline for the Fake News Challenge Stance Detection Task," *arXiv preprint arXiv:1707.03264*, 2017.

28 Ruchansky N., Seo S., and Liu Y., "CSI: A Hybrid Deep Model for Fake News Detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, 2017, p. 797–806.

In the next section, we describe how we have applied ANN-based text models to identify texts distinguished by a highly subtle and complex feature set and to process an extremely large corpus of texts using minimal amounts of manual labor.

4. Case Study: Using ANN to Detect Innovation Mentions in Company Websites

4.1. Aims and Scope

Our team has recently undertaken a study of factors that contribute to innovation development in small and medium enterprises. To this end, we first had to acquire data on what share of enterprises are developing innovations and the intensity of their innovation development activities. Currently, the best source of data on innovation activities of enterprises is Eurostat's "Community Innovation Survey."[29] However, it has several significant flaws, namely the data-lag between the data collection and the publication of the dataset, which can go up to four years. This flaw meant that we could not use this data source for our analysis and had to acquire these data ourselves.

Kinne and Axenback[30] argue that company websites constitute a viable and useful data source on company innovation. We decided to base our data collection on their approach and acquire company innovation data from company websites. More specifically, we sought to determine whether a company has introduced a new product or service within the last 12 months. We carried out two rounds of data collection and compared the website text between the two measurements. If the new text described a new (previously unmentioned) product or was related to a new product launch, we counted that as an

[29] Eurostat: "Community Innovation Survey," <https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey>, 10 10 2018.

[30] Kinne J., and Axenbeck J., "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," *ZEW Discussion Papers*, No. 18–033, Mannheim, 2018.

innovation announcement. We then disambiguated these announcements to arrive at the "true" innovation count.

However, the number of such texts that we had to check for being related to new products/services or innovation announcements was too large (approx. 600 000) for our team to classify manually. Therefore, we needed a way of automating the text classification, which was a relatively complex task. Though for a human reader to identify an innovation related text is fairly simple, figuring out the exact judgment criteria for this task is difficult. Usually, there is a lot of variation in how the innovations are announced and what words/phrases are used to describe innovations/new products. Even if innovations are announced by similar actors and relate to similar developments, they may not share a single common feature. Table 1 contains two sample innovation announcements from pharmaceutical companies active in developing drugs to treat various cancers and related symptoms. While both texts undoubtedly introduce new product innovations, they have very few features (words and phrases) in common, demonstrating that innovations can be announced in very different styles.

Table 1. ***Sample innovation announcements.***

-COMPANY NAME-, a leader in the use of -THERAPHY NAME- to treat cancer, today announced that the U.S. Food and Drug Administration (FDA)'s Office of Orphan Products Development has granted orphan drug designation for the company's -THERAPHY NAME- targeting -ANTIGEN NAME- for the treatment of soft tissue sarcoma, a solid tumor cancer.	-PRDUCT NAME-, specifically designed by merging two well-established analgesics into a new co-crystal structure, was shown in a phase II clinical study to achieve effective pain relief at lower doses compared to -THERAPHY NAME- alone, and with an improvement in overall tolerability.

Source: created by the authors.

Overall, we collected data from 1301 company websites (see Table 2). After identifying the language of the website and keep-

ing only the English language content, this amounted to 567 distinct web pages. Out of that, 163 companies were randomly selected and their content labeled manually; this labelled sub sample amounted to 31 898 web pages. This labelled sub-sample was split into two parts: the training set (75%), on which the model was trained, and the test set (25%), on which the model performance was evaluated.

Table 2. ***Descriptive data on the whole dataset and labelled sub-sample.***

	Collected Data	Labelled Sub-sample
Companies	1 301	163
Web pages	567 554	31 898

Source: created by authors.

4.2. Methods and Results

Since identifying innovation-related content was a complex task that goes beyond simple keyword and/or word pattern matching, we looked for possible ways how such complex concepts could be detected automatically. This prompted us to consider various supervised machine learning algorithms and ANNs in particular, due to their ability to detect and recognize complex patterns.

As discussed above, supervised machine learning works by first training an algorithm on a labelled set of data (i.e., data for which both input and output values are known), a train-set, and then testing its performance on previously unseen labelled data, a test-set. Once the algorithm learns the relationships between the input data and output data based on the training set well enough to be able to accurately predict the test data, it can be deployed to infer the output characteristics on the rest of the data. In other words, machine learning models require manually analyzing and assigning labels to a sub-set of data, which is later split into training and test sets. An algorithm is trained on the training set, and its performance is evaluated on the test set. Once the performance is satisfactory, the algorithm starts assigning labels to previously unseen data on its own.

