Journalism Studies

ISSN: 1461-670X (Print) 1469-9699 (Online) Journal homepage: https://www.tandfonline.com/loi/rjos20

(What) Can Journalism Studies Learn from Supervised Machine Learning?

Frederik De Grove, Kristof Boghe & Lieven De Marez

To cite this article: Frederik De Grove, Kristof Boghe & Lieven De Marez (2020) (What) Can Journalism Studies Learn from Supervised Machine Learning?, Journalism Studies, 21:7, 912-927, DOI: 10.1080/1461670X.2020.1743737

To link to this article: https://doi.org/10.1080/1461670X.2020.1743737

Published online: 13 May 2020.

JOURNALISM STUDIES
2020, VOL. 21, NO. 7, 912–927
https://doi.org/10.1080/1461670X.2020.1743737

Routledge
Taylor & Francis Group

(What) Can Journalism Studies Learn from Supervised Machine Learning?

Frederik De Grove, Kristof Boghe and Lieven De Marez

Department of communication sciences, Ghent University - imec-mict-UGent, Gent, Belgium

ABSTRACT

In recent years, scholars have explored the applicability of supervised machine learning (SML) within journalism studies. While such computational methods could be of added value to the field, the rationale for employing these supervised models harbors some assumptions that deserve further inspection. This paper seeks to specify under which conditions SML could be useful for journalism scholars and where the field stands in exploiting its potential benefits. We start with an introduction to SML and give an overview of its applications within journalism studies. Next, we identify challenges for the field in its adoption of such techniques. These include overstating the time and financial savings caused by automatic coding, neglecting proper sampling methods, the danger of algorithmic determinism and the limited generalizability of predictive modeling across different domains, contexts and time periods. At the same time, we distinguish several opportunities. These include sharing classifiers, standardizing coding schemes and adopting general purpose techniques. Most importantly, in order for SML to contribute to the epistemological advancements in the field, SML could be used to explain how long-standing theories in journalism are changing. In turn, this might help us to disentangle the inner workings of our contemporary complex news ecosystem.

KEYWORDS

supervised machine learning; computational methods; big data; content analysis; digital methods; predictive modeling

Introduction

In line with other communication scientific fields, journalism studies has taken up the use of computational methods. Amongst those methods, several authors have focused their attention on the possibilities offered by supervised machine learning (SML). The main advantage of such methods lies in their predictive abilities. Broersma and Harbers (2018), for instance, discussed the viability of supervised learning in predicting news genres to study the transformation of news within a historical framework. Similarly, other journalism scholars used SML to predict news values (Burggraaff and Trilling 2017) or the prevalence of generic news frames (Burscher et al. 2014).

Whilst these studies shed light on important questions for journalism scholars, several issues remain to be dealt with. When looking at current studies on SML in the field, they lack a discussion on what machine learning actually does. This feeds the false notion that these algorithms operate inside a black box. Furthermore, several claims and practices

CONTACT Frederik De Grove frederik.degrove@ugent.be

surrounding the current use of SML need closer attention. A case in point is the belief that automated content analysis is to be preferred above manual content analysis (see e.g., Trilling, Tolochko, and Burscher 2017; Scharkow 2013). In addition, we need to examine the conditions under which machine learning can contribute to the field. To this day, researchers rarely contemplate on how and in which cases the predictive capabilities of SML might inform our theoretical understanding of journalism. Ultimately, the goal of this paper is to critically look at where the field of journalism studies stands on SML and how we might go forward.

We start with a non-technical introduction on SML by comparing it to inferential statistics as this is an approach most journalism scholars are familiar with. Next, we give a short overview of supervised learning within journalism studies, followed by a critical discussion of current practices and assumptions. We conclude our discussion with a reflection on how and in which circumstances SML can inform our theoretical understanding of journalism as such.

A Non-Technical Introduction of Supervised Machine Learning

Prediction is arguably the most important element that distinguishes SML from other approaches such as inferential statistics. In general, prediction refers to classification or regression problems. The former aims to determine in which category a specific dependent variable should be assigned (e.g., predicting whether a news article contains a human interest angle or not). The latter deals with dependent variables that are continuous (e.g., predicting the willingness to pay for online news content). To further our understanding of SML, we approach linear regression from the idea of inferential statistics. Reconsider the variable willingness to pay for online news content. From an inferential point of view, our main aim would be to test for a set of independent variables to what extent each variable contributes to explaining variation in our dependent variable. We would build a parsimonious model and assess performance in terms of variance explained, *p*-values and confidence intervals. This helps us to determine the relevance (significance) of our (unbiased) parameter estimates. In other words, this allows us to explain and understand with a certain degree of confidence why some people—on average—are more likely to pay for online news than others. For this reason, inferential statistics and explanatory modeling go hand in hand. Although we can compute prediction intervals and predict new data using this model, it would in practice seldom be used to such end. And unless the relation between the independent variables and the dependent one was really linear, our model performance would be subpar. This is where SML comes into play. Where inferential statistics focuses on building a parsimonious model that explains as much variance as possible, SML tests a multitude of models to find a model that performs best in predicting new data. Performance is mainly measured in how far away predictions for new data points are from the actual values. This raises two questions: how do we obtain a multitude of models and where do new data points come from? Comparing a multitude of models can be considered as another essential part of a SML approach. Even when using multiple linear regression, the number of models one can test grows rapidly as the number of independent variables increases. For example, a study using three independent variables (A, B, C) to predict willingness to pay allows for seven different models (A, B, C, AB, AC, BC, ABC). When using twelve independent variables, we could test 4095 different

