The Pulse of Public Opinion

Using Twitter Data to Analyze Public Perception of Reform in El Salvador

Skipper Seabold

Alex Rutherford

Olivia De Backer

Andrea Coppola

WORLD BANK GROUP

Macroeconomics and Fiscal Management Global Practice Group

August 2015

Abstract

This study uses Twitter data to provide a more nuanced understanding of the public reaction to the 2011 reform to the propane gas subsidy in El Salvador. By soliciting a small sample of manually tagged tweets, the study identifies the subject matter and sentiment of all tweets during six one-month periods over three years that concern the subsidy reform. The paper shows that such an analysis using Twitter data can provide a useful complement to existing household survey data and even potentially replace survey data if none were available. The findings show that when people tweet about the subsidy, they almost always do so in a negative manner; and there is a decline in discussion of topics about

the reform subsidy, which coincides with increase in support for the subsidy as reported elsewhere. Therefore, the study concludes that decreasing discussion of the subsidy reform indicates an increase in support for the reform. In addition, the gas distributor strikes of May 2011 may have contributed to public perception of the reform more than previously acknowledged. This study is used as an opportunity to provide methodological guidance for researchers who wish to undertake similar studies, documenting the steps in the analysis pipeline with detail and noting the challenges inherent in obtaining data, classification, and inference.

This paper is a product of the Macroeconomics and Fiscal Management Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at jsseabold@gmail.com.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

The Pulse of Public Opinion: Using Twitter Data to Analyze Public Perception of Reform in El Salvador

Skipper Seabold	Alex Rutherford
American University	United Nations Global Pulse
Olivia De Backer	Andrea Coppola
United Nations Global Pulse	The World Bank*

JEL Classification: C55,C8,H2.

Keywords: Political Economy of Reform, Fuel Subsidy, Big Data

*Corresponding author e-mail: jsseabold@gmail.com. This is a preliminary draft. The authors would like to thank Diana Lachy Castillo for translating as well as Marcelo Echague Pastore and Juliana Torres for tagging tweets.

1 Introduction

Changes in economic policy, especially those concerning subsidies of staple goods and utilities, are often controversial. There is often a negativity bias in public perceptions towards changes in policy arising from resistance to a change in the status quo or a lack of understanding of the effects of these changes, which may even arise among those who stand to gain [Fernandez and Rodrik, 1991]. Calvo et al. [2014] recently investigated the public perceptions of a specific program of gas subsidy reform which was implemented in April 2011 in El Salvador. Household survey data were used to illuminate the dimensions underlying this perception such as political partisanship, level of information about the reform, and trust in the government’s ability to deliver the subsidy after the reform. Survey data were also analyzed from the period following the reform’s implementation showing how the role of these different factors evolved.

The reform considered for this work is an example of a reform that was initially unpopular despite the fact that the majority of the population stood to gain from it [Tornarolli and Vazquez, 2012]. The reform involved changing the subsidy from producers to consumers. Instead of subsidizing prices at the point of sale, the new mechanism delivered an income transfer to a large set of eligible households. As a result of this change the consumer price increased from \$5.10 (the subsidized price) to \$13.60 (the price without subsidy). Individual households received a transfer \$8.50 per month provided they were eligible. The eligibility requirement was consuming less than 200 Kwh in electricity per month, a criterion that was meant to exclude the highest income brackets of the population from receiving the gas subsidy. Households that lacked electricity needed to register at a governmental office and provide their address so that the household received a card (*tarjeta*) that entitled it to collect monthly the \$8.50.

The evolution of sentiment regarding the reform was investigated using household surveys conducted by *La Prensa Grafica*, the largest newspaper in El Salvador. The surveys were conducted in 6 different time periods and covered demographic questions such as income and political views. It was demonstrated that the overall sentiment towards the reform could be effectively accounted for by considering both the individual’s perception of the government’s ability to enact the reform and political affiliation.

In recent years social media has emerged as a novel and promising alternative means to extract societal level information. These data are useful for a variety of purposes including measuring brand perception, stock trading [Bollen et al., 2011] and civic participation [Bond et al., 2012]. More recently, such data sources and appropriate analysis techniques have been co-opted in order to improve the wellbeing of vulnerable popu-

lations through development and humanitarian programs. These include public health [Stoové and Pedrana, 2014, Garcia-Herranz et al., 2014], perceptions on vaccination programs [UNICEF, 2013], forecasting migration [Zagheni et al., 2014] and commuting flows [Lenormand et al., 2014], early-warning of epidemics [Garcia-Herranz et al., 2014] and information sharing during disaster response [Imran et al., 2013] and criminal violence [Monroy-Hernández et al., 2013].

The advantages of such public social media signals are clear. Large quantities of passively produced data may be collected in real-time or near real-time. Often social media content is augmented with user meta-data such as geographic location and demographic information such as gender and ethnicity may be inferred [Mislove et al., 2011, Pennacchiotti and Popescu, 2011, Sakaki et al., 2014]. Although such novel signals are not without their shortcomings, such as a bias towards young and urban populations, the potential for these streams to augment traditional information collecting processes such as individual or household level surveys is clear.

