Language Models in Sociological Research: An Application to Classifying Large Administrative Data and Measuring Religiosity*

Jeffrey L Jensen[†] Daniel Karell[‡] Cole Tanigawa-Lau[§]

Nizar Habash[¶] Mai Oudah[‖] Dhia Fairus shofia Fani[**]

June 29, 2021

Keywords:

Language model; Classification; Administrative data; Religiosity; Indonesia

*The first three authors equally share lead authorship and are listed alphabetically. We thank Blaine Robbins and our team of Indonesian research assistants.

[†]Division of Social Science, New York University Abu Dhabi, Abu Dhabi, UAE; jeffrey.jensen@nyu.edu

[‡]Corresponding Author. Department of Sociology, Yale University, 493 College Street, New Haven, Connecticut, 06511, USA; daniel.karell@yale.edu

[§]Department of Political Science, Stanford University, USA; coletl@stanford.edu

[¶]Division of Science, New York University Abu Dhabi, Abu Dhabi, UAE; nizar.habash@nyu.edu

[‖]Division of Science, New York University Abu Dhabi, Abu Dhabi, UAE; mai.oudah@nyu.edu

[**]Independent Researcher; dhiafsfani@gmail.com

Abstract

While computational methods have become widespread in the social sciences, probabilistic language models have been relatively underutilized. We introduce language models to a general social science readership. First, we offer an accessible explanation of language models, detailing how they predict a piece of language, such as a word or sentence, based on the linguistic context. Second, we apply language models in an illustrative analysis to demonstrate the mechanics of using these models in social science research. The example application employs language models to classify names in a large administrative database; the classifications are then used to measure a sociologically important phenomenon—the spatial variation of religiosity. This application highlights several advantages of language models, including their effectiveness in classifying text that contains variation around the base structures, as is often the case with localized naming conventions and dialects. We conclude by discussing language models' potential to contribute to sociological research beyond classification through their ability to generate language.

1. Introduction

Probabilistic language models are well-known in computational linguistics (Jurafsky and Martin 2019) and familiar to anyone using tools for speech recognition, automated spelling and grammatical correction in word processing applications, and the automatic completion of email messages. However, they remain relatively uncommon in social science research. For example, a search in November 2020 for the bigram term "language model" in articles published in *American Sociological Review*, *American Journal of Sociology*, *Social Forces*, *Sociological Methodology*, and *Sociological Methods & Research* did not return any references to probabilistic language models (hereafter, LMs).[1]

In this paper, we help introduce LMs to a general social science audience. To do so, we begin by offering a non-technical, accessible explanation of LMs. Then, we illustrate in detail the mechanics and utility of LMs through two related tasks common in social science research: first, the classification of a large volume of data, which, in our case, are names in an administrative database; second, the use of the classified data to measure a sociological concept, such as religiosity.[2]

We selected our example application for two related reasons. First, the increasing availability of large administrative databases provides promising new opportunities for quantitative analyses of data-scarce settings. For instance, in contexts in which individuals' names signal an identity marker (*e.g.*, race, ethnicity, religiosity), classifying the names in an administrative database into identity groups can shed light on a setting's group-level social composition (*e.g.*, Grofman and Garcia 2014; Harris 2015; Susewind 2015; Enos 2016; Imai and Khanna 2016; Hofstra and de Schipper 2018; Abramitzky et al. 2020). Yet, the full potential of administrative databases is sometimes attenuated by the information they contain. For example, if values of a variable, such as names, can appear in varying and unexpected ways—say, "Mustafa" and "Mustofa"—machine learning classification methods commonly used by social scientists may struggle to produce accurate classifications.

This challenge leads to our second reason for introducing LMs with a name-classification

task: it highlights a distinguishing feature of LMs. From data, they learn patterns of language use, such as how specific characters form types of words or particular words form types of phrases. Then, they estimate the probability of linguistic units (*e.g.*, words, phrases, or sentences) in a corpus. These probabilities can be used to predict the next chunk of language in a sequence, as in messaging apps' auto-complete function, as well as for classification. Indeed, our illustrative analysis shows that estimating the probabilities of linguistic units conditional on their context can effectively classify larger units, such as using characters to classify words, particularly when the larger units vary around a base form (*e.g.*, "Mustafa", "Mustofa").

The illustrative analysis begins with using LMs to classify all names from an Indonesian administrative database as having an Arabic linguistic origin or not. A large body of research has shown that individuals' names can signal meaningful demographic characteristics and features of identity, such as age, class, gender, and race (*e.g.*, Lieberson and Mikelson 1995; Bertrand and Mullainathan 2004; Sue and Telles 2007; Gerhards and Hans 2009; Olivetti and Paserman 2015; Goldstein and Stecklov 2016; Gaddis 2017a;b; Johfre 2020). This scholarship, in turn, draws on the insight that given names are selected through meaningful cultural processes unfolding in relation to social context (*e.g.*, Zelinsky 1970; Alford 1987; Lieberson and Bell 1992; Lieberson 2000; Elchardus and Siongers 2011; Seguin et al. 2021). For example, giving the name "Mary" to a child born in a small Christian community located in a predominantly Muslim country provides telling information about the parents', community's, and, likely, child's culture and beliefs. By the same logic, "Mary" provides less informative signals in an Anglophone, predominantly Christian context.

In our case, we analyze full names in an original database of administrative records comprising over a million residents in the Indonesian regency of Indramayu, an area of West Java that is home to approximately 1.7 million residents. In Indramayu, nearly all residents are culturally Muslim, but their names have origins in various languages, primarily Javanese and Indonesian (Bahasa). Hence, in this setting, the choice of parents to give their child an Arabic name, rather than a name from the regional Javanese or

Indonesian linguistic groups, signals that the parents held strong religious beliefs since Arabic is tightly coupled with Islam (Kuipers and Askuri 2017). Moreover, if the Arabic name invokes a name from the Quran, the Islamic holy book, it suggests a more intense degree of religiosity.

After classifying names as having Arabic or another origin (*e.g.*, Javanese, Indonesian)—and, if Arabic, quantifying how similar they are to Quranic names—we use the classifications to construct a fine-grained measure of geographic variation in religiosity. This extension beyond classification helps us fully demonstrate the usefulness of LMs for social science research. To create the religiosity measure, we link the classified names to residential addresses. We then use the spatially-located names to calculate the degree of Islamic religiosity across Indramayu at various levels of geographic aggregation. Finally, we conduct supplementary analyses using two other other databases, another administrative record capturing residents' religious behavior and a government registry of enrollment in religious schools, to evaluate the validity of the measure.

Our introduction and application of LMs have a number of other features worth highlighting. First, they expand social scientists' tool-kit for computational text analysis methods. While LMs fit computational social science's (largely) inductive approach of "sequential and iterative inferences" (Grimmer et al. 2021: 396; see also Nelson 2020; 2021), their advantages differ from more common methods. For example, whereas sociologists typically use another way to model language, word embeddings, to identify focal words' synonyms or syntactic variants and then use these to infer meaning (*e.g.*, Kozlowski et al. 2019; Stoltz and Taylor 2019; 2021), LMs provide the probabilities of linguistic units—which can be words, but also characters, phrases, or sentences—appearing in particular contexts. Social scientists can use these probabilities for classification, which can outperform familiar machine learning (ML) methods in certain cases, such as when using short texts, when linguistic units exhibit small variations around base forms (*e.g.*, when working with dialects and related languages), and when the subunits of larger units offer important information (*e.g.*, characters' relations to words, words' relations to phrases). Social scientists can also use LMs' estimated probabilities to *generate* language, or predict

a linguistic unit given a sequence of text. We do not examine this application here, but in the final section of this paper we briefly discuss the potential usefulness of LMs' generated language for sociological research, including vignette experiments and the formal analysis of culture.

Second, we contribute to the body of research showing how to exploit the increasing availability of voluminous highly detailed individual-level administrative data, which is driven in part by digitization efforts and digital record keeping. This is particularly important in data-scarce environments where few surveys exist and are expensive to administer, and government-provisioned public databases are limited or of poor quality. Specifically, the illustrative analysis shows that administrative databases can be sources of revealed preferences, and potentially lead to less measurement error than often occurs when surveys ask subjects questions of a sensitive nature.[3,4]

Another advantage of our application is that it scales very inexpensively. Surveys with representative samples are both expensive per observation and difficult to acquire at low levels of geographic aggregation (*e.g.*, county-level representative surveys in the United States). Given the increasing availability of administrative data across the world, we provide an economical method—especially as the number of observations rises—for measuring attitudes and beliefs at various levels of spatial or administrative aggregation. For example, our research setting, Indramayu Regency, has four sub-regency administrative levels; we can generate continuous measures of religiosity for each one after discerning the religiosity of individual names. This is noteworthy because the lowest administrative level, the *Rukun Tetangga*, has more than 10,000 units in Indramayu, each comprising, on average, approximately only 100 adult residents.

A third contribution is that our example application advances the literature analyzing and using names as signals of attitudes, beliefs, and non-lineage based identity. In particular, it demonstrates how to construct an "unobtrusive" (Webb et al. 2000) measure of the "sensitive and intricate" domain of religious belief and practice (Finke and Bader 2017: 3), which can complement standard survey-based indices.[5] Recent research has demonstrated the potential for names to capture religious group composition when each

name is already linked to a religion (Susewind 2015). Expanding this work, our analysis demonstrates how LMs are helpful when researchers are starting only with names—which unexpectedly vary in spelling—and knowledge of the social context.

Finally, our illustrative analysis demonstrates how names can be used in various kinds of social science studies. For instance, Arabic and Quranic names can help measure religiosity not only in Indonesia, but also in other non-Arabic Muslim populations (*e.g.*, Albania, Bangladesh, Bosnia, Malaysia, Pakistan, Turkey, the Caucasus, Central Asian countries). Indeed, our approach could be used when studying any population in which group members give their children distinctive names to signal membership (*e.g.*, Fryer and Levitt 2004). In addition, the analysis shows how culturally-significant artifacts containing "dictionaries" of names, such as spiritual texts and canonical works of literature, can act as tools for gleaning insights into names and attitudes, beliefs, and identity.

In the next section, we explain LMs and further discuss their applicability to social science research. In the subsequent section, we present our illustrative analysis. We begin this subsequent section by introducing the LMs that allowed us to classify names based on how they evoke religion. Then, we explain how we used these values to estimate geographic variation in the incidence of religiosity and religious intensity across each of these administrative districts in the Indramayu Regency. We end the paper with a brief review of a preliminary comparison of ML methods and LMs, as well as a discussion of LMs' usefulness for a range of sociological research.