We sought to evaluate the performance of the ANN model not only by its absolute performance (i.e., overall precision, accuracy, and recall) but also by its relative performance – by benchmarking it to other commonly used machine learning models. We considered a number of options and algorithms for this benchmarking exercise. Namely, we tested the following algorithms: logistic regression, random forests, support vector machines from a commonly used "SKLearn" library,[31] and an MLP with an additional set of convolutional and embedding layers, similar to the one described in study by Kim.[32] These algorithms were shortlisted because they are the most commonly used to carry out various text classification tasks. The implementation of these algorithms was done in the Python (3.6) programming language.

Prior to using the models, we pre-processed the texts to in order to turn them into numerical matrixes, which can be analyzed and interpreted by these algorithms. In doing so, we followed conventional pre-processing steps:

- Striping punctuation marks;
- Converting to lowercase;
- Stop word removal – removing words that occur frequently but do convey any relevant information (e.g., "also," "already," "anyway," etc.);
- N-grams – identifying word sequences that are used together in the sample texts and converting them to a single word (e.g., "European," "union" → "european_union");
- Dictionary encoding – turning a natural language text into a sequence of numbers (vectors) based on a dictionary-schema.

See an example of these pre-processing techniques applied step-by-step on a quote from Shakespeare in Table 3.

---

31 Pedregosa F. et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12 (Oct), 2011, p. 2825–2830.

32 Kim Y., "Convolutional Neural Networks for Sentence Classification," *arXiv preprint arXiv:1408.5882*, 2014.

Table 3. ***Natural language pre-processing.***

Normal text	"A fool thinks himself to be wise, but a wise man knows himself to be a fool."
Tokens	'fool', 'thinks', 'be', 'wise', 'wise', 'man', 'knows', 'fool'
Dictionary	'fool': 1, 'thinks': 2, 'be' : 3, 'wise' : 4, 'man' : 5, 'knows' : 6
Vector	1, 2, 3, 4, 5, 6, 1

Source: created by the authors.

Once the vectors were created, we inputted them directly into the ANN model. However, such mode was not suitable for other models, so for other models, we converted the vectors to bag of words encoding.[33] The comparison of the algorithms' results is presented in Table 4.

Table 4. ***The performance of different machine learning algorithms on the test set.***

N: 163 Companies; 31 898 webpages Training set: 23 923 (75%); Test set: 7 975 (25%)		**Logistic regression**		**Random forest**		**Support Vector Machine**		**ANN**	
		Model prediction		Model prediction		Model prediction		Model prediction	
		Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.
Actual Value	Neg.	7 397	52	7 403	46	7 448	1	7 442	7
	Pos.	166	360	159	367	448	78	41	485
F1 score:		0.767		0.781		0.257		0.952	

Source: created by the authors.

The results in Table 5 show that the ANN model performed the best out of four algorithms. The ANN model has achieved a F1 score (a harmonic mean of accuracy and recall) of around 0.95. More im-

[33] Python Programming Language Blog, "Bag of Words" <https://pythonprogramminglanguage.com/bag-of-words/>, 10 10 2018.

portantly, it has identified the highest number of true positives out of the compared algorithms, which indicates that it has successfully managed to "learn" the feature patterns that distinguish the innovation announcements from other texts.

Although logistic regression and a random forest classifier also demonstrated good results, both of these algorithms had significant shares of false negatives, i.e., instances when the text actually contained a mention of the innovation that was not picked up by the model. Since innovation mentions occurred relatively rarely (only in about 7% of the websites), missing a significant portion of these mentions could be considered serious shortcomings in model performance. In this regard, the performance of the ANN model, though better than the alternatives, still had a significant false positive rate of about 9%.

After running the model to identify the innovation-related content in company websites, we compared the structural characteristics of the manually labelled and predicted samples, as summarized in Table 5.

Table 5. ***Structural characteristics of manually and automatically labelled sets.***

	Labelled sample	**Predictions in the overall population**
Share of innovative companies (with at least one innovation-related page in their domain)	66%	55%
Share of innovation content in company (share of innovation pages in the company domain)	10.9%	8%
N (companies; web pages)	163; 31 898	1 301; 567 554

Source: created by the authors.

As indicated in Table 6, the model predicts that the share of innovative enterprises and innovation-related content overall is smaller in the general population than in the manually labelled sample. How-

ever, after randomly selecting and inspecting a thousand positive and a thousand negative model predictions, as well as additional 50 companies, we concluded that the model performance in the general population has maintained roughly the same false negative rates as in the labelled sample. This leads to conclude that the observed differences between the manually labeled sample and the general population innovation rates result from the characteristics of the sample (it contained a slightly larger share of innovative companies) rather than the performance of the model.