2 Objective

In this study, we ask whether we can replicate the results of the more traditional *La Prensa Grafica* household surveys in El Salvador and use social media data over the same period to provide a deeper analysis of public sentiment. To accomplish this we obtain a number of Spanish language tweets containing certain keywords of interest and filter these tweets to those originating in El Salvador—a process known as *geolocation*. We then perform exploratory analysis of the data and refine these results until we are satisfied that we have captured much of the available relevant discourse. We will then classify these tweets by subject and by sentiment. This is done in two stages. First, domain experts identify the subset and sentiment of a subset of the tweets manually. Then appropriate statistical classifiers are used to estimate the content matter and sentiment of the remaining tweets. This workflow is described in Figure 1. The following sections describe this process in greater detail. The data gathering process, geolocation, and manual tagging are described in Section 3. Section 4 details how we identified the ground truth for the topic and sentiment of a subset of tweets. Section 5 describes the classification process. We present the results in Section 6. We provide some avenues for further study in Section 7, and Section 8 concludes.

Objective Overview

Figure 1: An overview of the analysis pipeline used in the present study using Twitter data to understand the evolution of the public perception of the El Salvador gas subsidy reforms of 2011.

3 Data

3.1 Source

We consider the historical archive of the Twitter firehose of public tweets available through a paid service.

3.2 Taxonomy

In order to filter relevant content from the period of interest, a taxonomy of Spanish keywords related to the topic was constructed. This step is of critical importance. The taxonomy must contain all permutations of words relevant to the topic of interest including slang, abbreviations and synonyms. For this reason, several domain experts and native Spanish speakers were consulted to advise on taxonomy content. Further, if the

taxonomy is too broad there is a risk of including irrelevant content. Therefore an iterative process is required whereby the results of the filtering process is examined by eye and further logical rules combining more than one single word are applied if necessary of the form

IF *word_A* AND NOT *word_B*

Broadly, the taxonomy included terms relevant to several different thematic areas identified in Calvo et al. [2014] – gas and electrical prices, political actors and entities and the subsidy itself. Several iterations of the taxonomy were considered and duplicate content was removed and further filtering applied to remove terms introducing irrelevant content.

First Iteration

subsidio OR *tambo* OR *GLP* OR *propano* OR *focalizacion*

OR (*reforma* AND (*gas* OR *propano* OR *GLP* OR *tambo* OR *focalizacion*))

OR (*precio* AND (*gas* OR *propano* OR *GLP* OR *tambo* OR *focalizacion*))

Second Iteration

Gas/Subsidy

This includes alternative terms for gas canisters

(ANY(*cilindro*, *gas*) AND (*precio* OR *reforma*)) OR #*SubsidioGas* OR *subsidi**¹

Electricity

Includes the acronyms for electrical companies and regulatory bodies

AND(*electricidad*, *recibo*) OR AND(*recibo*, *luz*) OR (ANY(*caess*, *aes*, *cne*, *eeo*, *clesa*, *dui*, *nic*, *cenade*) AND OR(*precio*, *reforma*, *subsidio*, *pagar*))

Politics

Includes the names of prominent political parties and public figures who commented on the subsidy

ANY(*fmln*, *arena*, *minec*, *sigfrido reyes*, *daqueramo*) AND ANY(*reforma*, *precio*, *fraude*)
ANY(*archbishop*, *Jose Luis Escobar Alas*) AND ANY(*reforma*, *precio*)

¹The asterisk is a wildcard operator and matches all zero or more characters. Therefore, *subsidi** matches any word that starts with the root *subsidi*.

Food

ANY(*alimentos, comida*) AND OR(*precio, reforma*)

Additional Iteration

After consultation with domain experts an additional term was included because of its relevance in the design of the reform

tarjeta

3.3 Time Period

Tweets were extracted from the following periods corresponding to the dates of the surveys conducted by *La Prensa Grafica*, the week the subsidy was first introduced as well as a control week September 2013:

- Jan 2011
- May 2011
- August 2011
- 1st-7th September 2011
- May 2012
- August 2012
- September 2013

3.4 Geolocation

Individual tweets were geolocated to a country level using string matching on the user's declared location and comparing to an open-source database of place names². In addition a small proportion of accounts include an automated GPS location which was extracted and the corresponding country was identified. Only content which was identified to have originated from El Salvador was included.

3.5 Crowdsourcing

In order to identify the subject matter and the sentiment expressed in the social media content, it is necessary to first identify by hand the subject matter and sentiment in a

²<http://www.geonames.org/>

subset of the tweets so that we may automate the classification of the text of the rest of the tweets. The manual identification of the subjects and sentiment is a process known as labeling. The labels in this case are the subjects and sentiments that are assigned to each tweet. The automatic classification can be done via *supervised learning*, a topic discussed in greater detail in Section 5.2. Using a proportion of the content labeled by hand, a suitable computer algorithm examines these labeled examples and constructs rules that can classify new text content. Again, this process is discussed in greater detail below. First, however, it is necessary to understand this process of labeling some of the tweets.

The first step is to decide how many tweets we need to classify and to select a reasonably representative sub-sample. There is a large literature on optimal experimental design for classification problems [Onwuegbuzie and Collins, 2007, Figueroa et al., 2012, Beleites et al., 2013]. We chose, however, to apply some admittedly *ad hoc* heuristics to produce a representative sample for labeling. We had three goals in mind when selecting relevant tweets. First, we wanted to make sure that each topic is sufficiently represented. That is, we wanted to avoid giving the labelers all irrelevant tweets. Second, we wanted to make sure that the *vocabulary* of the sample was approximately as big as the *vocabulary* of all of the tweets. We introduce and discuss the *vocabulary* in greater detail below, but the general idea is that we wanted the language in the labeled sample to be the same as that in the unlabeled sample. Otherwise, we would not be able to predict the meaning of new words in the unlabeled sample. Third, we wanted each category to have about 100 tweets. We assumed, based on past experience, that this would give us enough information to accurately identify the content of the tweets.