2. Language models

Language models estimate the probability of a piece of text—characters, a word, or a string of words—given the context (Brants et al. 2007; Mikolov et al. 2010; Jurafsky and Martin 2019). Doing this can be relatively simple. Estimating the probability of a word, w, given a sequence of words, s, or $P(w|s)$, can be accomplished by counting the number of times w follows s in a corpus and dividing the sum by a count of s. However, relative frequency calculations quickly run into challenges. For instance, new sequences of words are continuously created so if we are interested in $P(w|s')$, there may be little or

no data on how w and s' are related. In addition, for large corpora, estimating the joint probability of a sequence of words, s'', can be resource intensive because the count of s'' instances must be divided by the number of all possible sequences of the same length as s''. A common strategy for handling these challenges is to use the Markov assumption to estimate the next character, word, or string of words given only the recent history (Jurafsky and Martin 2019). From here, LMs can become increasingly sophisticated, such as by building them on recurrent neural networks, but the fundamental logic remains the same: LMs estimate the probability of a piece of language based on the present pieces of language.[6]

Using LMs for classification, as we do in our example application, generally proceeds in two stages. First, as with most ML approaches, a training set is used to develop, or train, an LM for each class of interest. In our analysis, the training set is a sample of names from Indramayu's complete database of registered voters that a team of local research assistants labeled by class—whether a name is of Arabic origin (ARABIC) or of non-Arabic origin (OTHER). One of our LMs uses the ARABIC labeled subset to model the probability of an ARABIC name using its sequence of characters; the other LM does the same for the OTHER subset.

The second stage commonly involves using the LM's calculation of probability to determine a classification. The widespread practice is to use perplexity. For example, given an input name of unknown class, the LM trained for Arabic origin names examines the sequence of characters in the name and assigns a probability to whether that character sequence is "spelling out" a name belonging in the ARABIC class. Similarly, the LM trained for non-Arabic origin names examines the sequence of characters in the name and assigns a probability to whether that sequence is generating a name in the OTHER class. Then, each prediction's perplexity score—the inverse probability of, in this example, the name, normalized by the number of characters in the name—can be compared to discern the better prediction (Gamallo et al. 2017; Jauhiainen et al. 2017; Ramisch 2008; Vatanen et al. 2010). In other words, since minimizing the perplexity is the same as maximizing the probability, the LM assigning the lowest perplexity score to its prediction

is typically considered superior, and (in our case) the name is assigned the LM's class label, ARABIC or OTHER.[7]

There are two aspects of this process that warrant elaboration and highlight useful properties of LM approaches. First, LMs can be used for multi-class classification by employing multiple models, each focused on linguistic units' similarity with a distinct category. In studies with two classes, this can help address research design questions. Recall that in our illustrative example, we have two classes of interest, ARABIC and OTHER. With the LM approach, we could have built a single model that made predictions for the ARABIC class. Then, whenever the predictions' perplexity scores passed a threshold, we would have assigned the ARABIC label; when the scores were under the threshold, we would have assigned OTHER. Yet, while this design is intuitive, it raises questions. What determines a valid threshold? What if we had more than two classes? Are origins of names (and other features of linguistic units) best conceptualized as exclusive? That is, what if it is useful to know how much a name evokes an Arabic origin and how much it (concurrently) evokes another origin, which, in our case, would usually be Indonesian or Javanese?

One way to address such questions is to take advantage of the LM approach to build one model per class, as we do. This way, researchers would no longer be examining how probable a linguistic unit is to be in a particular class, but rather how probable the linguistic unit is to appear in each class. In our case, the multi-class ability of LMs would allow us to discern how similar a name is to names of Arabic origin and, at the same time, how similar the name is to names of Indonesian or Javanese origin. The design is extendable to any number of classes—a useful property for multi-ethnic or multi-racial settings—and helps avoid the selection of thresholds determining a dichotomous classification. Instead, researchers can simply select the strongest prediction by comparing perplexity scores, or they can incorporate information on the strength of each class's prediction into their analyses to shed light on membership in multiple or overlapping groups.

The second aspect warranting elaboration is that the estimation of probability and

perplexity-based classification can be combined with a simpler and more familiar dictionary technique. That is, classification can start by performing "look-up", or simply checking if the linguistic unit to be classified appears in the training data (*i.e.*, "in-vocabulary", or INV). If it does, it can be assigned its (usually human-generated) label. Then, for the remaining linguistic units—those not in the training set, or "out-of-vocabulary" (OOV)—the perplexity-based classification method can be used to impute labels. In addition to being conceptually straightforward, beginning with the dictionary technique is computationally efficient.

In sum, the LM approach to classification requires researchers to accomplish three tasks: (a) compile a list of labeled data; (b) use the labeled data in a dictionary approach; and (c) build and refine the models to classify unknown cases, which involves tuning hyperparameters, such as the n-gram length, or the number of characters in a sequence the models should consider to make the best predictions. The illustrative analysis demonstrates the details involved in accomplishing these tasks.

3. Example application: Language model classification

We demonstrate the mechanics and utility of LMs by using names from an administrative database to construct a measure of spatial variation in religiosity. Specifically, the example application draws on data from the Indonesian regency of Indramayu, classifies Indramayu residents' names as having an Arabic origin or an origin in another language with character n-gram LMs (as well as calculations of its "Quranic-ness"), and links the classifications to spatial locations.

This illustrated approach can be especially useful in environments in which government-provided data is scarce, administrative data is available, and names can provide important information on group compositions. Yet, there are other important context-specific factors that should be considered. The accuracy of our measure increases (or decreases) as: (a) the signal of the name is less (more) noisy, (b) the intergenerational transmission of these attributes (*e.g.*, beliefs, attitudes) rises (falls), and (c) the more (less) likely parents and children are to reside in the same relevant administrative area. Researchers should

consider these factors when determining whether it is appropriate to use this methodology to measure an attribute conveyed by given names, such as ethnicity, race, or religiosity.

3.1. Empirical case: Indramayu Regency, Indonesia

In Indonesia, a country of more than 260 million people and 17,000 islands, a regency is the second highest administrative level (below a province). Indramayu, part of West Java, the country's most populous province, is less than 200 kilometers from Jakarta (the capital), Bekasi, and Bandung, three of the four largest cities in Indonesia. Yet, despite its proximity to these major urban centers, it is poorer and more rural than the average regency in Indonesia. Since 99% of the regency is Muslim, our database on individuals residing in Indramayu consists nearly entirely of Indonesians identified as Muslims.[8]

While naming conventions in Indonesia vary by geography and ethnicity, most Indonesian names do not include a family name. Instead, full names, which commonly contain two to four single names (*e.g.*, "Abdul Hamid bin Mustofa"), usually include only *given* names, one of which is often derived from a name in a person's father's full name. Both these "inherited" given single names and the other given single names typically reflect the parents' cultural, ethnic, and religious identities, kinship, or geographic locations. For example, particular spellings, such as ending a single name with an "o", often indicates a Javanese male name (Uhlenbeck 1969). The small minority of Christians will often give their children names from the Bible (*e.g.*, Mary, John). Parents with a globalized orientation sometimes opt to give a name evocative of "Western" culture (Kuipers and Askuri 2017).

Despite the prevalence of Islam, especially in West Java, and the importance of names in Islam,[9] giving a child an Arabic or Quranic name is far from ubiquitous among Indonesian Muslims. Arabic names, however, are associated with Islam. They signal Muslim piety (Kuipers and Askuri 2017) in a country where ethnic and cultural identities remain strong and compete with religious identities (Pepinsky et al. 2018).[10] It is this signal given by Arabic (and Quranic) names that we use to construct a measure of religiosity.

3.2. Overview of data

Our data on Indramayu's residents come from the complete list of registered voters as of the 2015 elections, procured after a formal request to the Committee of Elections of Indramayu Regency.[11] In Indonesia, all individuals are automatically registered to vote if they are eligible to vote and have the national identity card, the Kartu Tanda Penduduk (KTP). As a result, the database omits three kinds of people who might have been residing in Indramayu. First, individuals who correctly do not have the KTP, such as children and foreigners.[12] Second, individuals who should have the KTP but do not. Consisting primarily of immobile and extremely impoverished residents, this group is estimated to be a very small minority. Not only is the KTP required for voting, it is critical to access all public services in Indonesia, including public housing and other anti-poverty programs (OECD 2019). Lastly, this database omits adults who live in Indramayu but are registered to vote in a different regency. We expect this population to also be very small because Indramayu experiences little in-migration.[13] Thus, we are confident that the registered voter database, containing the full person names for more than 1.3 million individuals, is a nearly complete record of adult residents of Indramayu.

In addition to individuals' full names, this registered voter database provides each person's residential street address. The address information allows us to place individuals within Indramayu's various sub-regency administrative boundaries (*i.e.*, district (*Kecamatan*), village (*Desa*), sub-village (*Rukun Warga*, or RW) and even residential block level (*Rukun Tetangga*, or RT)). The database includes 30 districts, 298 villages, 2,380 RWs and 10,192 RTs. The ultimate goal of our example analysis is to create a measure resembling those commonly used in social science research. To do so, we will obtain values of religiosity for each of these administrative levels using the LM-classified names and addresses of the residents.

3.3. Data selection and annotation

The voter registry contains 1,314,707 full names (513,527 unique full names), ranging in length from one single name to a sequence of 12 single names (*e.g.*, "Abdul", "Abdul Hamid bin Mustofa"). In order to increase the size of the data used to train and evaluate our LMs, we supplemented this main list of names with a roster containing 72,691 full names of expatriate Indramayu migrants provided by the regency's Department of Migrant Labor. The total corpus of names contained 568,195 unique full names comprising 1,381,923 non-unique single names (149,283 unique single names). The names are written in Roman alphabet, the standard script for administrative and commercial business in Indonesia.

We manually annotated three subsets of single names that differ in how they were sampled from the corpus of names. The first subset, "Train_{Rand}", contained a random sample of about 10,000 names from the list of unique single names. The second subset, "Train_{Freq}", was selected to maximize the coverage of the names. This subset comprised approximately the 10,000 most frequent unique single names, corresponding to 75% of all the non-unique single names. Finally, because we want to ultimately label *full names*, we randomly selected approximately 10,000 names from the list of unique full names (about 2%), then converted these names into a list of unique single names. We refer to this list as "Set_{Full}". Since some of the names in Set_{Full} appear in Train_{Rand} or Train_{Freq}, both of which we use for training, so Set_{Full} is not a "test set" in the usual ML sense. However, it captures a natural distribution of the phenomenon of interest, and we use it to evaluate our models (along with a second test set of OOV names) because it provides a good indicator of performance for the downstream measurement task. The second and third columns in Table 1 show how many unique names were in each dataset and how many (non-unique) times they appear in the corpus.

TABLE 1 ABOUT HERE

A team of five Indramayu residents coded the sampled names. All five were recent college graduates raised in Indramayu; three self-identified as Muslim and two self-identified

as Christian. The annotators assigned each single name to either the class ARABIC, if locals consider the name as having an Arabic origin, or OTHER, if the name has a non-Arabic origin, such as Indonesian or Javanese. After an initial round of annotation, the team members re-annotated 2,589 names that had received votes of three to two (without knowing the labels other team members had assigned). During this second round the annotators changed an average of 866 votes each. In the end, each of the five annotators had assigned a label that was in the majority for between 88% and 96% of the names, indicating a high level of agreement.

After the second round of annotation, each sampled single name received a final classification based on the majority vote. For example, if a name was coded as ARABIC three times and OTHER two times, it received the label ARABIC. Roughly 65% of single names were unanimously annotated OTHER and 3% were unanimously annotated ARABIC. Half of the ARABIC labels were assigned by a 4 to 1 vote. Nearly split voting (3-2) was rare, occurring for only 7% of all annotations. The last two columns in Table 1 show how many single names in each sample were assigned to each class.