models, not including interaction effects or polynomials for every variable. In practice, we use techniques such as ridge or lasso regression to shrink parameter estimates to efficiently control the number and impact of variables of a model. Yet, the central idea remains the same as if we would test 4095 models. SML aims to find the best model there is to make predictions. Now, where do we find and how do we test new data points exactly? Although several flavors of the same approach exist, the general idea is to split your data set in a training set, a validation set and a test set (Bishop 2006). The training set is used to build the models. In the example of multiple linear regression, one would run 4095 regressions on the training data thereby obtaining 4095 sets of parameters (one set for each model you run). Subsequently, the validation data set is used to compute for each model how well it performs. For regression, this amounts to computing for each data point in the validation set the difference between the expected outcome using the model parameters and the actual outcome in the data set on the dependent variable (prediction error for one data point). Performance for this model is generally computed by taking the average of the squared prediction errors. In reality, however, the validation data set is generally used to assess the effect of manipulating hyperparameters rather than the number of variables in the models. Once it is decided what the best performing model is, its performance is checked against the test set to inspect whether the model performs well with unseen data. The danger here is that the algorithm's flexibility during the validation phase is leveraged to explain the peculiarities of the data used to build the model. After all, if left unchecked, the model will add increasingly complex interactions so it can correctly predict all cases in the training and validation set. Consequentially, the model has little to no predictive power since it merely captures the noise instead of the true predictive signal present in the data. This would be a case of what they call model overfitting. Therefore, applying the model to a test set can be considered as a litmus test for the generalizability of the model.

SML models are supervised because the outcome they are supposed to predict is clearly defined. The model learns from examples in the training set and looks for features that are suitable for predicting the label or value for new data points. This stands in contrast with unsupervised machine learning (UML). In UML, the model learns from patterns and associations inherently present in the data. Instead of predicting a predefined category or continuous variable, the goal here is to let the categories arise from the data without much human intervention. Cluster analysis is a typical example here (de Mello and Ponti 2018). For example, one could perform a cluster analysis of news readers based on the articles they visit online. The algorithm could then point to different types of news readers who tend to manifest a distinctive reading habit. It is then up to the researcher to evaluate whether these clusters signify a meaningful categorization of the elements under study. Unlike these unsupervised methods, supervised models are able to learn from human-coded and thus interpreted data, which makes them suitable for measuring theoretically relevant and predefined constructs (Scharkow 2013). For example, if one wants to classify news content as being dramatized, one can train a model specifically for this purpose (Opperhuizen, Schouten, and Klijn 2019). The researcher could subsequently use these labeled texts to test a specific theory-driven hypothesis.

In sum, SML builds on the idea of predicting a dependent variable based on a set of example data. It does so by finding the best performing model amongst a set of

different models. Asking what journalism studies can learn from SML can thus be reformulated to asking whether and when journalism studies can learn from predictions. To answer this, we first turn to current research on SML in journalism studies.

Supervised Machine Learning in Journalism Studies

The application of SML within journalism studies is a relatively recent phenomenon. As Günther and Quandt (2015) noted, journalism scholars were—and to some extent still are—rather hesitant in adopting such quantitative predictive methods in their research. Yet, after some authors called for an adoption of computational techniques in journalism studies, these methods gained some traction among a relatively small subset of scholars (Flaounas et al., 2013; Scharkow 2013). Notwithstanding the progress being made, the uptake of supervised learning in the field has been fairly limited. Thus far, researchers employed these models exclusively as a technique for performing the automatic content analysis (ACA), which encompasses a plethora of techniques focused on leveraging computational power to analyze text corpora. In its simplest form, ACA includes the mere counting of certain keywords in a digital database (e.g., Qin 2015). However, as Deacon (2007) argues, this dictionary-based approach to text analysis fails to grasp the interpretative subtleties necessary for assigning meaning to texts. SML, then, is seen as an opportunity to grasp more latent and implicit variables in large news corpora that go beyond crude word counts (Boumans and Trilling 2016).

Given its current association with content analysis, supervised learning within journalism studies is applied as a content classification exercise. The goal of the learning algorithm is to infer the appropriate content label based on a set of manually-coded articles. Since the reliability of these predictions are heavily dependent on the quality of the training data, researchers tend to resort to classification schemes that are already well-established within journalism studies and are relatively undemanding to code. Moreover, there is a tendency to focus on binary classification schemes. Most prominent in this regard are models trained to detect the presence or absence of certain news values or biases, such as negativity, personalization and dramatization (Opperhuizen, Schouten, and Klijn 2019), the presence of conflict (Trilling, Tolochko, and Burscher 2017), infotainment (Burggraaff and Trilling 2017) or the gender of actors mentioned in a news article (Leavy 2018; Flaounas et al., 2013). Since news content implicates the analysis of textual data, most if not all machine learning applications in journalism studies start from a bag of words model (Zhang, Jin, and Zhou 2010). This technique encompasses a transformation of the raw textual information where each feature is represented as a combination of single or multiple words. These N-grams constitute the independent variables. The assumption here is that the presence or the combination of words are predictive for labeling the text. By far the most popular techniques in journalism studies for analyzing these models are Support Vector Machines (SVM) (Flaounas et al., 2013; Leavy 2018; Opperhuizen, Schouten, and Klijn 2019; Broersma and Harbers 2018) and Naive Bayes classifiers (Broersma and Harbers 2018; Burggraaff and Trilling 2017; Leavy 2018; Scharkow 2013). Both models are well-established, well-documented and commonly-used machine learning methods in the literature (James et al. 2013, 337). Their attractiveness lies in their relative simplicity and robustness, being able to perform well in a multitude of settings. Other techniques used in journalism studies are the

decision tree (Leavy 2018) or the related random forest (Broersma and Harbers 2018) classifier.