Through some preliminary exploratory analysis, we identified keywords and features in our tweets that should give us a good chance of selecting a tweet from a certain category, including irrelevant tweets. We then randomly selected tweets using our heuristics, making sure that each time period is represented equally. In the end, we selected about 30% of our sample to be labeled and the results of this labeling conform to expectations vis-a-vis coverage of the categories. We are confident that our sub-sample is adequately representative. The results of the labeling are discussed in the next section. First, we describe our strategy for having the tweets manually labeled.

In this case we have two separate classification tasks. The first is to classify the tweets collected for the analysis based on broad categories that encompass the majority of the social media content. These categories are the following:

- Lack of Information
- Partisanship

- Distrust of Institutions
- Personal economic impact
- Other
- Irrelevant

These categories were decided upon based on the *La Prensa Grafica* survey and some preliminary exploratory analysis³. The second task is to classify the sentiment regarding the reform expressed in the tweets according to the following:

- Strongly positive
- Positive
- Neutral
- Negative
- Strongly negative

In order to label this content, the tweets were uploaded to Amazon Mechanical Turk⁴, a crowdsourcing platform whereby tasks may be completed by distributed teams through an online marketplace.

Mechanical Turk offers several standard template tasks including labeling of images, assigning sentiment to a piece of text, etc. However, these templates are not very flexible. In our case we wanted to have all the instructions in Spanish not to exclude non-English speakers. Also it was necessary to allow a tweet to fall into more than one category, this was not allowed in the standard template.

Therefore, we decided to create a custom task. In this case the user designs the task with questions, tick/check boxes and instructions either with a user interface or directly adding in the HTML code. The disadvantage of this is that the custom task has an extra overhead. In order to create a task with the text of all the tweets in an automated way, a simple script was created to paste the header and footer HTML code as text together with the tweet text.

Creation of the task required a title for the Human Intelligence Task (HIT), a set of instructions and a means to record the results. The instructions are available in Appendix A. A time estimate must be provided (it is recommend to be generous with this time

³This preliminary analysis included the use of topic models fit via Latent Dirichlet Allocation (LDA) Blei et al. [2003]. LDA is a clustering algorithm of sorts that helps discover the different “topics” contained a collection of documents. This approach was also used during the taxonomy refinement stage to discover new keywords. More details are available upon request.

⁴<https://www.mturk.com/>

estimate), a reward (\$US) and any constraints on the user (e.g., the user must have completed such a task before). The potential users look at the HITs on offer and see a preview of the task. We suspect that the instructions are too complex and users are hesitant to accept the task.

Several test tasks were offered, but none were accepted. It is necessary to decide the optimal price, time estimate and number of tweets to be tagged in each task (a small task vs a big task).

- \$16, 3h (100 tweets)
- \$4, 1h (10 tweets)
- \$6, 45m (10 tweets)
- \$10, 45m (10 tweets)
- \$20, 45m (10 tweets)

We concluded that the task was too complex and that the preview would deter potential Turkers from accepting the task and thus not to be suitable for this platform. Typically crowdsourcing platforms support simple tasks such as identifying the gender of a person from an image of their face.

Ideally, we would have two or more labelers look at the same tweets and decide to which category they belong to and which sentiment they show, keeping only those tweets for which the labelers agree. However, due to time and resource constraints, we decided instead that a pair of Spanish speaking domain experts would classify about 500 tweets each and that these would be used as training data. This has the advantage that the labelers have a high affinity with the task, and, therefore, we believe, will give consistent labels across non-overlapping samples.

4 Tagged Tweets

This section briefly describes the results of having our domain experts complete the labeling task described above and further in Appendix A. In all 931 tweets were labeled. These 931 tweets were assigned to 995 different categories. Table 1 gives an overview of the subject distribution over time. Admittedly, it is difficult to draw any strong conclusions from the tagged tweets alone. It is possible that random sub-sample of tweets selected to be tagged is not wholly representative of all the tweets in the period. Furthermore, given the sample sizes it is difficult in some cases to say that from period to

Manually Tagged Tweets: Subjects							
Date	Count	Distrust Institutions	Irrelevant	Lack of information	Other	Partisan- ship	Personal economic impact
Jan 2011	142	24.2	8.9	12.1	45.2	20.2	10.5
Apr 2011	142	24.7	7.7	13.4	27.5	19.0	22.5
May 2011	142	21.1	11.3	6.0	44.4	13.5	18.0
Aug 2011	91	22.6	29.8	9.5	36.8	6.0	8.3
May 2012	142	10.1	42.4	5.0	28.8	14.4	5.0
Aug 2012	131	27.1	32.9	15.3	64.7	16.5	10.6
Sep 2013	141	10.1	23.9	5.8	58.0	5.8	2.2
Total	931	19.3	21.5	9.2	42.6	13.8	11.2

Table 1: Results of manually categorizing the tweets by the domain experts. Numbers are percentages except the counts column. Rows do not sum to 100%, as tweets can be given more than one category. Source: Author’s calculations.

period we have significant changes. Nevertheless, we make a few observations of trends that conform to the discussion in Section 2.1 of [Calvo et al. \[2014\]](#).

First, we see a general decline in the “personal economic impact” category starting around May 2011. This mirrors the change in opinion regarding the subsidy reform observed in the survey data that was thought to be driven in part by the initial belief that the changes would not benefit everyone. Second, we see a similar decline in the “lack of information” and the “distrust of institutions” categories. However, August 2012 shows an increase in both of these categories to previous levels. We will have to wait until we have labeled all of the tweets to be sure that this is not a sampling aberration. These observations are discussed further in Section 6.