3.4. Analyzing names and measuring religiosity

Our example application presents, by design, a challenge commonly faced by researchers using large databases: the manual coding or annotation of the database is not feasible in terms of time and cost. It also illustrates a second, more subtle, challenge. Namely, some data in large databases, especially administrative and natural-language databases, might vary in small, unexpected ways. These variations can be due to, for example, the same name being spelled a dozen different ways, word order changing by dialect, or to something as simple as misspellings. In these situations, LMs are especially useful for classification.

In this section, we present the underlying reasoning and four technical steps for moving from LM-based classification to constructing a measure of a sociological phenomenon. The first step entails a supervised classification of single names in our database using the training datasets and LMs. The second step involves using weighted voting combinations

of single-name classes. The third step calculates Quranic name similarity calculation using cross-linguistic name matching. The fourth step constructs the measure of spatial variation in religiosity. The final step evaluates the validity of this measure.

3.4.1. Related work and the advantages of language models

Both the substantial computational linguistics literature that undertakes language and dialect identification in text (*e.g.*, Gamallo et al. 2017; Jauhiainen et al. 2017; 2019; Ramisch 2008; Salameh et al. 2018; Vatanen et al. 2010; Zaidan and Callison-Burch 2014), as well as the smaller portion examining questions of name origins (Fu et al. 2010; Chen et al. 2006), largely rely on supervised ML techniques to learn from a training dataset how to classify a name. Moreover, they usually use as text features character n-grams, a representation of words in terms of sequences of subword characters.

Following the research using character n-grams in ML, we employ character n-gram LMs to accomplish the first step of our analysis, classifying all single names. These LMs estimate the probability of a sequence of characters given the context, which, in our case, is a preceding sequence of characters (Brants et al. 2007; Jurafsky and Martin 2019; Mikolov et al. 2010). Our use of character-level LMs is specifically motivated by, first, previous studies showing that LM inferences of probability are very helpful in the task of language and dialect identification (Gamallo et al. 2017; Vatanen et al. 2010; Ramisch 2008), and, second, Jauhiainen et al.'s (2017) findings that character-based calculations of probability are best when working with short texts, as we do when classifying single names.

We are further motivated to use character-level LMs by how they address variation in naming practices. Indonesians, like any other society, have developed numerous spelling variations of names. A name of Arabic origin, for instance, may appear in Indramayu with several different spellings, each differing by only one or two characters. Research in computational linguistics indicates that because character-level LMs leverage subword information, they are particularly useful at capturing local variations of base word structures.[14] For example, character-level LMs, unlike word-level models, do not face the

challenge of OOV tokens—units that do not appear in the training data—which often occurs when analyzing non-formalized areas of language, such as naming practices and dialects, at the level or words or phrases (Habash et al. 2012).

3.4.2. Step 1: Classifying single names

Before training the LMs, we merged Train_{Rand} and Train_{Freq} to create a single training dataset ("Training"). We kept Set_{Full} separate to use for optimizing the choice of hyperparameters and evaluating the system's performance. Table 2 presents the number of (non-unique) times single names in Training and Set_{Full} that occur in the corpus. The percentages of ARABIC names in Training and Set_{Full} are comparable at around 20%, but the ratios of ARABIC to OTHER labels within the datasets are unbalanced. The imbalance suggests that metrics like precision and recall should be used to understand the effect of different hyperparameters and model choices, which we explain in more detail below. Table 2 also includes the details of "Test", which is the portion of Set_{Full} not found in Training (*i.e.*, OOV names), and is analogous to the "test sets" used in common ML classification approaches. The size of Test is about 23% of Set_{Full}.

TABLE 2 ABOUT HERE

For our baseline model, we label the names in Set_{Full} in two steps. First, if a name appears in the training data, we assign it its label. Second, we assign all remaining names (*i.e.*, those not in Training) the majority class OTHER, or non-Arabic. As a result, none of the Arabic-origin names in Set_{Full} that are OOV (*i.e.*, in Test) would be identified. As we show below, this "majority baseline model" performed rather strongly on Set_{Full} in terms of evaluation metrics that ignore the imbalanced nature of the data since the majority class is prevalent.

We built our LMs using the SRILM Toolkit, a publicly available collection of C++ libraries and programs that provide LM infrastructure (Stolcke 2002). (Python bindings are available for SRILM, and the Natural Language Toolkit (NLTK) can be used to build n-gram LMs.[15].) Then, we focused on determining the optimal number of characters, or n-gram window, the models should use when making predictions. We began by evaluating

the results of a character window of one and systematically moved up to an n-gram window of 10.[16] We also considered a number of discounting techniques provided by the SRILM toolkit for modeling probability.

We evaluated our models and the various hyperparameter settings using a number of standard metrics (Power 2011). The first metric was precision, or the ratio of correct class predictions to the total number of class predictions. We report ARABIC precision, OTHER precision, and average precision of the two classes. The second was recall, or the ratio of correct class predictions to the total number of observed class names. We report ARABIC recall, OTHER recall, and average recall of the two classes. Third was F-score, or the harmonic mean of the precision and recall. We report on ARABIC F-score, OTHER F-score, and average F-score of the two classes. Finally, we also used accuracy, or the ratio of correctly predicted classifications to the total number of observed classifications.

Table 3 presents the results of the majority baseline model and the best LM over Set_{Full} and its OOV subset, Test. We see that on Set_{Full}, the accuracy and average F-score of the majority baseline model and our best LM are almost the same. Average precision is higher for the baseline, while average recall is higher for the LM. These first results appear to be a simple trade-off of average precision and average recall.

However, when we examine the LM's ARABIC metrics, we see a major increase in ARABIC recall from 86% to 97% paired with a drop comparable in magnitude for ARABIC precision. The effect on the majority class OTHER is less intense. These latter results indicate that the LM model is successfully identifying the majority of observed Arabic-origin names at the cost of misclassifying some non-Arabic names as ARABIC. Furthermore, the LM results are significantly better than the majority baseline model results when evaluating with the Test dataset. In seven out of the ten metrics we use, the LM is superior to the baseline when using Test, including when classifying ARABIC names, the class of primarily interest. These latter results show that the best LM performs well on the data representing the full names in the administrative database, Set_{Full}, and that it does so by correctly predicting some OOV Arabic names (compared to the baseline which predicts no OOV Arabic names).

TABLE 3 ABOUT HERE

The results indicate that the LM approach provides good overall performance while ensuring that we are capturing many more Arabic names relative to the baseline. We thus selected the best performing n-gram window specification—a trigram model with Witten-Bell discounting—and trained a final LM on all the annotated data (*i.e.*, a merge of Training and Set_{Full}). We then used this model to classify the remaining single names as originating from Arabic or not. Of the 1,381,923 (non-unique) single names in our data, we labeled 214,464 (15.5%) as ARABIC and 1,167,459 as OTHER.

3.4.3. Step 2: Full name classification

Recall that single names combine to form full names. In our application's second step, we leveraged this fact by using the relatively small number of manually coded single names, and the LM-classified remaining single names, to gain insight into numerous full names. Not only was this strategy efficient, but it also recognized the social reality that individuals' preferences, attitudes, and identities—whether signaled in names or not—often comprise a mix of cultural elements and practices. In our case, this mixing manifests as full names that might have some single names representing non-religious Indonesian heritage and other single names reflecting an Islamic faith. That said, the choice of variable type as categorical or continuous, and threshold(s) that place names into any categories, should be both theoretically-driven and consistent with the context. In our case, both the social science literature (Lieberson and Mikelson 1995; Sue and Telles 2007) and Indonesian-specific work (Kuipers and Askuri 2017; Llewellyn 2018) show that names are used to signal membership (or not) in a group, which reflects a binary, if noisy, distinction, rather than a continuous one. For example, in our case, an individual with an Arabic name is among the pious.

We used a majority rule to label each full name in the voter registry based on single names' classifications. For example, if a full name with three single names had two single names of Arabic origin, the full name was classified as ARABIC. In a context in which more than 99% of the population is Muslim, a threshold of only one Arabic single name

would likely label too many people as strongly religious. Furthermore, since the setting is *not* Arab—Indonesian culture is predominant, after all—requiring all single names to be Arabic is too stringent. If full names had equal numbers Arabic and non-Arabic single names, we summed the perplexity scores across the ARABIC single names and the perplexity scores across the OTHER single names, and then assigned the class with the lower total score. For instance, the name "Eni Sukarni bin Muhamad", which has two non-Arabic names, (*i.e.*, "Eni" and "Sukarni") and two Arabic-origin names (*i.e.*, "bin" and "Muhamad") would be classified as OTHER because the models have learned to judge the two non-Arabic names as more strongly not Arabic than the Arabic names as being Arabic. This technique selected the classification that had the better overall predictions. We labeled a total of 47,988 unique full names as ARABIC (8.4% of the registry's unique full names) and 465,539 unique full names as OTHER. Future research building on our approach could use each full name's mix of single name types to assign a continuous value rather than a categorical one, if appropriate for the study context.

3.4.4. Step 3: Quranic name similarity

The third step of our illustrative analysis measured how evocative full names are of proper names in the Quran. In addition to demonstrating the flexibility of LMs, we interpret this "Quranic-ness," or the extent to which individual full names contain Arabic *and* Quranic single names, as an indication of religious intensity. After all, the full names with Quranic-like names comprise terms taken from Islam's core spiritual text.

To calculate full names' Quranic-ness, we first computed the similarity between each ARABIC single name in each full name and each of the 98 person names mentioned in the Quran[17] using Freeman distance (Freeman et al. 2006). Comparing names written in the Roman alphabet to names originally written in Arabic is difficult because of varying transliteration conventions, resulting in a range of character representations for each phoneme. This problem is compounded by the fact that Arabic script typically does not represent short vowels. Freeman distance offers a solution to the challenge of matching Romanized and Arabic names primarily by allowing an expanded number of

cross-language matches for many characters (Freeman et al. 2006). Doing so has been found to result in better matching performance across several commonly used string-matching techniques (Freeman et al. 2006; Habash 2008). Freeman scores range from zero to one, with one being a perfect match.

After calculating the distance between each ARABIC single name and the Quran's names, we averaged the highest distance values for each full name's terms over the total number of single names in the full name. For example, if a full name had three ARABIC terms with highest distance values of 0.6, 0.7, and 0.8, the full name's overall similarity to Quranic names would be 0.7. Non-Arabic single names in a full name with ARABIC single names contributed values of zero to the average; full names with no ARABIC single names were scored as having zero similarity to Quranic names. The end result was a continuous value of full names' Quranic-ness, which we understood as a signal of religious intensity—and which, because of our approach, could be measured across Indramayu at different levels of geographic resolution.

3.4.5. Step 4: Measuring spatial variation in religiosity

The fourth step of our illustrative analysis shows how LM-powered analyses can provide a foundation for common aspects of social science research. We used the LM classifications and Quranic scores to construct two name-derived variables measuring religiosity at different levels of spatial aggregation. To do so, we linked the results of the preceding steps to spatial administrative units using residents' home addresses, which were recorded in the voter registry. The first variable captures the proportion of registered voters with a full name labeled as ARABIC in a given area. The second comprises the mean Quranic score of all full names in a given area.