While most researchers opt for a single machine-learning technique, some use a combination of models to establish which technique performs best (Broersma and Harbers 2018; Leavy 2018) or even construct an ensemble model, which combines the predictive power of multiple supervised models and techniques at once (Burscher et al. 2014). Taken together, these supervised learning models serve as an extension of or even a substitution for human-coded content classifications. Ultimately, though, the human coder is still considered as the gold standard (Boumans and Trilling 2016). Yet, there is unanimous agreement and optimism among scholars that SML is well suited for automatic content analysis, yielding acceptable performance measures in most cases (Scharkow 2013). Implicit in most discussions on SML amongst researchers is the view that the loss in precision is outweighed by the benefit of automatically coding vast amounts of text.

Indeed, journalism scholars tend to apply these supervised models to an impressive amount of news articles ranging from around 2000 (Opperhuizen, Schouten, and Klijn 2019) to two and a half million (Flaounas et al., 2013) units. With samples this large, many argue that applying machine learning algorithms for coding means a significant reduction in time and financial costs (Boumans and Trilling 2016; Burscher et al. 2014). Human-coded training sets can be as small as 100 in some cases (Opperhuizen, Schouten, and Klijn 2019), although some employ rather impressive training sets of tens of thousands annotated articles (see for example Broersma and Harbers 2018).

This scaling up of content analysis in general is seen as a necessity within journalism studies for several reasons. For one, the increased availability of (big) data is seen by some as an argument in itself (Scharkow 2013; Broersma and Harbers 2018). More substantially, many refer to the allegedly inherent advantages of working with bigger samples. In this regard, one argument goes that machine learning allows the detection of unforeseen or more ambiguous patterns that go unnoticed when using qualitative methods (Boumans and Trilling 2016; Leavy 2018). Others make the case that big data sets render sampling methods redundant, since one can analyze the entire corpus without additional costs (Scharkow 2013). Related to this is the aspiration to establish stronger evidence for generalizing or extrapolating findings within journalism studies (Flaounas et al., 2013). The goal, then, is to establish macro societal-level patterns as described in Lazer et al. (2009) influential article on the promises of computational social sciences. Other scholars go beyond the perceived benefits of bigger samples and frame this scaling up of journalism studies as an opportunity to explore new methodological approaches, such as longitudinal content analysis (Opperhuizen, Schouten, and Klijn 2019; Broersma and Harbers 2018). Furthermore, the perceived cost reduction made possible by machine learning spurred others to propose an integration of content analyses with other methodological approaches (Burscher et al. 2014). This echoes earlier arguments for opening up framing research and, for example, combine content analysis of frames with survey research (Kinder 2007). Finally, others see the predictability of algorithms as an ideal way to increase the reliability of journalism research (Boumans and Trilling 2016, 17). The line of reasoning here is that researchers could share their trained classifiers. This would ensure that certain labels are operationalized in exactly the same manner, leaving no room for noise stemming from the variability between human coders.

A Critical Look at Current Practices and Assumptions

A common rationale behind adopting SML is the advantage provided by using larger sample sizes. There are several reasons to be critical of this assumption. First, very large samples run the danger to yield low data quality. In fact, there is an inherent tradeoff between automation and reliability (Mahrt and Scharkow 2013). As Scharkow (2013) noted, predictions generated by SML are about 20 percent less reliable than those based on smaller human-coded sample data. In addition, when using large samples for automatic coding but within an inferential framework it is important to point out that statistical power increases only marginally after reaching a certain threshold. Moreover, while bigger samples exhibit more statistical power, they also increase the danger of drawing exceedingly insignificant conclusions (Gigerenzer 2004). Furthermore, while it is true that one can apply a trained model to an unlimited amount of new data points, it is rarely considered that training and validating the model as such already necessitates a considerable amount of manually coded examples. For instance, Burscher, Vliegenthart, and de Vreese (2015) noted how the performance of a news topic classifier stabilizes only after incorporating around 2000 articles in the training and validation set. Not only does this coding exercise entail a costly operation in terms of human resources, but the training data necessary to train the model harbors plenty of statistical power on its own to establish relationships that are theoretically and practically relevant. When trying to bypass this coding exercise, only a few alternatives are available such as the use of open-source annotated datasets such as the Reuters corpus (Lewis et al. 2004). However, while these corpora harbor an impressive number of annotated articles, the labels attached to them are fairly limited (e.g., general topic classifications such as sports versus politics). Hence, such corpora are only useful to predict crude categorizations.

In addition, the advantage of working with bigger samples is likely overstated given that social science methodology is already well-developed in sampling theory (Riffe, Lacy, and Fico 2014). A well-thought out sampling design and procedure making use of a relatively small random sample tends to produce better results compared to a large non-random sample (Kaplan, Chambers, and Glasgow 2014). A case in point is the study by Burggraaff and Trilling (2017) where they trained a supervised model to conclude personalization is more prevalent in online news media than in traditional newspaper outlets. To obtain their results, a sample of almost 800.000 news articles was used. However, appropriate sampling methods and sizes specifically for gathering online news data are well-known (Hester and Dougall 2007). In fact, Schaudt and Carpenter (2009) successfully studied online news values by utilizing constructed week sampling methods. Hence, automatic coding comes with a significant cost, and it might be useful to consider the question when large sample sizes are necessary. A case in point where the theoretical added value could outweigh the methodological costs is research employing longitudinal or comparative designs. Indeed, such endeavors often require large sample sizes in order to discern trends over time or differences between (national) news industries or outlets (e.g., Opperhuizen, Schouten, and Klijn 2019).