Table 2 contains the results of the second tagging task for identifying sentiment expressed in tweets. The main takeaway from this exercise is that people do not seem to use Twitter to express approval. It is, of course, possible that the design for choosing tweets to classify for sentiment was poor. However, given the reasonably large number of tweets this seems unlikely. Overall, it appears that only around 3% of the labeled tweets contained any positive sentiment. This very large class imbalance makes classification of any further existing positive sentiment tweets difficult. We note some avenues for further research in this direction in Section 7. The paucity of positive sentiment tweets, aside, it does appear as though the number of Negative and Strongly Negative sentiment tweets decline somewhat over the period while those tweets that do not express any sentiment is increasing. This may be construed as increasing approval, though any conclusions should wait until we classify all of the remaining tweets.

Manually Tagged Tweets: Sentiment

Date	Count	Strongly Negative	Negative	Neutral	Positive	Strongly positive
Jan 2011	142	7.7	35.2	50.7	5.6	0.1
Apr 2011	142	11.9	47.2	38.0	2.1	0.1
May 2011	142	7.0	48.6	39.4	2.1	2.8
Aug 2011	91	9.9	44.0	44.0	2.2	0.0
May 2012	142	10.5	34.5	51.4	3.5	0.0
Aug 2012	131	0.1	34.3	61.1	3.1	0.1
Sep 2013	141	5.0	15.6	78.7	0.1	0.0
Total	931	7.5	36.7	52.2	2.8	0.1

Table 2: Results of manually tagging tweets for sentiment. Numbers are percentages except the counts column. Source: Author’s calculations.

We now turn to the issue of classifying the remaining tweets and describe our methodology for doing so.

5 Methodology

In this section, we describe the steps necessary to estimate and track both the subject matter of tweets and sentiment over time. First, we will describe how to represent text documents so that we can perform estimation.

5.1 Representation of Text Documents

Given a *corpus* of tweets, or *documents*, the first step for a text classification task is to transform each document into a feature vector that can be used in a classification algorithm. To do so, we make use of the traditional bag-of-words assumption. This assumption holds that the order in which words occur in a document is not very important in classifying the content of that document. Starting from this assumption, we remove all of the punctuation and digits from each document and normalize the unicode characters⁵.

At this point, we also perform some *feature engineering* based on some prior assumptions and some iterations based on our results. Feature engineering is, loosely defined, the process of extracting or creating useful features from our data. To this end, we have transformed several specific words into concepts. For example, we transformed

⁵See the discussion provided by The Unicode Consortium <http://www.unicode.org/faq/normalization.html>

Stemming Example

word	stem	word	stem
chetumal	chetumal	toreado	tor
chetumaleños	chetumaleñ	toreándolo	tor
chiapas	chiap	toreó	tore
chicharrones	chicharron	torrenciales	torrencial

Table 3: Example of Spanish stemmer. Source: Snowball project.

each mention of a currency value to a placeholder, DOLLAR_AMOUNT, under the assumption that there is some valuable information in the mentioning of a value if not the value itself. All positive and negative emoticons are transformed in to POS_EMOTICON and NEG_EMOTICON, respectively. We transform all @ mentions to AT_MENTION. Finally, we transform all links to a URL_LINK placeholder.

Next, we stem the words using the Spanish stemmer from the Snowball project⁶. Stemming is the process of reducing words to their root form and is used primarily to reduce the number of features and document sparseness as discussed below. Table 3 contains some examples from the Snowball project documentation for the Spanish stemmer.

After removing the punctuation and stemming, we split each tweet on spaces, creating tokens or n -grams where n can be any value. If $n = 1$, the tokens are unigrams. For $n = 2$, bigrams. For $n = 3$, trigrams, and so on. The use of tokens longer than unigrams can help preserve semantic meaning for phrases where the bag-of-words assumption might be unduly restrictive. For example, we might want to preserve the semantic meaning of phrases such as “not good” by using a bigram. Whether or not higher-order n -grams improve classification is an empirical question and is discussed further below.

The last important piece for the creation of the feature vector representation is selecting a *vocabulary*, V . The *vocabulary* can be thought of as the words we believe will allow a learning algorithm to discern the class of a document. It is not atypical in text classification problems for the size of the vocabulary, $P = |V|$, to exceed the number of observations, or samples, N , resulting in an underdetermined problem. Not all classification algorithms are able to handle this situation, so vocabulary selection can become quite important for these estimators. For this study, we remove Spanish *stop words* such as *la*, *en*, *y*, and *los*, any term that occurs in fewer than 3 tweets and any term that occurs in occurs in more than 70% of the tweets. This leaves us with a vocabulary size

⁶The main project site is <http://snowball.tartarus.org/index.php>. We used the Python bindings from PyStemmer <https://github.com/snowballstem/pystemmer>.

Example Tweets

1. This new computer is great.
2. Really upset with the President’s newest economic policies.

Table 4: Example of two fictional tweets.

Sample Vocabulary

1. new
2. computer
3. great
4. really
5. upset
6. president
7. economic
8. policies

Table 5: Sample vocabulary for tweets in Table 4.

of $P = |V| = 1354$ and $N = 995$ labeled tweets. While further strategies for vocabulary selection are available, we use regularization methods appropriate for underdetermined problems to select further the features that discriminate between classes as discussed below.

We present a concrete example of feature vector creation in Table 4. Consider the following set of fictional tweets

Ignoring for the time being stemming, Table 5 represents a possible unigram-only vocabulary for these tweets.