At the RT level, the smallest administrative unit, our two name-derived variables correlate at $r = 0.85$. The correlation rises above 0.90 at the village and district levels. Figure 1 shows the variation across districts (left, N=30) and villages (right, N=298) in the proportion of all resident voters with an Arabic full name. The values of this variable ranges from zero to 25%. Figure 2 shows the values of the mean Quranic similarity

variable, which range from zero to 20%, for the same areas.

FIGURES 1 AND 2 ABOUT HERE

Figure 3 presents frequency plots showing the proportion of individuals with Arabic full names in the villages (top left), RWs (top right), and RTs (bottom left) of Indramayu regency. Figure 4 shows the same frequency plots for areas' mean Quranic similarity scores.[18] As is evident, the incidence of individuals with an Arabic name is highly concentrated within each administrative unit, suggesting that religious Muslims, and the more pious among them, are highly spatially concentrated in this regency. In addition, both sets of plots show that the frequency of Arabic and Quranic naming practices exhibit substantial skew at the individual level, and that this skew attenuates at higher levels of aggregation.

FIGURES 3 AND 4 ABOUT HERE

3.4.6. Step 5: Evaluating validity

In the fifth and final step, we evaluated the validity of both LM-based measures using criterion validity tests. We present this portion of the illustrative analysis in Appendix B, and only summarize it here.

We located two village-level measures that should be related to spatial variation in religiosity. First, we found the number of residents in a village who were registered to attend the Hajj, which we normalized by village adult population. Second, we found the number of students enrolled in Islamic schools, which we divided by the village's total school enrollment. As we discuss in Appendix B, each validity measure has limitations, which constrained our ability to test the accuracy of our LM-based measures (*i.e.*, the convergent validity). Yet we were able to test whether these variables are are related in the ways we expect (*i.e.*, criterion validity).[19] Using a variety of methods, from simple bivarate plots to tobit models (to account for overdispersion in our validation variables), we found evidence that the validation variables are correlated to our LM variables.

Our validation exercise leads us to offer some suggestions for researchers using LMs for the specific task of classification in data-scarce environments. To the extent possible, existing data should be used to evaluate the direction and accuracy of any measure based on LM classifications. However, in data-scarce settings, such as Indramayu, valid constructs in which to test the accuracy of the LM measure may be unavailable. If existing data allow it, researchers should at least test the criterion validity of the LM measure (*i.e.*, does the LM measure relate to real world data in ways we would expect?). Finally, researchers should clearly state if their measure meets a particular validity check, and the limitations of their ability to assess various evaluations of validity.

4. Discussion and conclusion

We have introduced and demonstrated the usefulness of probabilistic LMs to social scientists. To so so, we offered a non-technical explanation of LMs and an illustration of how LMs can be used in social science research. The illustrative analysis highlighted some advantages LMs have for the automated classification of text. Among these are the calculation of linguistic units' probability of appearing in multiple classes and the fact that LMs can leverage information at the level of text below the primary level of interest, such as characters when words are of interest or words when sentences are of interest. The former feature can be helpful when social scientists are researching individuals' simultaneous membership in multiple groups (see Nelson 2021). For example, a study could examine how similarly an individual's social media posts concurrently align with the discourse of multiple online communities. The latter feature can be crucial when classifying text that contains variation around the base forms, as is often the case with localized naming conventions and dialects (Salameh et al. 2018). Thus, we see LMs as an important addition to the computational approaches available for estimating variation in difficult-to-measure quantities.

Our application of LMs additionally raised two questions about using LMs for common classification tasks. First, how exactly do LMs compare to more familiar ML methods? The answer depends on contexts, datasets, and research questions. However, a clear next

step to advance the use of LMs in social science is to systematically compare LM and ML performance across common use-cases. We have conducted a preliminary comparison using our data; we report our findings and insights in Appendix C. In brief, our analysis shows that LMs are likely to be worth the effort if researchers are specifically interested in the minority class (in our case, Arabic-origin names) and are unable to obtain a large amount of training data.

Second, how should researchers select among different LMs that can be used to construct a measure? We suggest that researchers consider what the measure will be used for; the ultimate application of the measure will likely guide which metric to use when choosing an LM. For example, if the measure will inform governmental polices on the provision of resources to an under-served population, it could be reasonable to use the metric of recall. Maximizing recall, or the identification of members of the population, would help ensure that all its members are counted. Precision (and, by extension, F-score) might be less important: imperfect precision could lead to some resources being provided to members of other groups, which may not be as bad of an outcome as missing members of the under-served population (although perhaps inefficient).

The illustrative analysis also offered a contribution to the sociological study of religion: it demonstrated an inexpensive and unobtrusive way to measure religiosity in a large population across a large geographic area, and at different levels of spatial aggregation. While administrative data can be costly to annotate, the approach we presented scales extremely inexpensively; only a few thousand annotations can be used to classify datasets of 100,000, 1,000,000, or 10,000,000 observations. Our example thus not only applies to the empirical study of religiosity, but also to any setting in which researchers want to unobtrusively measure attitudes and beliefs but surveys are prohibitively expensive and high-quality government provisioned data are scarce.

We close by emphasizing that LMs have the potential to contribute to social science research beyond classification. We are especially optimistic that their *generative* nature can be of great value. In generation tasks, LMs identify the best piece of language to come next in a linguistic sequence. This can be thought of as ranking all the possible

words in a language and selecting the word (or character, phrase, or sentence) that gives a linguistic sequence the highest probability of appearing in a corpus. Analyzing the set of "best guesses"—the linguistic units with probabilities higher than a particular threshold, or reasonable alternatives to one another—could additionally be useful and is a straightforward extension of using the single best guess.

Social scientists can potentially use LMs' generated language in a variety of applications. For example, word-level LMs could suggest names that are common in a given corpus for use in vignette experiments (*i.e.*, the most probable names). However, because names can conjure nonrandom ideas in experimental subjects, which could confound the results (Johfre 2020), a character-level LM could be used to generate (novel) names that seem natural in the context but do not signal any particular meaning. In studies of how robots' participation in conversations between people affect the human participants (*e.g.*, Traeger et al. 2020), such as the interjection of "bots" into social media threads, researchers could use phrase- or sentence-level LMs to generate the robots' contribution to the discussion. They could even build variants of these LMs to produce different styles of robot contributions as experimental conditions.

In addition, sociologists adopting the formal approach to studying culture (Edelmann and Mohr 2018) could complement the current workhorse method, word embeddings (*e.g.*, Kozlowski et al. 2019; Stoltz and Taylor 2019; 2021), with LMs. For example, instead of using LMs to simply select the best word to come next in a sequence, researchers could examine the set of words with high probabilities of coming next—say, the five "best guesses"—as signals of meaning and polysemy. This would be a probabilistic alternative to using embedding models to study meaning, as well as a way to instill confidence in embedding models' identification of synonyms or syntactic variants. LMs' predictions also have the potential to address one of Biernacke's (2012) main critiques of the formal approach to culture—that it struggles to analyze what was left unsaid. To identify moments of "unsaid" expression and meaning, sociologists of culture could use LMs to predict the best next word in a sequence, then compare the prediction to either what was actually said in the observed discourse or to an observed silence. The gap between an LM's

predicted word—what *could* or *should* have been said—and the observed word or silence might be indicative of "unsaid" meaning. In fact, constructing alternative, "unsaid" versions of text data could result in more ambitious projects: researchers can examine why the LM-generated, alternative versions of text did *not* occur to support efforts at shedding new, surprising light on a discourse and cultural system, and subsequent theoretical development (see Tavory and Timmermans 2014; Karell and Freedman 2019; Brandt and Timmermans 2021). In sum, while we have focused on how LMs can help improve the relatively common social science tasks of classification and measurement, we believe LMs are an exciting general addition to the computational social science suite of methods and are poised to open new avenues of research that can make use of generated language.

Notes

[1]The search did identify one article that made use of the term "language model", but the authors were referring to a conceptual representation of language use rather than to formal probabilistic LMs.

[2]For a recent review of measures of religiosity, or the expression of religious belief and practice, see Finke and Bader 2017.

[3]As Finke and Bader (2017: 3) point out, "[s]ocial surveys, the dominant method for collecting social data, suffer from a long list of 'artifacts' that can alter outcomes, such as question ordering, the response categories offered, the interviewers used, and the question wording. Moreover, asking respondents to reveal intimate experiences with the sacred requires a trust seldom established in a ten-minute phone interview." See also Brenner (2014) for issues of using surveys to measure religiosity in Muslim countries.

[4]Of course, using administrative records to identify individuals by a particular attribute can be, and has been, abused, sometimes with terrifying results. We discuss the ethics of this practice in Appendix A.

[5]According to Finke and Bader (2017: 4), "Approximately 95% of the data files currently held in the Association of Religion Data Archives (theARDA.com) were generated by surveys."

[6]For readers interested in the mathematical details of LMs, numerous computational linguistic publications are available. We suggest Jurafsky and Martin 2019 as a starting point.

[7]Comparing perplexity scores across models is generally not valid, unless the models share vocabularies, as they do in our example application.

[8]Our source of recent economic and demographic data is the website of the Indonesian Bureau of Statistics, or Badan Pusat Statistik (BPS). For example, we found that approximately 35% of Indramayu's residents are living in poverty, and the regency has the lowest Human Development Index of all regencies the province of West Java (http://bappeda.jabarprov.go.id/documents/rancangan-rpjmd-kabupaten-indramayu-2016-2021).

[9]For example, it is reported in the Hadith that the Prophet Muhammad said, "You will be called on the Day of Resurrection by your names and the names of your fathers, so have good names."

[10]For example, Llewellyn (2018) profiles a religious figure in the Indonesian city of Medan, Mohammad Hamdan, and notes that Hamdan's parents gave him the Arabic name meaning "praiseworthy" rather than "the Indonesian word for 'praise' which is *puji* (also a common Indonesian name) . . . to show their belief in Islam."

[11]Our use of identifiable data was approved by the [anonymized] IRB after we implemented protocols to protect the data and confirmed that no information about individuals would be disseminated.

[12]The KTP, which Tipple and Speak (2009: 177) call the "sole defining element of citizenship in Indonesia", is issued to every Indonesian citizen at the age of 17 or when they are married.

[13]According to the Indonesian Census, the population of Indramayu changed minimally between 2005 and 2015 (over the same period in which Indonesia's population increased 17%). This aligns with our knowledge of the regency's migration patterns: out-migration is common, whereas in-migration is not.

[14]Salameh et al. (2018) show that a character-level LM achieves higher accuracy than a Multinonial Naïve Bayes (MNB) classifier, a classification model social scientists have used more frequently than LMs, when examining Arabic dialects and a corpus containing six (rather than two) classes. The MNB, in turn, was found to outperform other classification models, such as Linear Support Vector Machines, Convolutional Neural Networks, and bi-directional Long Short-Term Memory models.

[15]See https://srilm-python.readthedocs.io/en/latest/ and https://www.nltk.org/api/nltk.lm.html

[16]When names had fewer characters than the n-gram window, the model would consider all their characters.