Another argument to use SML is that training a classifier is a long-term investment since the model can be reused and shared among researchers in the field. In this case, the possible loss in data quality is compensated by increasing the scalability of the model. While

this holds true in theory, the fact remains that nearly all publications in the field employ a study-specific model. If models are reused at all, researchers recycle their own (see e.g., Trilling, Tolochko, and Burscher 2017). This is especially regrettable given that the sharing of models might be a viable contribution of SML to the field. For one thing, relying on established models would imply a standardization of coding practices, which is still a critical issue within journalism (Lovejoy et al. 2014). This could facilitate comparative journalism research or simplify the study of longitudinal trends across different studies. Moreover, standardization could curb the theoretical fragmentation of journalism studies. To this day, inconsistencies in operationalization often signify a more fundamental disagreement on the theoretical boundaries of certain constructs. For example, there are a multitude of interpretations on what a media frame actually constitutes in an ontological sense (Scheufele and Tewksbury 2006). Sharing SML models could motivate researchers to deep-dive into common assumptions and could subsequently unify the field around some shared theoretical foundations and key concepts.

Despite the promise of inter-study reliability, the question remains whether SML models are really able to capture the conceptual depth and nuance of their target variables. In essence, SML techniques are unable to grasp the nuances inherent to human meaning-making. Indeed, the performance of current SML techniques tends to be limited to clear-cut concepts. For instance, predicting whether or not an article should be considered as entertainment news can be achieved with acceptable precision (Burggraaff and Trilling 2017). Model performance becomes less evident when aiming to predict labels sensitive to contextual knowledge such as certain news factors (Scharkow 2013) or general news genres (Broersma and Harbers 2018). It becomes even more challenging when trying to go beyond relatively simple categorizations such as determining the main event described in a news article (Hamborg et al. 2018). Next to these conceptual limitations, the accuracy of algorithmic predictions tends to diminish as data complexity increases. Models trained on multinomial-classification schemes underperform when compared with binary classifiers (Herrera et al. 2016). Related to this is the fact that many machine learning algorithms suffer in accuracy when they need to predict less-prevalent categories within unbalanced classification schemes (Krawczyk 2016). All these considerations encourage the adoption of crude but unambiguous and balanced binary classification schemes to increase model performance. This is problematic in the sense that such analytical limitations inevitably steer the type of research questions being asked. This could be considered a form of algorithmic determinism. Let us illustrate this with research on news diversity. As most SML models tend to perform best with clear-cut binary categories, chances are high that news diversity is conceptualized as actor diversity in terms of gender (e.g., Leavy 2018) or as sentiment (e.g., Opperhuizen, Schouten, and Klijn 2019). As it is much harder to detect differences in opinions or viewpoints, studies on news diversity would tend to favor conceptualization that are manageable by SML models. In the long run, this runs the danger of narrowing the meaning of diversity itself.

Next to the threat of algorithmic determinism, the promises of increased scalability and reliability are potentially thwarted by the domain, context, and time dependency of many constructs relevant to journalism studies. Domain dependency refers to the idea that models tend to perform poorly outside the context which they were designed for (Joshi et al. 2012). Take for instance extracting sentiment, which is a goal shared by marketing researchers, literary scholars, political scientists and journalism scholars. Sharing sentiment

classifiers between academic fields might thus seem beneficial. However, research suggests that even straightforward dictionary-approaches are tailored towards specific domains (Soroka, Young, and Balmas 2015). As such, it seems inevitable to develop separate models sensitive to the language characteristic for the type of text under analysis. The issue of domain specificity holds for other types of supervised models as well. For example, Burscher, Vliegenthart, and de Vreese (2015) showed how a classifier trained on predicting policy issues in news texts performs poorly when applied to parliamentary questions. For the same reason, researchers should be cautious to apply commercially available cloud computing platforms such as Amazon Web Services (AWS), which includes off-the-shelf sentiment analysis like Amazon Comprehend. Context dependency implies that constructs vary in meaning between different news industries and outlets. A prototypical example of the former is how constructs are products of a specific socio-cultural context. For example, concepts such as tabloidization (Esser 1999) and objectivity (Esser and Umbricht 2014) are not universal, taking on distinct meanings in different national news industries (Fisher 2016). On a much smaller scale, unaccounted differences between news outlets might skew the performance of a certain classifier as well. For example, research uncovered substantial differences in readability and subjective word usage between news outlets (Flaounas et al., 2011). This begs the question whether, for example, features relevant for predicting the human-interest frame in *The Sun* (high readability, high subjectivity) are relevant for predicting the same kind of construct in *The Guardian* (low readability, low subjectivity). Outlet differences such as these might explain why the classifier of Burscher, Vliegenthart, and de Vreese (2015) performed substantially worse when it was applied to a newspaper not included in the training data. Finally, predictive models may perform worse over time because many constructs relevant to journalism are not time invariant. Take for instance the changes in political news, which gradually incorporated more opinionated elements throughout the twentieth century (Steele and Barnhurst 1996). Shifts such as these have ramifications for longitudinal predictive models. For example, in their study to predict news genres across a period of 120 years, Broersma and Harbers (2018) were forced to develop a separate model for each time period to reach an acceptable model performance.

All these contingencies should prompt us to carefully consider the *de facto* costs of training a model. SML not only needs a considerable amount of training data, but may also underperform in several domains, socio-cultural contexts, outlets or time periods. These dependencies are also indicative of an academic discipline which is deeply committed to a contextual and holistic approach towards its study subject (Carlson et al. 2018). As such, SML is hardly a cure-all for standardizing coding practices.