The feature vector representation of each of these tweets could be the counts of each vocabulary word in each tweet

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad (1)$$

This example serves to illustrate several features of text classification. First, as mentioned above we have a high-dimensional input. Second, we have few irrelevant features. Given the removal of stop-words, a high majority of the remaining features will contain information that helps discriminate between classes. However, while the class features are dense, each observation is sparse. Only a few of the features will occur in any given observation. Any learning algorithm used to classify documents must be well suited to handle these characteristics.

While the representation in 1 is in terms of the counts, or term frequencies, of the text in each tweet, there are other alternative representations to consider. Two other possibilities are binary indicators of terms and term-frequency inverse-document-frequency (*tf-idf*). As tweets are not very likely to repeat words given the 140 character limit, term frequencies and binary indicators are not likely to vary much, so we do not consider binary indicators further. One potential problem with just using the term frequencies is that each term is considered equally important when in fact some may have little to offer in terms of distinguishing the content of a document. One remedy is to scale the term frequency by the *inverse document frequency* of the term. The document frequency of the term is simply the number of documents in which the term occurs. The inverse document frequency is computed

$$idf = \log \frac{N}{df_t} + 1 \quad (2)$$

where N is the number of documents in the corpus and df_t is the document frequency of term t . The *idf* is larger for rarer terms and smaller for more frequent terms and, thus, gives the desired downweighting effect. In turn, the *tf-idf* is computed as

$$tf-idf = tf * (idf + 1) \quad (3)$$

There are several different definitions used for *tf-idf*. In 3, *tf* may simply be the counts of the terms or a binary indicator as mentioned above. However, one may also calculate the logarithmically scaled *tf* as $1 + \log(tf)$ of the counts. Finally, the *tf-idf* representation of each document may optionally be normalized. When applied, common normalization schemes include dividing by the ℓ_1 or ℓ_2 norm. The ℓ_2 norm is particularly popular as it transforms each document into a unit vector and allows computing the cosine similarity between documents by a simple dot product.

In practice, using *tf-idf* on short texts such as tweets may result in noisy *tf-idf* numbers. However, which technique is most appropriate is something we will assess empirically. In Section 5.3, we address the question of how to choose which transformation to use and when. In the next section, we describe the estimators used for classification.

Before moving on, we show the results of applying the *tf-idf* transformation to our labeled data. Table 6 contains the top 20 unigram and bigram stems by *tf-idf* applied to each subject. While there are some noisy, non-informative words present, many of the words conform to what we would expect *a priori* to be discussed in these subjects. For example, the “distrust of institutions” category contains stems for *gobierno*, *focalizacion*,

Top 20 n-grams: Subjects

Distrust institutions	Irrelevant	Lack of information	Other	Partisan- ship	Personal economic impact
que	subsidi	que	que	que	me
no	dollar_amount	at_mention	dollar_amount	comedor	luz
at_mention	no	me	no	no	dollar_amount
gobiern	com	quit	pag	aren	no
gast	tarjet	cuand	recib	gas	que
focaliz	que	va	ser	popul	asi
hay	preci	no	agu	derech	sub
esta	gasolin	format	nuev	fmln	preci
tod	fmln	car	at_mention	ser	pag
com	gast	recib	tarjet	ha	hay
mal	tien	si	ahorr	manten	quit
dollar_amount	mas	cambi	millon	el	ayud
deb	total	sac	dic	ministr	gt
val	dic	part	propan	fun	hitl
pag	sub	chic	hast	pid	despues
sin	aren	caj	han	at_mention	ahorr
cuand	url_link	ya	ni	mas	twitt
nuev	regul	presidencial	url_link	gobiern	lleg
estan	empresari	pag	electr	dic	suicid
recib	me	general	tod	elimin	ja

Table 6: The top 20 unigrams and bigrams by *tf-idf* for each subject from our tagged tweets. Source: Author’s calculations.

and *recibir*. The “personal economic impact” category contains words such as *me*, *luz*, and *pagar*, and *precio*. There are no bigrams in this table. Quite often, bigrams are not repeated very often in a document or a corpus, so they are not very high in terms of *tf-idf*⁷

Table 7 shows the same results broken down by sentiment. These results are not nearly as illuminating. The word “no” occurs in almost every category except “positive.” We also see a few bigrams in the “Positive” and “Strongly Positive” categories. These are clearly not very general, and reflect the lack of information rather than any true sentiment content.

This lack of coherence demonstrates an important concept in text classification—co-occurrence. It is not the single occurrence of a word that dictates how a classifier identifies

⁷While words that occur in only one document receive a high *idf* score, the term frequency is so low, the composite *tf-idf* is low.

Top 20 n-grams: Sentiment

Strongly negative	Negative	Neutral	Positive	Strongly positive
no	que	que	util	luz
que	no	dollar_amount	gener	men
nos	gas	no	banc	cos
luz	dollar_amount	gas	dollar_amount	gran
mas	mas	url_link	propan	jajajaj
hay	com	tarjet	form	dollar_amount
gas	sub	com	millon	sol
porqu	si	dic	pag	luz gran
leñ	ni	at_mention	gas	quier
dollar_amount	pag	fmln	que	jajajaj dollar_amount
pag	at_mention	electr	subsidi dadan	ciu- mirecib
verg	tien	me	bien	at_mention
propan	gast	nuev	recib	me
cocin	gasolin	propan	ciudadan	cos sol
graci	gobiern	ser	cobr	preci
electr	popul	recib	nuev	luz at_mention
per	cuand	deb	ahorr	quier subsidi
com	aren	per	luz	sol men
preci	porqu	gobiern	url_link	men subsidi
dic	tamb	hoy	me	no

Table 7: The top 20 unigrams and bigrams by *tf-idf* for each sentiment from our tagged tweets. Source: Author’s calculations.

the contents of a document, it is the co-occurrence of several words together. We will now describe the classifiers used in more detail before presenting our results.