[17]The Quran's proper names can be identified using the proper name part of speech tag on the Quran corpus (Dukes et al. 2013), accessible at http://corpus.quran.com/

[18]We cannot plot the measures at the RW level (the administrative level immediately below the village, N=2,380) and the RT level (the administrative level below the RW, N=10,192) because shapefiles of these administrative units are not available.

[19]See, for instance, Adcock and Collier 2001; Bollen 2014 and Ying et al. 2021, for a discussion of validity tests.

References

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2020. "Do Immigrants Assimilate More Slowly today than in the past?" *American Economic Review: Insights* 2:125–141.

Adcock, Robert and David Collier. 2001. "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* pp. 529–546.

Agarwal, Kritika. 2016. "Doing Right Online: Archivists Shape an Ethics for the Digital Age." *Perspectives on History* November 1:https://www.historians.org/publications–and–directories/perspectives–on–history/november–2016/doing–right–online–archivists–shape–an–ethics–for–the–digital–age.

Alford, Richard. 1987. *Naming and Identity: A Cross-Cultural Study of Personal Naming Practices*. New Haven, CT: HRAF Press.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg more Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94:991–1013.

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*. New York: Palgrave MacMillan.

Bollen, Kenneth A. 2014. "Measurement models: The relation between latent and observed variables." *Structural equations with latent variables* pp. 179–225.

Brandt, Philipp and Stefan Timmermans. 2021. "Abductive Logic of Inquiry for Quantitative Research in the Digital Age." *Sociological Science* 8:191–210.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. "Large Language Models in Machine Translation." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* pp. 858–867.

Brenner, Philip S. 2014. "Testing the Veracity of Self-reported Religious Practice in the Muslim World." *Social Forces* 92:1009–1037.

Chen, Yining, Jiali You, Min Chu, Yong Zhao, and Jinlin Wang. 2006. "Identifying Language Origin of Person Names With N-Grams of Different Units." In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pp. I–I.

Dukes, Kais, Eric Atwell, and Nizar Habash. 2013. "Supervised Collaboration for Syntactic Annotation of Quranic Arabic." *Language Resources and Evaluation* 47:33–62.

Edelmann, Achim and John Mohr. 2018. "Formal Studies of Culture: Issues, Challenges, and Current Trends." *Poetics* 68:1–9.

Elchardus, Mark and Jessy Siongers. 2011. "First Names as Collective Identifiers: An Empirical Analysis of the Social Meanings of First Names." *Cultural Sociology* 5:403–422.

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us About the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60:123–142.

Finke, Roger and Christopher D. Bader. 2017. *Faithful Measures: New Methods in the Measurement of Religion*. New York: NYU Press.

Freeman, Andrew T., Sherri L. Condon, and Christoper M. Ackerman. 2006. "Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenstein Edit Distance Algorithm." *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* pp. 471–478.

Fryer, Roland G. and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* 119:767–805.

Fu, Yu, Feiyu Xu, and Hans Uszkoreit. 2010. "Determining the Origin and Structure of Person Names." In *Proceedings of the Seventh International Conference on Language*

Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Gaddis, S. Michael. 2017a. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–89.

Gaddis, S. Michael. 2017b. "Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." *Socius* 3:https://doi.org/10.1177/2378023117737193.

Gamallo, Pablo, Jose Ramom Pichel, and Iñaki Alegria. 2017. "A Perplexity-Based Method for Similar Languages Discrimination." In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 109–114, Valencia, Spain. Association for Computational Linguistics.

Gerhards, Jürgen and Silke Hans. 2009. "From Hasan to Herbert: Name-Giving Patterns of Immigrant Parents Between Acculturation and Ethnic Maintenance." *American Journal of Sociology* 114:1102–1128.

Goldstein, Joshua R. and Guy Stecklov. 2016. "From Patrick to John F.: Ethnic Names and Occupational Success in the Last Era of Mass Migration." *American Sociological Review* 81:85–106.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Grofman, Bernard and Jennifer R. Garcia. 2014. "Using Spanish Surname to Estimate Hispanic Voting Population in Voting Rights Litigation: A Model of Context Effects Using Bayes' Theorem." *Election Law Journal* 13:375–393.

Guillemin, Marilys and Lynn Gillam. 2004. "Ethics, Reflexivity, and 'Ethically Important Moments' in Research." *Qualitative Inquiry* 10:261–280.

Habash, Nizar. 2008. "Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation." *Proceedings of ACL-08: HLT, Short Papers (Companion Volume)* pp. 57–60.

Habash, Nizar, Mona Diab, and Owen Rambow. 2012. "Conventional Orthography for Dialectal Arabic." *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)* Istanbul:711–718.

Harris, J. Andrew. 2015. "What's in a Name? A Method for Extracting Information about Ethnicity from Names." *Political Analysis* 23:212–224.

Hofstra, Bas and Niek C. de Schipper. 2018. "Predicting Ethnicity With First Names in Online Social Media Networks." *Big Data & Society* 5:1–14.

Imai, Kosuke and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24:263–272.

Jauhiainen, Tommi, Krister Lindén, and Heidi Jauhiainen. 2017. "Evaluation of language identification methods using 285 languages." In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 183–191, Gothenburg, Sweden. Association for Computational Linguistics.

Jauhiainen, Tommi, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. "Automatic Language Identification in Texts: A Survey." *J. Artif. Int. Res.* 65:675–682.

Johfre, Sasha Shen. 2020. "What Age Is in a Name?" *Sociological Science* 7:367–390.

Jurafsky, Daniel and James Martin. 2019. *Speech and Language Processing*. Draft Manuscript.

Karell, Daniel and Michael Freedman. 2019. "Rhetorics of Radicalism." *American Sociological Review* 84:726–753.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84:905–949.

Kuipers, Joel C. and Askuri. 2017. "Islamization and Identity in Indonesia: The Case of Arabic Names in Java." *Indonesia* pp. 25–49.

Lieberson, Stanley. 2000. *A Matter of Taste: How Names, Fashions, and Culture Change*. New Haven, CT: Yale University Press.

Lieberson, Stanley and Eleanor O. Bell. 1992. "Children's First Names: An Empirical Study of Social Taste." *American Journal of Sociology* 98:511–554.

Lieberson, Stanley and Kelly S. Mikelson. 1995. "Distinctive African American Names: An Experimental, Historical, and Linguistic Analysis of Innovation." *American Sociological Review* pp. 928–946.

Llewellyn, Aisyah. 2018. "What's in a Name in Indonesia?" *Asia Times* February 6:https://asiatimes.com/2018/02/whats–name–indonesia/.

Mikolov, Tomas, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. "Recurrent Neural Network Based Language Model." *INTERSPEECH 2010* .

Nelson, Laura. 2020. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 49:3–42.

Nelson, Laura. 2021. "Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South." *Poetics* DOI: 10.1016/j.poetic.2021.101539.

OECD. 2019. *Social Protection System Review of Indonesia*. https://doi.org/https://doi.org/10.1787/788e9d71-en.

Olivetti, Claudia and M. Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review* 105:2695–2724.

Pepinsky, Thomas B., R. William Liddle, and Saiful Mujani. 2018. *Piety and Public Opinion: Understanding Indonesian Islam*. New York: Oxford University Press.

Power, David M. W. 2011. "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies* 2:37–63.

Ramisch, Carlos. 2008. "N-gram models for language detection." Technical report.

Salameh, Mohammad, Houda Bouamor, and Nizar Habash. 2018. "Fine-Grained Arabic Dialect Identification." *Proceedings of the 27th International Conference on Computational Linguistics* Sante Fe:1332–1344.

Schwarz, Judith. 1992. "The Archivist's Balancing Act: Helping Researchers While Protecting Individual Privacy." *Journal of American History* 79:179–189.

Seguin, Charles, Chris Julien, and Yongjun Zhang. 2021. "The Stability of Androgynous Names: Dynamics of Gendered Naming Practices in the United States 1880–2016." *Poetics* 85.

Stolcke, Andreas. 2002. "SRILM – An Extensible Language Modeling Toolkit." *Proceedings of the International Conference on Spoken Language Processing* pp. 901–904.

Stoltz, Dustin S. and Marshall A. Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement Via Word Embeddings in Texts." *Journal of Computational Social Science* 2:293–313.

Stoltz, Dustin S. and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics* DOI: 10.1016/j.poetic.2021.101567.

Subotić, Jelena. 2021. "Ethics of Archival Research on Political Violence." *Journal of Peace Research* 58:342–354.

Sue, Christina A. and Edward E. Telles. 2007. "Assimilation and Gender in Naming." *American Journal of Sociology* 112:1383–1415.

Susewind, Raphael. 2015. "What's in a Name? Probabilistic Inference of Religious Community from South Asian Names." *Field Methods* 27:319–332.

Tavory, Iddo and Stephan Timmermans. 2014. *Abductive Analysis: Theorizing Qualitative Research*. Chicago: The University of Chicago Press.

Tipple, Graham and Suzanne Speak. 2009. *The Hidden Millions: Homelessness in Developing Countries*. New York: Routledge.

Traeger, Margaret L., Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. 2020. "Vulnerable Robots Positively Shape Human Conversational Dynamics in a Human–Robot Team." *Proceedings of the National Academy of Sciences* 117:6370–6375.

Uhlenbeck, Eugenius Marius. 1969. "Systematic Features of Javanese Personal Names." *Word* 25:321–335.

Vatanen, Tommi, Jaakko J. Väyrynen, and Sami Virpioja. 2010. "Language Identification of Short Text Segments with N-gram Models." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Webb, Eugene J., Donald T. Campbell, Richard D Schwartz, and Lee Sechrest. 2000. *Unobtrusive Measures*. Thousand Oaks, CA: Sage, rev. ed. edition.

Ying, Luwei, Jacob Montgomery, and Brandon Stewart. 2021. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Forthcoming in Political Analysis* .

Zaidan, Omar F. and Chris Callison-Burch. 2014. "Arabic Dialect Identification." *Computational Linguistics* 40:171–202.

Zelinsky, Wilbur. 1970. "Cultural Variation in Personal Name Patterns in the Eastern United States." *Annals of the Association of American Geographers* 60:743–769.

Dataset	Count		Unique	Class			
				Arabic		Other	
All single names	1,381,923		149,283				
Train_{Rand}	43,239	3%	9,323	250	3%	9,073	97%
Train_{Freq}	1,035,882	75%	9,962	1,050	11%	8,912	89%
Set_{Full}	27,640	2%	11,446	1,297	11%	10,149	89%

Table 1: The single-name datasets. The table describes the three datasets we sampled for manual coding. For each dataset, the table presents the total count of single names (Count) and the number of unique single names (Unique). The percentages next to the counts are derived by comparing the subsets to the entire collection of single names. For the annotated datasets, the table presents the proportions of ARABIC and OTHER classes.

Class	Training		Set_{Full}		Test	
Arabic	241,592	22%	5,011	18%	726	11%
Other	837,529	78%	22,629	82%	5,745	89%
Total	1,079,121		27,640		6,471	

Table 2: Class-based non-unique single-name counts for the training dataset, the dataset derived from full names (Set_{Full}), and the portion of Set_{Full} that is out-of-vocabulary ("Test").