Appropriating SML for Journalism Studies

Up to this point, we have discussed how SML is being used within journalism studies. Or more precisely, in what forms SML is published in journals that are dedicated to journalism studies. However, the abundance and availability of online news data has led to a convergence between journalism studies and computer sciences that is largely taking place outside of the traditional venues for journalism research. Indeed, several authors have been working on predicting news-related variables such as news frames or bias (Jiang and Han 2019; Vasdev 2019; Grinberg 2018; Castillo et al. 2014; De Choudhury,

Diakopoulos, and Naaman 2012; Castillo, Mendoza, and Poblete 2011). Others have been looking at differentiating between opinions or facts (Vasdev 2019) or on classifying social media news sources (De Choudhury, Diakopoulos, and Naaman 2012). Predicting news popularity also attracted attention of a multitude of scholars (Wu and Shen 2015; Shreyas et al. 2016; Rizos, Papadopoulos, and Kompatsiaris 2016; Van Canneyt et al. 2018). A possible advantage of these studies lies in their use of more advanced data modeling techniques. In fact, these studies tend to incorporate SML algorithms that often outperform the prevalent SVM and Naïve Bayes applications found in journalism research. There is also a downside, however. These studies are first and foremost focused on the algorithm and its performance and less on how the predictions of this algorithm inform our understanding of journalism. This leads to a form of low-level empiricism that does not inform theory in a substantial manner. For example, when researching the perceived credibility of news, Castillo, Mendoza, and Poblete (2011) conclude that authors that have previously written a large number of messages are perceived as relatively more credible. Compare this to a study by Karlsson, Clerwall and Nord (2014) where a quantitative investigation into the predictors of news' trustworthiness is embedded within a broader theoretical framework of journalistic norms and the institutional legitimacy of journalism. This lack of theoretical interaction also leads to a shallow conceptual approach. For instance, in their recent review of automated media bias detection, Hamborg, Donnay, and Gipp (2018) remarked that the conceptual approach of the computer sciences towards media bias is highly superficial and does not exploit the profound theoretical understanding of the concept as construed by social scientists. In a sense, the news-related data that are employed in these studies are perfectly interchangeable with any other type of text and often little theoretical reflections are forwarded on how the model can help us to decipher the inner workings of journalism as a societal institution. The reason for this might be attributed to the fact that the computer scientist and journalist scholar are traditionally embedded within two different cultures of statistical modeling (Breiman 2001). Computer scientists are trained within the confines of an algorithmic culture, whilst the data modeling culture is the home turf of the social scientist. Both cultures tend to produce different kinds of knowledge. Whereas the former sees statistics as a tool for making accurate predictions, the latter sees statistics as a tool for uncovering causal relations between variables based on theoretical assumptions. Their different approach to theory also implies different standards of model evaluation. Whereas the data modeling culture aims for model parsimony and interpretability, the best-performing predictive model often belies straightforward interpretation. Contrary to this, the explanatory modeler aims for unbiased parameter estimates within a reductionist but transparent model of social reality to conclude whether a certain predictor has a significant impact on the outcome.

In order to use SML in a sustainable way, we argue that we need to make sure SML is integrated within the existing epistemological framework of journalism studies. As Steensen and Ahva (2015) have indicated in their review of the field, journalism publications increasingly adopt a theoretical lens in their aim to grasp the complexities of the news ecosystem. They see this as a sign that the discipline has entered an age of maturity. This should not be interpreted as a purist view on the field of journalism studies, which is an interdisciplinary field by nature (Zelizer 2004). However, outsourcing methodology runs the risk of hollowing out the theoretical and ontological foundations of the field. In this sense, we join the recent call of Margolin (2019) for a symbiosis between computational

and traditional methods within communication research in general. A true symbiotic relationship implies that computational techniques complement the existing methodological framework of the field.

Discussion: A Way Forward?

It is not unusual for new methodological and analytical approaches to find a way in through the fringes of a field and to be used as validation tools first. We hope to have described where the field stands and some of the challenges it faces. A much harder issue to deal with is how to go forward. How can SML drive conceptual advancements in the field of journalism studies? In what ways can we use the predictive power of SML so that it goes beyond mere validation exercises or algorithm evaluations?

The changing socio-technological environment has challenged journalism scholars to revise conceptual notions of news production through concepts such as liquid journalism (Deuze 2008), ambient journalism (Hermida 2010), network journalism (Heinrich 2011) and hybrid media (Chadwick 2013). This contemporary news ecology has also challenged a multitude of theories central to journalism such as agenda setting (McCombs and Shaw 1972) or gatekeeping (White 1950). Both theories blossomed during the previous century in a time when the media landscape was comparatively simple. Their assumptions and findings, however, do not easily generalize to the current media environment. Consequently, several scholars have tried to update these theories to the twenty-first century. Concepts such as secondary gatekeeping (Singer 2014) and gatewatching (Bruns 2005) are examples of such endeavors. We argue that finding a way to use SML for theory building might start from those theories that are most affected by the changes that typify our contemporary news ecology. An additional benefit in doing so would mean that when we better understand how these mid-level theories such as agenda setting or gatekeeping have changed, we would also gain insight in our fast-changing, complex media environment.

In order to make this more tangible, we illustrate how SML could help us to understand how the complex, hybrid, non-linear news flows have changed the idea of agenda setting. First of all, agenda setting implies a causal claim. Prediction and causal claims go hand in hand. SML could be used to build a model that accurately predicts the media agenda of tomorrow based on the agendas of the previous days or weeks. To do so, we would need a classifier that is able to determine what topics, issue or events belong together over different media formats and platforms. What is more, we would also need a classifier to determine the kind of actor that is producing information (e.g., media, public, politician). For both classifiers, we would greatly benefit from standardized coding practices as well as from methods that could go beyond domains and contexts. Once these classifiers have been built, we would need to construct a model that can deal with prediction and sequential data. This is where expertise from outside our field will be more than useful. A possible approach for this example could be the use of Long Short-Term Memory recurrent neural networks (Hochreiter and Schmidhuber 1997). The moment we have a model that successfully predicts the future agenda, we would need to dismantle that model to theorize about contemporary agenda setting. This might allow us to disentangle when and under what circumstance different actors and content can influence the future agenda. This could in turn enable us to better grasp our media ecology as we would be able to zoom in

on the complex interactions between media logics and social media logics. Of course, this is only one limited example and future research might start by considering those theories that would benefit most from using SML.