5.2 Classifiers

A large portion of the machine learning literature focuses on classification tasks using text data. Generally speaking, given a set of input text and associated classes, a classifier finds the relationship between the text and the class of text. The classifier may then be used to predict the class of new documents for which the class is not yet known. A few examples of applications include detecting whether an e-mail is spam or not, identifying the language of a document, the subject of a document, or the relevance of a document given a search query.

For the present purposes, we are interested in binary and multi-class classifiers. The probit and logit models are examples of binary classifiers familiar to econometricians. The multinomial logit is an example of a multi-class classifier. It assigns each observation, or *sample* in machine learning parlance, to one of several mutually exclusive categories. While we truly have what is known as multi-label data⁸ in this setting—samples do not have to belong to only one category—for simplicity we have chosen to approach the problem as a multi-class classification task. Many of the observations have the “other” category as their second (or third, etc.) category. These labels are simply discarded. Any tweets that are assigned more than one subject that is not “other” are treated as two separate observations with two separate *target values*. The term *target values* here refers to the outcome variable. It is sometimes called a *label*. This class of problems belongs to a broader type of machine learning problem known as *supervised learning*. That is, the target values are known for some set of the data in contrast to *unsupervised learning* tasks in which the classes of the data are unknown. Clustering is an example of an unsupervised learning tasks.

It is common in the machine learning literature to approach the multiclass problem as a combination of several binary choice problems. K different classifiers are built, one for each outcome class, and for the i^{th} class, the positive labels are those observations belonging to that class while the negative labels are all other classes. This is referred to as a One-versus-All (OVA), or One-versus-Rest, approach. Other approaches include a One-versus-One (OVO), or All-versus-All, approach where we build $\binom{K}{2} = K(K-1)/2$ classifiers to distinguish each pair of classes. More exotic approaches exist, but there is

⁸The econometrics literature uses the term multiple-response categorical variables (MRCVs) [Bilders and Loughin, 2003].

typically little gained from more complicated approaches in terms of accuracy [Hsu and Lin, 2002].

It is not possible to know *a priori* which classifier will perform best for any given task. As such, we explore the use of five estimators. The first four learning algorithms used fit within the Stochastic Gradient Descent (SGD) algorithm. The models that can be trained using SGD take the following general form. We have some binary target data Y where $y_i \in -1, 1$ and an input vector X where $X_i \in \mathbb{R}^p$. Using a linear predictor function

$$f(X) = \beta' X. \quad (4)$$

We seek β such that we minimize the training error as a function of the loss function L and a regularization term R , which penalizes model complexity by pushing the β coefficients to or towards zero. That is

$$\arg \min_{\beta} = \frac{1}{n} \sum_{i=1} nL(y_i, f(x_i)) + \alpha R(\beta) \quad (5)$$

where α is a non-negative hyperparameter controlling the strength of the regularization. SGD itself is a robust, performant optimization algorithm. The four learning algorithms solved via SGD, therefore, differ only in the loss function. These four learning algorithms are linear Support Vector Machines (SVMs), logistic regression, the modified Huber loss function [Zhang, 2004], and the perceptron.

The SVM loss, or hinge loss, is

$$L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i)) \quad (6)$$

The logistic loss function is

$$L(y_i, f(x_i)) = \log(1 + \exp(-y_i f(x_i))) \quad (7)$$

The modified Huber loss function is

$$L(y_i, f(x_i)) = \begin{cases} \max(0, 1 - y_i f(x_i))^2 & \text{if } y_i f(x_i) \geq -1 \\ -4f(x_i)y & \text{otherwise} \end{cases} \quad (8)$$

The perceptron loss is a slight modification of the SVM

$$L(y_i, f(x_i)) = \max(0, -y_i f(x_i)) \quad (9)$$

For each of these loss functions, we also vary the regularization method, considering the ℓ_1 norm, or lasso, which is able to shrink coefficients to exactly zero,

$$R(\beta) = \sum_{i=1}^p |w_i| \quad (10)$$

The lasso will select at most p non-zero coefficients in the case where $p > n$. This could be limiting. The ℓ_2 norm, otherwise known as Ridge Regression, on the other hand, shrinks coefficients towards zero⁹,

$$R(\beta) = \frac{1}{2} \sum_{i=1}^p w_i^2 \quad (11)$$

The final regularization penalty considered is the elastic net, which is a weighted combination of both norms

$$R(\beta) = \frac{\rho}{2} \sum_{i=1}^p w_i^2 + (1 - \rho) \sum_{i=1}^p |w_i|. \quad (12)$$

The elastic net tends to work well when there are groups of highly correlated variables.

The final classifier considered is the naive Bayes classifier. The naive Bayes classifier is a simple application of Bayes' theorem under the "naive" assumption of independence of the features. Given Bayes' theorem

$$P(y|x_1, \dots, x_p) = \frac{P(y)P(x_1, \dots, x_p|y)}{P(x_1, \dots, x_p)} \quad (13)$$

the (surely wrong) independence assumption implies that

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p) = P(x_i|y) \quad (14)$$

Using this assumption Bayes' theorem simplifies to

⁹This is an important distinction. No model coefficients will be set to exactly zero using the ℓ_2 norm.

$$P(y|x_1, \dots, x_p) = \frac{P(y) \prod_{i=1}^p P(x_i|y)}{P(x_1, \dots, x_p)} a \propto P(y) \prod_{i=1}^p P(x_i|y) \quad (15)$$