	Set_{Full}		Test	
	Baseline	Best LM	Baseline	Best LM
Accuracy	97%	97%	**89%**	86%
Average precision	**98%**	93%	45%	**70%**
Average recall	93%	**97%**	50%	**82%**
Average F-score	95%	95%	47%	**73%**
Arabic precision	**100%**	87%	0%	**42%**
Arabic recall	86%	**97%**	0%	**76%**
Arabic F-score	**92%**	91%	0%	**54%**
Other precision	97%	**99%**	89%	**97%**
Other recall	**100%**	97%	**100%**	87%
Other F-score	98%	98%	**94%**	92%

Table 3: The single-name classification evaluation results. The table presents the results of two systems on two datasets. The two systems are the majority baseline and the best LM determined empirically. The two datasets are Set_{Full} and the out-of-vocabulary portion of the Set_{Full} ("Test"). In addition to overall system results in terms of accuracy, and average precision, recall and F-score, the table includes the precision, recall and F-score for the ARABIC and OTHER classes. The superior result of each comparison is in boldface.

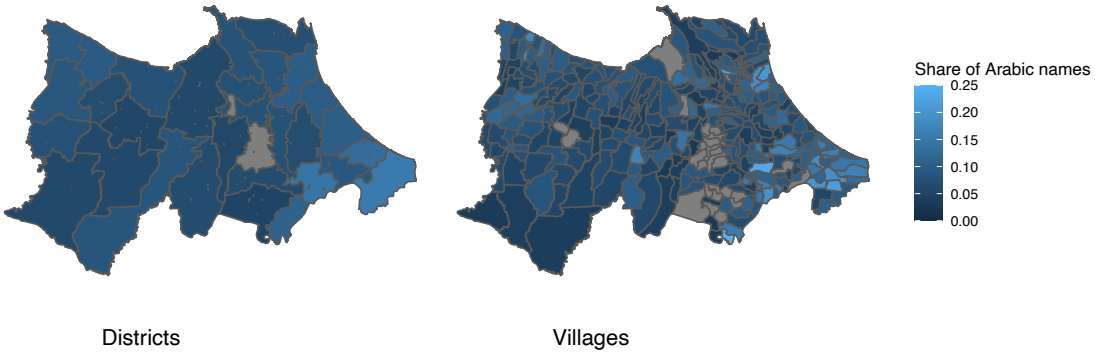

Figure 1: Spatial variation in Arabic full names at the levels of villages and districts

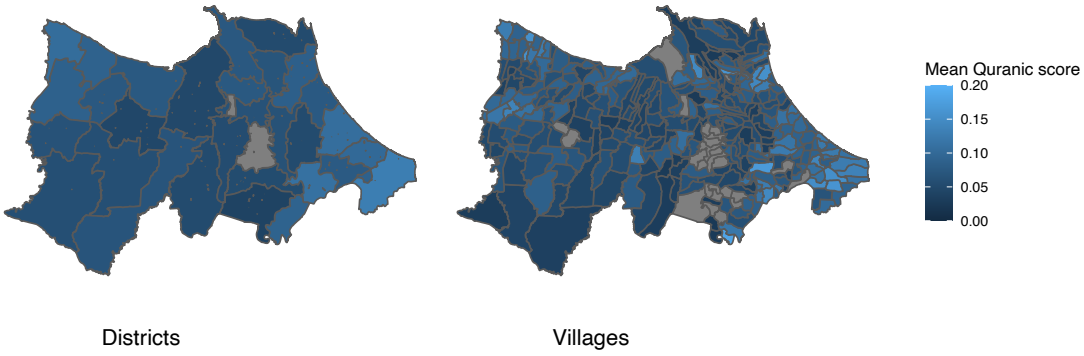

Figure 2: Spatial variation in full names' Quranic similarity across villages and districts

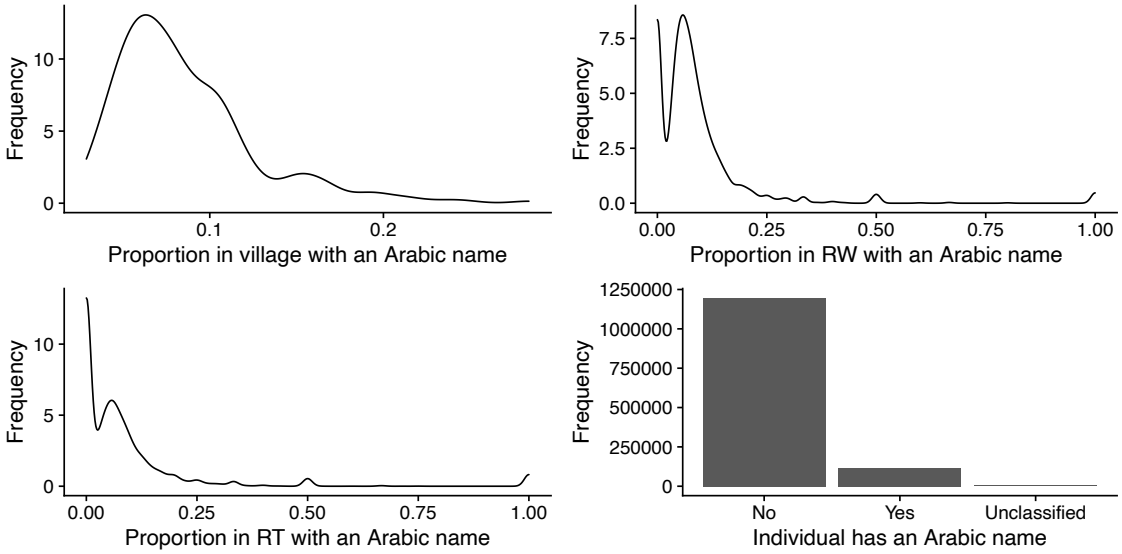

Figure 3: Low-level variation in Arabic naming

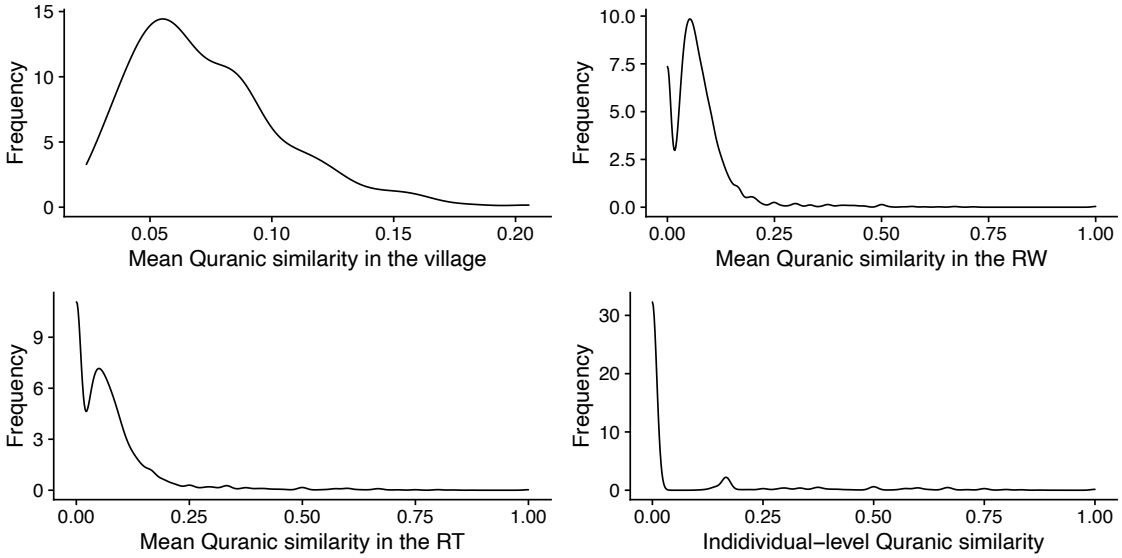

Figure 4: Low-level variation in Quranic naming

Online Appendices

Language Models in Sociological Research: An Application to Classifying Large Administrative Data and Measuring Religiosity

- Appendix A: Comments on ethics
- Appendix B: Validation of name-derived religiosity variables
- Appendix C: Preliminary machine learning comparison

A. Comments on ethics

In the main text, we mention that administrative data can be used to unobtrusively measure attitudes, beliefs, or characteristics of a sensitive or personal nature. Of course, the use of official records to identify people with a particular attribute has led to tragic and terrifying outcomes, particularly for members of minority and marginalized groups. However, these outcomes are not necessary: better identifying the location of minority and marginalized groups could improve the allocation, provision, and distribution of needed resources. For example, during a vaccination roll-out, administrative databases could help a local government establish a greater number of vaccination stations in areas with many residents who belong to groups that have traditionally not been provided adequate access to healthcare. Thus, the use of administrative data to measure potentially sensitive characteristics of people is not always "wrong" or "right".

To find guidance when performing research such as ours, scholars can begin with the 1979 Belmont Report on research involving humans, which provides the foundation for the United States' Institutional Review Boards (IRBs). The Belmont Report highlights criteria that inform ethical human subjects research, including two that are sometimes in tension: respect, or allowing the research subject to exercise their agency by consenting (or not) to the research, and beneficence, or conducting research that maximizes benefit while minimizing harm.

When researchers use administrative data, the people recorded in the data typically cannot give consent. In fact, they may suffer emotional distress (and even be exposed to danger) if researchers tracked them down to seek their consent. One way that IRBs sidestep this problem is by defining "human subjects research" as research that involves interaction with people or intervention into their lives. Using administrative records usually does not intervene into people's lives, and therefore is sometimes not considered "human subjects research" (for instance, as long as identifying information is properly handled).

However, simply classifying research using administrative records as not "human sub-

jects research" does not resolve important ethical concerns. It does not absolve researchers from weighing the benefits of research against the (speculated) desires of the people in the records. Therefore, we recommend that researchers consider discussions in their fields about the ethics of administrative data, as well as debates in History, Anthropology, and cognate fields on the ethical use of archival data, which often deals with records of individuals who (similarly) cannot give consent (*e.g.*, Schwarz 1992; Agarwal 2016).

We find especially useful guidance in a recent extension of these discussion to Political Science. Subotić (2021) argues that scholars working with "unobtrusive" records, such as historical records and administrative data, must still "think through the balance of harms and benefits not only in terms of current harms/benefits, but also in the light of possible future harms, such as when, for example, a changing political situation deems information seemingly innocuous at the time much more dangerous or damaging to the subjects later on" (349). Furthermore, even though researchers working with administrative data may not be working with "human subjects" (according to a possible IRB definition) there "are ways in which [the researchers] can conceptualize groups that could be owed benefits" (349).