The previous example shows that in order for SML to contribute to theory, it will be important to reconcile prediction with explanation. There are several ways to deal with the perceived dichotomy between explanatory and predictive modeling. Most notable, we could explore ideas such as "white-box" (n.d.) or "interpretable" (Doshi-Velez and Kim 2017) machine learning. What these techniques have in common, is a desire to make machine learning applications more transparent to aid in meaning- and decision-making processes. For journalism studies, supervised learners could point researchers to unexpected relationships which in turn might inform theory (Shmueli 2010). From this perspective, the a-theoretical nature of the model is its biggest strength. It observes reality empirically without fitting the ambiguity of social reality to a reductionist and predefined model. Supervised learners could then serve as a tool to generate new hypotheses (Nelson 2017). For example, using rule mining or decision trees, it could become apparent that news articles classified as being biased tend to contain a subset of negative moral vocabulary. Such finding might inform or refine conceptual boundaries of the term news bias and lead to new hypotheses. The work of Leavy (2018) already captures the spirit of this explanatory technique.

Another way to explore what is happening when using SML is to look at the prediction of a specific case or subset of cases. Instead of trying to formulate a crude summarization of the model as a whole, specific predictions could be informative. One could review a random sample of predictions with their corresponding feature weights in order to evaluate the model as a whole. A recently developed and popular algorithm in this regard is the LIME-technique (Ribeiro, Singh, and Guestrin 2016). Another option might be to evaluate a specific subset of cases such as ambiguous classifications with low confidence or outlier-cases. The work of Chen et al. (2018) on detecting conceptual ambiguity could serve as a primer here. In their study, several coders labeled a set of tweets using a standardized coding scheme. They subsequently trained a separate model for each coder. Diverging predictions between the two (or more) models hint at conceptual ambiguity, inviting the researchers to probe the features driving these inconsistent predictions. This could point the coders to conceptual nuances and inform further theorizing-efforts.

Something that also deserves attention is the importance of overcoming domain and context dependency. Take for instance the idea of secondary gatekeeping. In order to understand how news spreads and who is doing it, we would need a model that is able to predict who the most important gatekeepers are. Similar to the agenda setting approach, this would include identifying the kind of users that interact with certain news content. In theory, the same algorithms to identify users and content that were built for understanding agenda setting could be used. Recent advances in machine learning might attenuate the domain and context dependency issue in this case. Multi-domain learners (MDL), for example, are capable of incorporating the influence of one or more specific domains into a single model (Joshi et al. 2012). A domain here constitutes one or multiple classifications that may or may not have an impact on the prediction at hand, such as the type of news outlet (e.g., social media content vs legacy media content). Instead of building a separate model for each domain, the MDL leverages the common predictive power of features shared across the different domains, minimizing

the training data needed. Another promising avenue when it comes to text classification is the application of pre-trained word embeddings (Rudkowsky et al. 2018). Embeddings take on a relational linguistic perspective by incorporating the co-occurrence of features in the same text, paragraph or even sentence in the model. In essence, words are represented in a vector space where similar words tend to cluster together. This does not only improve the overall accuracy of the prediction, but it could diminish the negative impact of context-dependent word usage as well. Specifically, its relational approach simplifies generalizations to features that were infrequent or even absent from the training data.

In sum, embedding SML into the epistemological framework of journalism studies might best be tackled by starting from the challenges that are typical for our contemporary media environment. By implementing SML when researching theories such as agenda setting, gatekeeping, news sourcing or news values, we might gain more insight into the changes these theories have undergone and more generally into our contemporary complex news system. A way forward would thus start from a theoretical perspective rather than a validation one. In addition, prediction should be supplemented with explanation. This could be achieved by using methods implicit in predictive models or by using other methods in the toolbox of journalism research.

Conclusion

We set out to understand where the field of journalism studies stands when it comes to SML and how we could move forward. We noted that the assumed advantages of large datasets, scalability and cost efficiency do not hold under all circumstances. When using SML as a straightforward classification tool, the benefits of scalability have to outweigh the costs. Depending on the goal of a research project, a carefully constructed sample might be more appropriate. Having said that, automated content classifiers could play a pivotal role in standardizing coding practices among journalist scholars. However, given the contextual sensitivity of journalism studies the generalizing capabilities of such classifiers might be limited. Previous research has clearly demonstrated that trained classifiers are prone to errors when applied outside their training context. Furthermore, the performance of these classifiers could be improved by using more complex techniques and models such as neural networks.

More importantly, in order for SML to have a sustainable future in the field, it has to contribute substantially to the epistemological framework of journalism studies as a discipline. This endeavor implies a symbiosis between computer sciences and social scientist cultures. Applications of SML in journalism studies are still dominated by an algorithmic culture, which is the archetypal approach to statistical modeling within the computer sciences. This has left the field with plenty of interesting attempts to predict all kinds of news-related variables. However, in order to move forward as a field, we need to consider how SML can be incorporated into the theoretical inquiries typical for the discipline. Potential venues here might lie in leveraging the predictive power of SML to grasp the complexities of the digital news ecosystem, the use of more sophisticated methods or in setting up research that reconciles predictive power with explanatory research. Of course, It goes without saying that SML or, by extension, computational methods are no catch-all solutions for all the questions and challenges that the field of journalism

studies knows today. SML is only one analytical approach and is in no way a replacement for the already rich toolbox we have at our disposal as journalism scholars. Arguably, integrating SML with well-known methods and approaches might yield the most satisfying results.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Lieven De Marez http://orcid.org/0000-0001-7716-4079

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Singapore: Springer.

Boumans, J. W., and D. Trilling. 2016. "Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism Scholars." *Digital Journalism* 4 (1): 8–23.