4.3. Limitations

Our initial motivation to gather the data from company websites was related to the current limitations of the existing data sources, namely Eurostat's "Community Innovation Survey."[34] Our approach was generally a success, and the case study above serves to demonstrate that ANN algorithms could be of use to researchers attempting to undertake various large-scale projects, especially those which require assigning labels to large quantities of text data. However, a comparison with data from the Community Innovation Survey also helps to reveal the limitations of our approach.

Using ANNs has enabled our team to map out and label a vast amount of data, which would have been otherwise impossible. Furthermore, having a trained ANN model for labelling the data means that the subsequent data collection and labelling rounds can be highly automated, requiring minimal effort from the research team. However, despite these clear advantages, this has a major drawback – a lack of granularity/detail. While the Community Innovation Survey manages to distinguish between many different innovation activity types, we were able to achieve only a binary distinction between innovations and not (though the efforts to improve on this result are

[34] Eurostat, "Community Innovation Survey" <https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey>, 10 10 2018.

ongoing). This is partly connected to using websites as a data source, since the information there is limited to the messages the owners of the websites seek to communicate to particular target audiences. However, another limitation comes from the machine-learning model itself. Machine learning models are known to be sensitive to the class imbalance problem. That is, they tend to perform best when they have to distinguish between classes of roughly the same size, and they tend to perform worse when one of the classes is significantly smaller than the other(s). In the above example, innovation-related content constitutes less than 11% of all content, and developing a reliable model with such class imbalance is hard enough. Any attempt to introduce more classification categories would mean that classes would become even smaller, while complexity of the model would increase. Furthermore, the class imbalance would impede model performance with increasing severity. Therefore, while this approach is very well-suited for dichotomous large-scale classifications, more complex multi-class problems continue to remain difficult to tackle. On the technical side, ANN-based models should be able to solve most of these problems, but more elaborate classification schemes require significantly larger amounts of training data to deliver the desired performance.

Conclusions and Discussion

This article presented an overview of the main strengths and limitations of using artificial neural network algorithms in social science. These algorithms can identify and "learn" highly complex, non-linear patterns in data and, while doing so, they can solve certain problems that are beyond the capability of other algorithms used for similar tasks. This feature could make ANNs of very high value to social science, where the vast majority of relationships and underlying patterns in the data are highly non-linear. However, this boon comes at a cost: when using ANNs, the precise relationship between the input

and output variables becomes almost impossible to determine, and it is even difficult to know whether a particular input variable affects the outcome at all.

This feature makes ANNs of little use in answering the typical questions asked in social sciences and diminishes their utility in hypothesis testing. Nonetheless, we argue that social scientists can make use of ANN abilities by outsourcing laborious data coding and pre-processing tasks to these algorithms. Given enough training data, ANN-based models could perform complex data coding tasks and extract rich and detailed features from unstructured data sources. As such, they can allow to significantly scale up some of the qualitative research tasks, especially in the field of content or discourse analysis and deliver both analytic depth and scale, combining the best features from both qualitative and quantitative methodologies.

However, despite such clear advantages, these algorithms also suffer from several limitations. Though they offer a chance to identify complex patterns in vast amounts of data quickly, the complexity of the pattern is directly proportional to the amount of training data required for the model to achieve good results. In the case study example, the team had to manually analyze and code over 30 000 web pages until the model's performance was satisfactory. Furthermore, ANNs,, like other machine learning algorithms are sensitive to a whole range of problems, which could negatively impact the result. One of these problems – class imbalance – was already discussed in the previous section. Another notable problem is that these algorithms tend to inherit and replicate all the biases in the training data.[35] Therefore, prior to applying ANNs, researchers must study their data in great depth and address any biases therein, especially if the end product of their research will serve to inform decision-making in any sphere. Yet, we maintain that despite these limitations, ANN algo-

[35] Baer T. and Kamalnath V., "Controlling Machine-Learning Algorithms and Their Biases," *Mckinsey Blog*, <https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases>, 10 10 2018.

rithms have much to offer to the social science community by empowering small research teams to take on more ambitious projects, which would otherwise remain outside their capacity.

References

Baer Tobias and Vishnu Kamalnath, "Controlling Machine-Learning Algorithms and Their Biases," *Mckinsey Blog*, <https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases>, 10 10 2018.

Bansal S., "How Autoencoders Work – Understanding the Math and Implementation," <https://www.kaggle.com/shivamb/how-autoencoders-work-intro-and-usecases>, 10 10 2018.

Blei David M., Andrew Y. Ng, and Michael I., "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, 2003, p. 993–1022.