To provide examples of Subotić's points, we offer some context of our study. First, we considered the potential for harm resulting from our particular study. We assessed the risk as particularly low since we focused on a majority religion. That is, our analysis locates strongly religious Muslims in a region that is nearly entirely Muslim. In Indramayu (and in Indonesia) Muslims are not a marginalized group, threatened minority, or protected class. Second, our study is the first step in a larger research project designed to confer benefits to residents of Indramayu—the individuals in the records and whom we "conceptualize [as the group] owed benefits." The larger project explores why some low-skilled Indonesians (many of whom live in Indramayu) migrate to different parts of the world, such as the Gulf Cooperation Council states, which has serious effects on their earnings and quality of life. (The LM classification we present here feeds into an effort to determine whether migrants' religiosity influences their decision to migrate to Muslim countries instead of, say, East Asian countries like South Korea and Taiwan). We have

been in contact with Indonesian government agencies overseeing migration affairs, and have already agreed to report the findings of our broader project.

Yet, as earlier mentioned, our approach *could* be used in ways that do not confer benefits, and even potentially bring harm to the people in administrative databases. Thus, we emphasize two further points: (1) not all administrative data research is the same and (2) "scholars need to make ethical choices at multiple 'ethically important moments' that arise during research" (Subotić 2021: 351; see also Guillemin and Gillam 2004). In other words, researchers using our approach—and similar approaches—should (1) carefully consider the risks and benefits of their own particular data and (2) which of their research decisions (made at "ethnically important moments") might confer harm or benefits. For example, perhaps spatially locating some minority groups will bring positive changes while locating members of another group would bring them harm. In this case, the researchers do not have to locate the latter group or could use spatial aggregations that mitigate the risk.

We expect that administrative data will only become more available in coming years. These data could bring important scholarly insights, which could potentially benefit the individuals and groups recorded in the data. However, the data could also bring risk and harm. We do not think there one clear answer for how to proceed in all cases, but we do believe that researchers must carefully consider the risks inherent in their case and data, as well as have an idea how their research could benefit the people who have become encoded in the data. We look forward to greater discussion on this topic among social scientists using newly accessible data and computational approaches.

B. Validation of name-derived religiosity variables

In this appendix, we evaluate our two names-based religiosity variables against two criterion variables of spatial variation in religiosity. A criterion validity test examines the extent to which our names-based operationalization of religiosity is related to an outcome to which it should be theoretically correlated (Pedhazer and Schmelkin 2013).[20]

Our first criterion variable is the proportion of registered voters who were registered to perform the Hajj between the years 2011 and 2017, measured at the village level. Since we were able to obtain multiple years of Hajj registration data, we collapsed this variable into a seven-year average for each administrative level.

Due to the demand of Muslims around the world to perform the Hajj, each country receives an annual quota of visas. The limited number of annual visas for a country the size of Indonesia means that people are required to register far in advance. Saudi Arabia typically grants Indonesia approximately 220,000 annual visas, and, according to the Indonesian government, the average period between registering and being granted a visa is now 20 years. The Indonesian government allocates a portion of the total quota to each regency; as a result, the waiting lists are administered at this level. We made a request to the Indramayu office of the Ministry of Religious Affairs for the administrative data on the regency's registrants to perform the Hajj. This data set included their name and approximate address, which we linked to villages. (Placing individuals in lower administrative units was possible, but resulted in numerous units with no observed Hajj registrants or Arabic names. Moreover, our second criterion variable is only observed at the village level.) We divided the total number of Hajj registrants in each village by the corresponding number of registered voters.

Our second criterion variable is the proportion of total students enrolled in an Islamic school. For the denominator, we acquired from the regency the total public and private enrollment for all primary and secondary schools in each village. We located these data at the Indonesian Bureau of Statistics, or Badan Pusat Statistik (BPS). This dataset

[20]Specifically, our tests are assessments of concurrent validity because the criterion variables are measured at the time of our names-based variables' measurement.

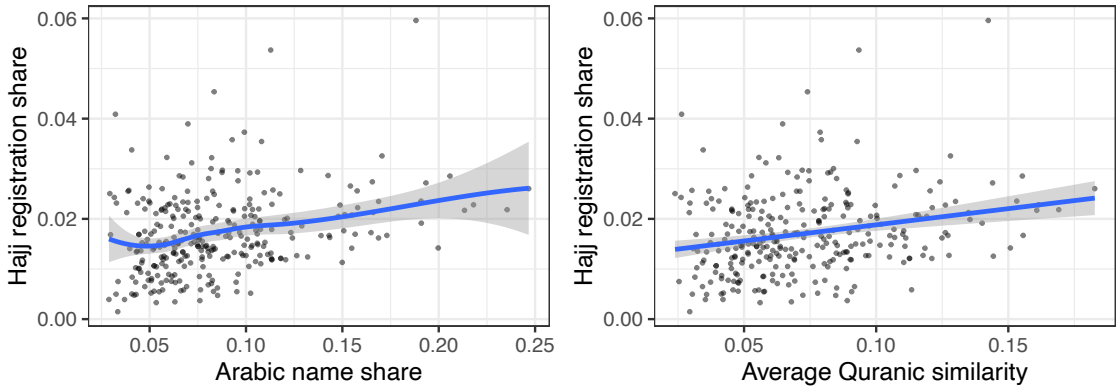

Figure B.1: Validation test 1: Hajj-registration share at the village level

also provided the number of students attending Islamic schools, which we used as the numerator.

Each criterion variable has strengths and limitations, which, in general, reflects the difficulties in measuring religiosity both across individuals and space. On the one hand, the Hajj registration variable is a direct measure of a decision by individuals to pursue a significant and religious act, the Hajj. Yet, on the other hand, we lack the individual-level data to control for the available means of each individual to perform this pilgrimage (*e.g.*, disposable income, time). We would expect that in a developing country, especially in a relatively impoverished regency like Indramayu, the costs associated with the Hajj would be prohibitive. This registry also omits all people who either have already performed the Hajj or those who intend to but have not yet registered. Figure B.1 shows that relationship between the share of a village's adults who are registered to attend the Hajj and Arabic name share (left) and average Quranic similarity score (right), respectively. To put this in perspective, a one standard deviation increase in Arabic name share is associated with a 0.29 standard deviation increase in the proportion of adults registered for the Hajj (and for average Quaranic similarity it is correlated with a 0.32 standard deviation increase).

Our second criterion variable has similar trade-offs. The choice of parents to send their children to an Islamic school is a strong signal of piety. This is especially true in an educational environment in which there are many private, non-Islamic alternatives to

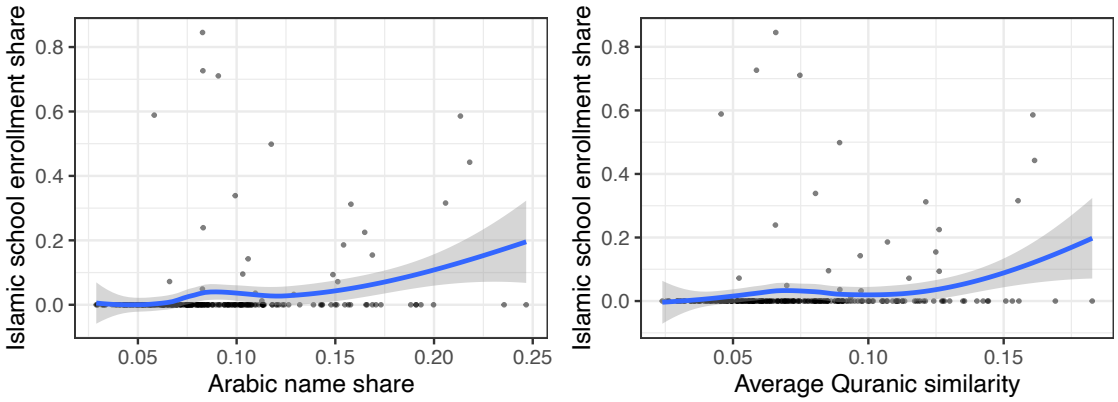

Figure B.2: Validation test 2: Islamic-school enrollment share at the village level

public schools. Yet, we assume that the composition of schools in a village reflects the decision of parents in the village and are largely serving the children in those villages. We cannot test for this assumption. More importantly, the vast majority of villages do not contain even one Islamic school, resulting in over-dispersion. This is apparent in the Figure B.2. The plot clearly demonstrates that the enrollment share is zero in the vast majority of villages (*i.e.*, there are very few Islamic schools). While we have used Tobit models in attempt to address this problem, the results with this variable should be interpreted cautiously.

Despite their limitations, these variables in combination form suitable benchmarks to test the validity of our name-based variables. In particular, they capture distinct aspects of individuals' religious behavior independently of our research activities. In addition, they cover the same geographic and temporal range as our name-based variables. As for the problem of over-dispersion in the schooling variable, we employ methods that model the structure of these variables. Table B.1 presents the correlation coefficients for each variable at the village level.[21] Table B.2 provides basic descriptive statistics for both names-based variables and each criterion variables at all administrative levels for which the data are available.

[21]Given the paucity of districts in Indramayu (N=30), we did not conduct validity tests at this administrative level. However, we do use district-level fixed effects in a robustness check, as explained below.

	Islamic School (%)	Islamic Enroll. (%)	Hajj Reg. (%)	Hajj Reg. (#)	Arabic Name (%)
Islamic School (%)					
Islamic Enroll. (%)	0.95				
Hajj Reg. (%)	0.12	0.08			
Hajj Reg. (#)	0.21	0.17	0.71		
Arabic Name (%)	0.26	0.22	0.29	0.32	
Quranic Similarity Score	0.19	0.16	0.23	0.32	0.91

Table B.1: Village-level correlations (N=292)

B.1. Main results

Table B.3 presents the estimates from simple linear regressions of our names-based measures and the two criterion variables. The unstandardized coefficient is shown first, followed by the standard error in parentheses and the standardized coefficient in brackets. Column 1 shows the OLS estimates when Hajj registration is regressed on Arabic name share and Quranic similarity score. We find that both of our names-based religiosity measures are positive and significant at the 99% level. Column 2 reports the estimates when using the second criterion variable, the proportion of students enrolled in Islamic schools. It shows that the coefficients are also positive and significant at the 99% level for Arabic name share and at the 95% level for Quranic similarity, respectively. The standardized coefficients for each also indicate a moderate effect size. Complete results for each model from Table B.3 are shown in Table B.4 (unstandardized) and Table B.5 (standardized).

B.2. Robustness results

The two criterion variable tests provide support for the validity of our names-based religiosity measures. To increase our confidence further, we conducted two additional tests of the robustness of these relationships. First, we addressed the concern with excess zeroes in the schooling criterion variable by fitting a Tobit regression.[22] We also re-

[22]As explained earlier, the schooling variable is continuous and highly rightward skewed due to censoring at zero. This can make the truncated portion of OLS estimates biased. The Tobit (or censored

		Administrative unit	
	Statistic	Village	District
Name-derived variables			
Arabic name share	Mean	0.09	0.08
	Std. Dev.	0.04	0.02
Average Quranic similarity	Mean	0.07	0.07
	Std. Dev.	0.03	0.02
Count of registered voters	Mean	4,416.41	40,437.72
	Std. Dev.	1,972.28	19,746.66
Criterion variables			
Hajj registration share	Mean	0.02	0.001
	Std. Dev.	0.01	0.001
Islamic school share	Mean	0.03	0.04
	Std. Dev.	0.10	0.05

Table B.2: Descriptive statistics of the name-derived and criterion variables of religiosity by administrative unit

estimated the Hajj registration relationship with a Tobit model. The results are reported in Appendix Table B.6. In both models, the coefficient increased and the standard error decreased compared to the corresponding OLS models.