Breiman, L. 2001. "Statistical Modeling: The two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Broersma, M., and F. Harbers. 2018. "Exploring Machine Learning to Study the Long-Term Transformation of News: Digital Newspaper Archives, Journalism History, and Algorithmic Transparency." *Digital Journalism* 6 (9): 1150–1164.

Bruns, A. 2005. *Gatewatching: Collaborative Online News Production*. New York: Peter Lang.

Burggraaff, C., and D. Trilling. 2017. "Through a Different Gate: An Automated Content Analysis of how Online News and Print News Differ." *Journalism*. doi:10.1177/1464884917716699.

Burscher, B., D. Odijk, R. Vliegenthart, M. de Rijke, and C. H. de Vreese. 2014. "Teaching the Computer to Code Frames in News: Comparing two Supervised Machine Learning Approaches to Frame Analysis." *Communication Methods and Measures* 8 (3): 190–206.

Burscher, B., R. Vliegenthart, and C. H. de Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize Across Contexts?" *The ANNALS of the American Academy of Political and Social Science* 659 (1): 122–131.

Carlson, M., S. Robinson, S. C. Lewis, and D. A. Berkowitz. 2018. "Journalism Studies and its Core Commitments: The Making of a Communication Field." *Journal of Communication* 68 (1): 6–25.

Castillo, C., M. El-Haddad, J. Pfeffer, and M. Stempeck. 2014, February. "Characterizing the Life Cycle of Online News Stories Using Social Media Reactions." In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 211–223. Baltimore, Maryland: ACM.

Castillo, C., M. Mendoza, and B. Poblete. 2011, March. "Information credibility on twitter." In Proceedings of the 20th international conference on World wide web, 675–684. Hyderabad, India: ACM.

Chadwick, A. 2013. *The Hybrid Media System: Politics and Power*. Oxford, UK: Oxford University Press.

Chen, N.-C., M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon. 2018. "Using Machine Learning to Support Qualitative Coding in Social Science." *ACM Transactions on Interactive Intelligent Systems* 8 (2): 1–20. doi:10.1145/3232718.

Deacon, D. 2007. "Yesterday's Papers and Today's Technology." *European Journal of Communication* 22 (1): 5–25.

De Choudhury, M., N. Diakopoulos, and M. Naaman. 2012, February. "Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories." In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 241–244. ACM.

de Mello, R. F., and M. A. Ponti. 2018. *Machine Learning: A Practical Approach on the Statistical Learning Theory*. Cham, Switzerland: Springer.

Deutch, D., and N. Frost. (n.d.). Explaining White-box Classifications to Data Scientists. Retrieved from https://www.cs.tau.ac.il/~danielde/WhiteBoxFull.pdf.
Deuze, M. 2008. "The Changing Context of News Work: Liquid Journalism for a Monitorial Citizenry." *International Journal of Communication* 18 (2): 848–865.
Doshi-Velez, F., and B. Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.
Esser, F. 1999. "`Tabloidization' of News: A Comparative Analysis of Anglo-American and German Press Journalism." *European Journal of Communication* 14 (3): 291–324.
Esser, F., and A. Umbricht. 2014. "The Evolution of Objective and Interpretative Journalism in the Western Press: Comparing six News Systems Since the 1960s." *Journalism & Mass Communication Quarterly* 91 (2): 229–249.
Fisher, C. 2016. "The Advocacy Continuum: Towards a Theory of Advocacy in Journalism." *Journalism* 17 (6): 711–726.
Flaounas, I., O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. 2013. "Research Methods in the Age of Digital Journalism." *Digital Journalism* 1 (1): 102–116. doi:10.1080/21670811.2012.714928.
Gigerenzer, G. 2004. "Mindless Statistics." *The Journal of Socio-Economics* 33 (5): 587–606.
Grinberg, N. 2018. "Identifying Modes of User Engagement With Online News and Their Relationship to Information Gain in Text." In Proceedings of the 2018 World Wide Web Conference, 1745–1754. International World Wide Web Conferences Steering Committee, April.
Günther, E., and T. Quandt. 2015. "Word Counts and Topic Models." *Digital Journalism* 4 (1): 75–88. doi:10.1080/21670811.2015.1093270.
Hamborg, F., K. Donnay, and B. Gipp. 2018. "Automated Identification of Media Bias in News Articles: An Interdisciplinary Literature Review." *International Journal on Digital Libraries*. doi:10.1007/s00799-018-0261-y.
Hamborg, F., S. Lachnit, M. Schubotz, T. Hepp, and B. Gipp. 2018, March. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. In International Conference on Information, 356–366. Springer.
Heinrich, A. 2011. *Network Journalism: Journalistic Practice in Interactive Spheres*. New York: Routledge.
Hermida, A. 2010. "Twittering the News: The Emergence of Ambient Journalism." *Journalism Practice* 4 (3): 297–308.
Herrera, F., F. Charte, A. J. Rivera, and M. J. del Jesus. 2016. *Multilabel Classification: Problem Analysis, Metrics and Techniques*. doi:10.1007/978-3-319-41111-8
Hester, J. B., and E. Dougall. 2007. "The Efficiency of Constructed Week Sampling for Content Analysis of Online News." *Journalism & Mass Communication Quarterly* 84 (4): 811–824.
Hochreiter, S., and J. Schmidhuber. 1997. "LSTM Can Solve Hard Long Time Lag Problems." In Advances in neural information processing systems, 473–479.
James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.
Jiang, L., and E. H. Han. 2019. ModBot: Automatic Comments Moderation. https://drive.google.com/file/d/10bkVVPwqMolzEm_MSsmkecoWWq496pw7/.
Joshi, M., W. W. Cohen, M. Dredze, and C. P. Rosé. 2012. "Multi-Domain Learning: When Do Domains Matter?" In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 1302–1312. Association for Computational Linguistics, July.
Kaplan, R. M., D. A. Chambers, and R. E. Glasgow. 2014. "Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias." *Clinical and Translational Science* 7 (4): 342–346.
Karlsson, M., C. Clerwall, and L. Nord. 2014. "You Ain't Seen Nothing yet: Transparency's (Lack of) Effect on Source and Message Credibility." *Journalism Studies* 15 (5): 668–678.
Kinder, D. R. 2007. "Curmudgeonly Advice." *Journal of Communication* 57 (1): 155–162.
Krawczyk, B. 2016. "Learning From Imbalanced Data : Open Challenges and Future Directions." *Progress in Artificial Intelligence* 5 (4): 221–232.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann. 2009. "SOCIAL SCIENCE: Computational Social Science." *Science* 323 (5915): 721–723. doi:10.1126/science.1167742.