Conroy Niall J., Victoria L. Rubin, and Yimin Chen, "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology* 52 (1), 2015, p.1–4, <https://doi.org/10.1002/pra2.2015.145052010082>.

Crawford Kate, "The Hidden Biases in Big Data," <https://hbr.org/2013/04/the-hidden-biases-in-big-data>, 10 10 2018.

Esuli Andrea, and Fabrizio Sebastiani, "Machines That learn How to code Open-Ended Survey Data," *International Journal of Market Research* 52 (6), 2010, p. 775–800, <https://doi.org/10.2501/s147078531020165x>.

Eurostat, "Community Innovation Survey," <https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey>, 10 10 2018.

Giorgetti Daniela, and Fabrizio Sebastiani, "Automating Survey Coding by MultiClass Text Categorization Techniques," *Journal of the American Society for Information Science and Technology* 54 (14), 2003, p. 1269–1277, <https://doi.org/10.1002/asi.10335>.

Giorgetti Daniela, Irina Prodanof, and Fabrizio Sebastiani, "Automatic Coding of Open-Ended Questions using Text Categorization Techniques," *Proceedings of the 4th International Conference of the Association for Survey Computing,* ASCIC 2003, p. 173–184.

Janciauskas Marius, and Franklin Chang, "Input and Age-Dependent Variation in Second Language Learning: A Connectionist Account," *Cognitive science* 42 (S2), 2018, p. 519–554, <https://doi.org/10.1111/cogs.12519>.

Judge Timothy A., and Cable Daniel M., "The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model," *Journal of Applied Psychology*, 89 (3), 2004, p. 428, <https://doi.org/10.1037/0021-9010.89.3.428>.

Kim Y., "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

King Gary, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton: Princeton University Press, 1994, <https://doi.org/10.1007/978-3-531-90400-9_58>.

Kinne Jan, and Axenbeck Janna, "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," *ZEW Discussion Papers*, No. 18-033, Mannheim, 2018, <https://doi.org/10.2139/ssrn.3240470>.

Kouloumpis Efthymios, Theresa Wilson, and Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!" *Fifth International AAAI conference on Weblogs and Social Media*, 2011.

Manning Christopher D., and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge: The MIT Press, 1999.

Michigan Institute of Technology, "The Thinking Machine" (1961) – MIT Centennial Film", <https://techtv.mit.edu/videos/10268-the-thinking-machine-1961---mit-centennial-film>, 10 10 2018.

Mittelstadt Brent Daniel et al., "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3 (2), 2016.

Nils J. Nilson, *The Quest for Artificial Intelligence: The History of Ideas and Achievements*, Cambridge: Cambridge University Press, 2010.

Nithyanand Rishab, Brian Schaffner, and Phillipa Gill, "Online Political Discourse in the Trump Era," *arXiv preprint arXiv:1711.05303*, 2017.

Park Souneil et al., "The Politics of Comments: Predicting Political Orientation of News Stories with Commenters' Sentiment Patterns," *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ACM, 2011, p. 113–122, <https://doi.org/10.1145/1958824.1958842>.

Pedregosa Fabian et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12 (Oct), 2011, p. 2825–2830, <https://doi.org/10.1002/9781119557500.ch5>.

Python Programming Language Blog, "Bag of Words," <https://pythonprogramminglanguage.com/bag-of-words/>, 10 10 2018.

Riedel Benjamin et al., "A Simple but Tough-to-Beat Baseline for the Fake News Challenge Stance Detection Task," *arXiv preprint arXiv:1707.03264*, 2017.

Rosenblatt Frank, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* 65 (6), 1958, p. 386, <https://doi.org/10.1037/h0042519>.

Ruchansky Natali, Sungyong Seo, and Yan Liu, "CSI: A Hybrid Deep Model for Fake News Detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, 2017, p. 797–806, <https://doi.org/10.1145/3132847.3132877>.

Rumelhart David E., Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors," *Cognitive Modeling* 5 (3), 1988, p. 1.

Saleiro Pedro et al., "Popmine: Tracking Political Opinion on the Web," *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Comput-*

ing; Pervasive Intelligence and Computing, IEEE, 2015, p. 1521–1526, <https://doi.org/10.1109/cit/iucc/dasc/picom.2015.228>.

Tacchini Eugenio et al., "Some Like it Hoax: Automated Fake News Detection in Social Networks," *arXiv preprint arXiv:1704.07506*, 2017.

"TensorFlow Playground", <https://playground.tensorflow.org/>, 10 10 2018.

"What Does tf-idf Mean?", <http://www.tfidf.com/ >, 10 10 2018.

Wang William Yang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2017, <https://doi.org/10.18653/v1/p17-2067>.