This latter term is our classifier. We can calculate the *maximum a posteriori* (MAP) estimates of both terms. The MAP estimate of $P(y)$ is given by the observed frequencies. The MAP estimate of $P(x_i|y)$ is found by assuming y is multinomial distributed such that

$$y_k \sim MN(\theta_k, N_k)$$

for each class k . The parameters are estimated via smoothed maximum likelihood (relative frequency counting)

$$\hat{\theta}_{ki} = \frac{N_{ki} + \alpha}{N_k + \alpha|V|}$$

where N_{ki} is the number of times term i appears in an observation of class k . N_k is the total number of terms in class k . $|V|$ is the vocabulary size as above. The α term is a smoothing parameter to avoid the division by zero problem. For $\alpha = 1$ this is Laplace smoothing, and $\alpha \in [0, 1)$ is known as Lidstone smoothing,

Given this set of potential classifiers, we must select the “best” classifier and the appropriate model parameters for the classifier. In the current setting “best” means the estimator that avoids overfitting and generalizes to give the best out-of-sample predictive power. We assess this via a cross-validation scheme, which we describe in the following sub-section.

5.3 Evaluating Classification

The evaluation of potential classifiers is done via cross-validation. This involves splitting the labeled dataset—the tweets that were manually tagged—into a training and a testing set. We fit the classifier on the training data and assesses its predictive performance on the held-out testing data to get a sense of its out-of-sample performance. This is done a number of times and the performance metric of each fit is averaged. In addition, if researchers are particularly data-rich, you might first split the data in to a training and a holdout set, perform cross-validation on the training set, and then judge the performance on the holdout set, which has never been seen by the learning algorithm. This gives a

		Predicted class		
		1	0	Total
True value	1	True Positive	False Negative	$tp + fn$
	0	False Positive	True Negative	$fp + tn$
Total		$tp + fp$	$fn + tn$	

Table 8: An illustration of the terms that make up a binary confusion matrix. Multi-class confusion matrices are described in the same way. Source: Author’s calculations.

reasonable assurances that we have avoided overfitting the sample data and will have good generalization performance for the unlabeled tweets. Before discussing the results of this out-of-sample prediction, we describe the cross-validation approach used.

There are a number of different strategies for splitting the data to apply cross-validation. For this exercise, we use stratified K-folds cross-validation with $K = 5$. The data is split in to 4 *folds* with the final fold being the complement of the rest. The stratified qualifier indicates that the percentage of each class in the dataset is preserved in each sub-sample. The algorithm is trained on the complement of each single fold and then a score is computed for this single held-out fold. We now discuss the choice of score function.

In choosing a score function, it is necessary for the researcher to identify what is most important criteria for the task. Several metrics are available, which are based on the confusion matrix. Table 8 shows a confusion matrix for a binary classification problem. There are four common measures based on the confusion matrix that can be generalized to the multiclass classification problem. Sensitivity, or the true positive rate or recall, measures the number of observations correctly identified as belonging to that class out of all that truly belong to that class

$$TPR = \frac{TP}{TP + FN} \quad (16)$$

Specificity, or the true negative rate, measures the number of observations correctly

identified as not belonging to that class out of all that truly belong to that class

$$TNR = \frac{TN}{FP + TN} \quad (17)$$

Precision, or positive predictive value, is the number of correctly identified observations belonging to that class out all predicted as belonging to that class

$$PPV = \frac{TP}{TP + FP} \quad (18)$$

Negative predictive value is

$$NPV = \frac{TN}{TN + FN} \quad (19)$$

The F_1 -measure is a measure of accuracy that combines precision and sensitivity. Is is defined as the harmonic mean between the two measures

$$F_1 = 2 \frac{PPV \cdot TPR}{PPV + TPR} \quad (20)$$

Another common measure is a generalization of the F_1 -measure called the F_β -measure. It allows researchers to put differing weights on precision and recall

$$F_\beta = (1 + \beta^2) \cdot \frac{(1 + \beta^2) \cdot PPV \cdot TPR}{\beta^2 PPV + TPR} \quad (21)$$

The results presented in the next section are based on the F_1 -measure to balance our desire for both high precision and high recall.

6 Results

6.1 Cross-Validation Results

We performed 5-fold cross-validation to select the best transformation and estimator. For the transformation to the feature vector, we considered both unigrams and bigrams, binary indicators, counts, and *tf-idf* of the n -grams, as well as ℓ_1 , ℓ_2 , and no normalization for each document. For each of the SGD classifiers, we ran 100 iterations. We

varied the α parameter and the regularization penalty function, trying the ℓ_1 , ℓ_2 , and the elastic net. We set α to a grid of size 25 from $1e - 6$ to 1000 in log space. For the elastic net, we let the ρ be $[\text{.05}, \text{.15}, \text{.25}, \text{.5}, \text{.75}, \text{.85}, \text{.95}]$. Finally, we also varied the weights for each class. We tried both without weights and setting the weight of each class to the inverse of its observed frequency given that we do not observe a uniform distribution of classes in either the categories or the sentiment. For the Naive Bayes classifier, we used the same feature vector transformation options, and we used a grid of size 10 from .1 to 1 for α .

Using the $f1$ -measure to evaluate performance, we select a feature vector transformation that uses only the frequency of unigrams rather than *tf-idf* for the subject matter classification. The chosen classifier is the SGD classifier with the modified Huber loss function and an ℓ_2 regularization penalty with $\alpha \approx .1$. We also select to use the class weights that are inversely proportional to the observed class frequency.