Second, we addressed concerns due to the lack of important covariates, such as income per capita, by estimating fixed-effects models with a dummy variable for the district level, the next highest administrative level. If unobserved factors, such as income per capita, are heterogeneously distributed across space, then fixed effects should help control for this unmeasured heterogeneity. The estimates for these fixed-effects regressions are reported in Appendix B.7. We find that the inclusion of fixed effects produces broadly similar results to the OLS and Tobit models for both Hajj registration share and Islamic school enrollment share.

regression) model is appropriate when a continuous dependent variable is bounded at one of the extremes (in this case zero), significant clustering around that extreme value, and is highly rightward skewed or unbounded at the other extreme (Wooldridge 2010).

	Village-level Validity Tests	
	Hajj registration sh. (1)	Islamic school enrollment Sh. (2)
Arabic name share	0.062^{***} (0.013) **[0.298]**	0.600^{***} (0.224) **[0.207]**
Quranic similarity	0.064^{***} (0.016) **[0.243]**	0.556^{**} (0.264) **[0.149]**
Mean of outcome	0.017	0.026
N	292	275

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table B.3: Validation analysis results at the village level. Simple linear regression estimates presented as unstandardized coefficients, (robust standard errors), and **standardized coefficients**. Significance levels are constant across all unstandardized and standardized coefficients.

	Hajj registration share		Islamic school enrollment share	
	(1)	(2)	(3)	(4)
Arabic name share	0.062^{***}		0.600^{***}	
	(0.013)		(0.224)	
Quranic similarity		0.064^{***}		0.556^{**}
		(0.016)		(0.264)
Constant	0.012^{***}	0.012^{***}	-0.027	-0.016
	(0.001)	(0.001)	(0.017)	(0.019)
Mean of outcome	0.017	0.017	0.026	0.026
N	292	292	275	275
Adjusted R^2	0.085	0.055	0.044	0.021

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table B.4: Village-level validation with unstandardized estimates. Robust standard errors in parentheses.

	Hajj registration share		Islamic school enrollment share	
	(1)	(2)	(3)	(4)
Arabic name share	0.298***		0.207***	
	(0.061)		(0.077)	
Quranic similarity		0.243***		0.149**
		(0.061)		(0.071)
Constant	−0.001	−0.001	−0.018	−0.018
	(0.056)	(0.057)	(0.056)	(0.057)
Mean of outcome	0	0	0	0
N	292	292	275	275
Adjusted R^2	0.085	0.055	0.044	0.021

*p < .1; **p < .05; ***p < .01

Table B.5: Village-level validation with standardized estimates

	Hajj registration share		Islamic school enrollment share	
	(1)	(2)	(3)	(4)
Arabic name share	0.329***			
	(0.081)			
Quranic similarity		0.272***	1.845***	1.845***
		(0.083)	(0.614)	(0.614)
Constant	−0.261***	−0.261***	−7.858***	−7.858***
	(0.099)	(0.099)	(1.691)	(1.691)
Mean of outcome	0	0	0	0
N	292	292	275	275
Log Likelihood	−304.535	−307.526	−112.069	−112.069
Wald Test (df = 1)	16.616***	10.831***	9.014***	9.014***

*p < .1; **p < .05; ***p < .01

Table B.6: Village-level validation with Tobit models

	Hajj registration share		Islamic school enrollment share	
	(1)	(2)	(3)	(4)
Arabic name share	0.239***		0.268***	
	(0.069)		(0.079)	
Quranic similarity		0.224**		0.231***
		(0.082)		(0.081)
Mean of outcome	0	0	0	0
N	292	292	275	275
Adjusted R^2	0.315	0.307	0.025	0.007

*p < .1; **p < .05; ***p < .01

Table B.7: Village-level validation with district-fixed effects

C. Preliminary machine-learning comparison

Our goal has been to provide an accessible explanation and simple example of LMs to help social scientists decide whether LMs are appropriate for their particular case and data. A clear next step to advance the use of LMs in social science is to systematically compare LM and ML performance across common use-cases. These comparisons would help develop a set of guidelines for researchers deciding between LMs or ML methods.

We have conducted a preliminary comparison of LMs and ML methods using our data, and we report our findings and insights in this appendix. The LM portion of the comparison used the techniques, steps, and results described in the main text. The ML portion involved the following steps.

C.1. Analysis

First, we trained three common ML algorithms—support vector machine (SVM) (implemented with `kernlab`), random forest (RF) (implemented with `ranger`), and extreme gradient boosting (XGB) (implemented with `xgboost`)—on Train_{Rand} and Train_{Freq} before they were merged. (See the main text for explanations of these datasets). Using the separate (hence smaller) datasets allowed for a faster comparison of many models and more efficient model-tuning. The models were tested on a holdout set.

For all algorithms, the re-sampling strategy was five-fold cross validation repeated five times. We generated hyperparameter combinations stochastically by sampling from each algorithms' hyperparameter space in a procedure that maximizes coverage (*i.e.*, spatial entropy). The SVM and RF tuning processes tested 10^3 combinations of hyperparameters. Since XGB's hyperparameter space has many more dimensions, we tested 10^4 candidate combinations of the learning rate, number of trees, maximum tree depth, the number of predictors in each split, minimum observations per splitting node, required loss reduction to justify a further split, and the sample size of each iteration. At the end of each tuning procedure, we chose the hyperparameter combination with the best F-score in out-of-sample classification. Ultimately, we found RF to be the clear winner.

Next, we trained the RF model on the full "Training" dataset, or the combination of Train_{Rand} and Train_{Freq} using the best of 81 hyperparameter combinations chosen through 5-fold cross validation (no repeats). The final hyperparameters for our RF model were 15 predictors in each split and no fewer than 9 observations at the end of each node. We judged performance using the Set_{Full} and Test datasets.

We used two types of predictors. The majority of the predictors were based on character patterns common throughout Indonesian (Bahasa) transliterations of Arabic words. For example, for each name in the data, we counted the presence of character patterns like "KH", "DH", "Q", "RR", "AW", "LL" that are common in Arabic-transliterated words but relatively infrequent in words of Javanese and Indonesian origin. The second type of predictor was based on each name's Jaro-Winkler string distance from names known to be of Arabic origin. Specifically, we measured each name's similarity to (1) a set of common Indonesian names of Arabic origin, (2) a set of common Arabic names transliterated into English, and (3) a set of proper nouns from the Quran transliterated into English. We obtained these lists from various sources, including Wikipedia and our Indrayamu-based research assistants. Permutation-based variable-importance metrics suggest that the most useful predictors for our models were the name's similarity to English names of Arabic origin, the name's similarity to common Indonesian names of Arabic origin, the name's similarity to a Quranic proper nouns, and the presence of certain character patterns like "KH", "Q", and "[[vowel]][HTD]".

Note that the lists of names known to be of Arabic origin were not part of the LM analysis. Therefore, our comparison is not between the LMs and ML methods using strictly the same data. Instead, our comparative analysis consists of comparing the performance of our LMs and ML methods on the same data, but when the latter uses extra data, which we assume would be part of reasonable ML strategies.

C.2. Results

Table C.1 displays our best RF results alongside the same metrics for our baseline model (the "majority baseline model") and the best LM, which are reported in the main text's

	Set_{Full}			Test		
	RF	Baseline	Best LM	RF	Baseline	Best LM
Average precision	95%	**98%**	93%	64%	45%	**70%**
Average recall	92%	93%	**97%**	57%	50%	**82%**
Average F-score	93%	95%	95%	58%	47%	**73%**
Arabic precision	93%	**100%**	87%	39%	0%	**42%**
Arabic recall	84%	86%	**97%**	16%	0%	**76%**
Arabic F-score	89%	**92%**	91%	23%	0%	**54%**
Other precision	97%	97%	**99%**	90%	89%	**97%**
Other recall	99%	**100%**	97%	97%	**100%**	87%
Other F-score	98%	98%	98%	92%	**94%**	92%

Table C.1: Results of the language model and machine learning comparison. The table presents the results of three systems on two datasets. The three systems are the best random forest (RF) model, the majority baseline model, and the best LM determined empirically. The two datasets are Set_{Full} and the out-of-vocabulary portion of the Set_{Full} ("Test"). In addition to overall system results in terms of accuracy, and average precision, recall and F-score, the table includes the precision, recall and F-score for the ARABIC and OTHER classes. The superior result of each comparison is in boldface.

Table 3. We see that the RF model performed relatively well and matched both the baseline and best LM in some metrics. However, the LM is superior when predicting the linguistic units of primary interest: out-of-vocabulary ("Test") minority class terms, or unlabeled Arabic names. The RF obtained a precision and recall of 39% and 16%, respectively, compared to the LM's 42% and 76%. The RF's F-score was 23% while the LM's was 54%.

C.3. Discussion

Our preliminary comparison of LMs and ML methods indicates that the ML approach achieves relatively good results, but the LM was better for predicting the linguistic units of main interest: out-of-vocabulary minority class terms, which, in our case, are unlabeled Arabic names. This finding leads us to emphasize two questions and insights for researchers to consider:

1. How important is the classification performance on the minority class? If it is very

important, as in our case, then we recommend researchers consider LMs, especially if using short texts, when linguistic units exhibit small variations around base forms (*e.g.*, when working with dialects and related languages), and when the subunits of larger units offer important information (*e.g.*, characters' relations to words, words' relations to phrases).

2. How much training data are available? We have a lot of training data, and the RF makes use of all of it. Yet, this model still achieved an F-score of only 23% on out-of-vocabulary minority-class names (compared to the LM's 54%). If researchers are unable to obtain a large training dataset, they may want to consider LMs. The LM results may not be spectacular, but our analysis suggests they will be better than results from ML approaches.

In sum, our analysis shows that LMs are likely to be worth social scientists' effort if they are specifically interested in the minority class and face budget constraints that limit the size of the training dataset. However, our comparison is only a preliminary effort—more systematic comparisons across typical use-cases are needed to develop comprehensive and detailed guidelines.

References

Agarwal, Kritika. 2016. "Doing Right Online: Archivists Shape an Ethics for the Digital Age", *Perspectives on History*. https://www.historians.org/publications-and-directories/perspectives-on-history/november-2016/doing-right-online-archivists-shape-an-ethics-for-the-digital-age

Guillemin, Marilys and Lynn Gillam. 2004. "Ethics, Reflexivity, and 'Ethically Important Moments' in Research", *Qualitative Inquiry* 10(2): 261–280.

Pedhazur, Elazar J. Liora Pedhazur Schmelkin. 2013. *Measurement, Design, and Analysis: An Integrated Approach*. New York: Psychology Press.

Schwarz, Judith. 1992. "The Archivist's Balancing Act: Helping Researchers While Protecting Individual Privacy", *Journal of American History* 79(1): 179–189.

Subotić, Jelena. 2021. "Ethics of Archival Research on Political Violence", *Journal of Peace Research* 58(3): 342–354.

Wooldridge, Jeffrey M. 2020. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.