Leavy, S. 2018. "Uncovering Gender Bias in Newspaper Coverage of Irish Politicians Using Machine Learning." *Digital Scholarship in the Humanities* 34 (1): 48–63.

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li. 2004. "Rcv1: A new Benchmark Collection for Text Categorization Research." *Journal of Machine Learning Research* 5: 361–397.

Lovejoy, J., B. R. Watson, S. Lacy, and D. Riffe. 2014. "Assessing the Reporting of Reliability in Published Content Analyses: 1985-2010." *Communication Methods and Measures* 8 (3): 207–221.

Mahrt, M., and M. Scharkow. 2013. "The Value of Big Data in Digital Media Research." *Journal of Broadcasting & Electronic Media* 57 (1): 20–33.

Margolin, D. B. 2019. ": Computational Contributions: A Symbiotic Approach to Integrating big, Observational Data Studies Into the Communication Field." *Communication Methods and Measures*. doi:19.1080/19312458.2019.1639144.

McCombs, M. E., and D. L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly* 36 (2): 176–187.

Nelson, L. K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research*. doi:10.1177/0049124117729703.

Opperhuizen, A. E., K. Schouten, and E. H. Klijn. 2019. "Framing a Conflict! How Media Report on *Earthquake Risks Caused by gas Drilling*." *Journalism Studies* 20 (5): 717–734.

Qin, J. 2015. "Hero on Twitter, Traitor on News." *The International Journal of Press/Politics* 20 (2): 166–184.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016, August. "Why Should I Trust You?: Explaining the Predictions of any Classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144. ACM.

Riffe, D., S. Lacy, and F. Fico. 2014. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. 3rd ed. Mahwah: Routledge.

Rizos, G., S. Papadopoulos, and Y. Kompatsiaris. 2016. "Predicting News Popularity by Mining Online Discussions." In Proceedings of the 25th International Conference Companion on World Wide Web, 737–742. International World Wide Web Conferences Steering Committee, April.

Rudkowsky, E., M. Haselmayer, M. Wastian, M. Jenny, Š Emrich, and M. Sedlmair. 2018. "More Than Bags of Words : Sentiment Analysis with Word Embeddings." *Communication Methods and Measures* 12 (2–3): 140–157.

Scharkow, M. 2013. "Thematic Content Analysis Using Supervised Machine Learning: An Empirical Evaluation Using German Online News." *Quality and Quantity* 47 (2): 761–773.

Schaudt, S., and S. Carpenter. 2009. "The News That's fit to Click: An Analysis of Online News Values and Preferences Present in the Most-Viewed Stories on Azcentral.com." *Southwestern Mass Communication Journal* 24 (2): 17–26.

Scheufele, D. A., and D. Tewksbury. 2006. "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models." *Journal of Communication* 57 (1): 9–20.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Shreyas, R., D. M. Akshata, B. S. Mahanand, B. Shagun, and C. M. Abhishek. 2016. "Predicting Popularity of Online Articles Using Random Forest Regression." In 2016 s International Conference on Cognitive Computing and Information Processing (CCIP), 1–5. IEEE, August.

Singer, J. B. 2014. "User-generated Visibility: Secondary Gatekeeping in a Shared Media Space." *New Media & Society* 16 (1): 55–73.

Soroka, S., L. Young, and M. Balmas. 2015. "Bad News or mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 108–121.

Steele, C. A., and K. G. Barnhurst. 1996. "The Journalism of Opinion: Network News Coverage of U.S. Presidential Campaigns, 1968–1988." *Critical Studies in Mass Communication* 13 (3): 187–209.

Steensen, S., and L. Ahva. 2015. "Theories of Journalism in a Digital Age." *Journalism Practice* 9 (1): 1–18.

Trilling, D., P. Tolochko, and B. Burscher. 2017. "From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics." *Journalism & Mass Communication Quarterly* 94 (1): 38–60.

Van Canneyt, S., P. Leroux, B. Dhoedt, and T. Demeester. 2018. "Modeling and Predicting the Popularity of Online News Based on Temporal and Content-Related Features." *Multimedia Tools and Applications* 77 (1): 1409–1436.

Vasdev, S. 2019. "Can Machine Learning Help us Measure the Trustworthiness of News?" Presented at the Computation + Journalism Symposium, February. https://drive.google.com/file/d/1_Qrp2pGhl3eu7r-3BU6nwG_XN32jv7T_/view.

White, D. M. 1950. "The 'Gatekeeper': A Case Study in the Selection of News." *Journalism & Mass Communication Quarterly* 27 (4).

Wu, B., and H. Shen. 2015. "Analyzing and Predicting News Popularity on Twitter." *International Journal of Information Management* 35 (6): 702–711.

Zelizer, B. 2004. *Taking Journalism Seriously: News and the Academy*. Thousand Oaks: SAGE Publications.

Zhang, Y., R. Jin, and Z. H. Zhou. 2010. "Understanding bag-of-Words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics* 1 (1–4): 43–52.