For the sentiment feature vector transformation, we select the *tf-idf* of unigrams with an ℓ_2 normalization for each tweet. For the classifier, we select the SGD classifier with the hinge loss function and an ℓ_2 regularization penalty with $\alpha = .01$. All classification was performed using the scikit-learn library for Python version 0.15.2.[[Pedregosa et al., 2011](#)]. Any parameter not mentioned was left at its default value.

6.2 General Classification Results

Table 9 shows the results of using the classifier to predict the subject of the tweets over time. It is similar to Table 1 with a notable exception. There is a smaller percentage of tweets categorized as “distrust of institutions.” We do not put too much weight on the August 2011 and August 2012 months due to the small sample¹⁰. Only a few misclassified results will change the percentage considerably. However, we do not have these problems with the other months. We also still observe the increase in the months of April and May of 2011, confirming the survey results from *La Prensa Grafica* reported in [Calvo et al. \[2014\]](#).

Similarly, we observe a drop-off in the predicted “lack of information” and “personal economic impact” categories with the same caveat about August 2011 and 2012. This gives us confidence that these results are capturing general public opinion about the gas subsidy reform. Table 10 presents select data from the *La Prensa Grafica* [[Calvo et al., 2014](#)]. We see a similar trend when we look at the “distrust of institutions”, “lack of

¹⁰We speculate that perhaps August is month of low Twitter usage in general owing in part to the *Fiestas Agostinas*.

Classification Results: Subjects

Date	Count	Distrust institutions	Irrelevant	Lack of Information	Other	Partisan- ship	Personal economic impact
Jan 2011	275	6.9	4.7	10.9	52.0	18.5	6.9
Apr 2011	863	14.1	7.9	9.0	47.0	6.0	15.9
May 2011	310	14.5	12.6	7.4	43.2	7.7	14.5
Aug 2011	118	11.9	27.1	7.6	34.7	8.5	10.2
May 2012	570	5.4	36.1	3.3	39.8	11.1	4.2
Aug 2012	168	7.1	20.8	10.1	50.0	4.2	7.7
Sep 2013	680	3.8	27.5	4.6	58.4	3.7	2.1

Table 9: The classification results for the categories of all the tweets. Source: Author’s calculations.

information”, and “personal economic impact” categories. As public sentiment is shifting in favor of the subsidy reform, people mention these categories less. As we see below, the appearance of these tweets that fall within these categories are almost always negative, so “no news is good news” here.

The results in Table 11 for the sentiment analysis are not as clear. Given the extraordinarily high class imbalance at the expense of the positive category, the classifier is unable to predict even one positive category in- or out-of-sample. This is likely an artifact of using the SGD algorithm, which will not do well on highly imbalanced classification tasks with rare events and also the lack of discriminating information in the small number of positive training examples. However, we do notice a decline in overall negative sentiment coinciding with the change in the survey sentiment. However, as we see in Table 12, this is mainly due to the increase in the “other” and “irrelevant” categories, which generally have fewer tweets that express any sentiment.

One of the benefits of using Twitter data is that we can take a deeper look and identify what exactly is driving these results. Given the nature of surveys, it is often prohibitively costly if not impossible to ask different questions after a general picture emerges from the collection of an original survey. That is, with surveys it is much more important to get it right the first time. Figures 2-9 give a general picture of the coefficients that are important in predicting whether or not a tweet belongs to a certain category or expresses a certain sentiment. These are the coefficients in Equation 4. The use of words with positive coefficients suggests that the tweet belongs to that class. The negative coefficients suggest that the tweet belongs to some other class in the OvA scheme.

La Prensa Grafica Survey Answers

Date	Being satisfied conditioned on support of ARENA party	Being satisfied conditioned on support of FMLN party	% answering “satisfied” or “very satisfied”
Jan 2011	18.8	44.1	30.0
May 2011	33.8	57.7	43.2
Aug 2011	44.2	50.5	44.9
May 2012	42.1	57.8	50.2
Aug 2012	52.7	76.9	66.0
Sep 2013	55.0	71.3	64.3

Table 10: Selected questions from the *La Prensa Grafica* survey as reported in [Calvo et al. \[2014\]](#). All numbers are percentages. We observe similar timings in the shifts in the topics being discussed on Twitter with respect to the gas subsidy. Source: [Calvo et al. \[2014\]](#)

Classification Results: Sentiment

Date	Count	Negative	Neutral	Positive
January 2011	275	44.4	55.6	0.0
April 2011	863	59.4	40.6	0.0
May 2011	310	63.9	36.1	0.0
August 2011	118	64.4	35.6	0.0
May 2012	570	37.4	62.6	0.0
August 2012	168	43.5	56.5	0.0
September 2013	680	19.0	81.0	0.0

Table 11: The classification results for the sentiment of all the tweets. Source: Author’s calculations.

Classification Results: Sentiment by Category

Date	Category	Negative	Neutral	Positive
January 2011	Distrust institutions	94.7	5.3	0.0
April 2011	Distrust institutions	98.4	1.6	0.0
May 2011	Distrust institutions	97.8	2.2	0.0
August 2011	Distrust institutions	92.9	7.1	0.0
May 2012	Distrust institutions	100.0	0.0	0.0
August 2012	Distrust institutions	100.0	0.0	0.0
September 2013	Distrust institutions	100.0	0.0	0.0
January 2011	Irrelevant	38.5	61.5	0.0
April 2011	Irrelevant	66.2	33.8	0.0
May 2011	Irrelevant	66.7	33.3	0.0
August 2011	Irrelevant	62.5	37.5	0.0
May 2012	Irrelevant	41.3	58.7	0.0
August 2012	Irrelevant	45.7	54.3	0.0
September 2013	Irrelevant	11.8	88.2	0.0
January 2011	Lack of information	96.7	3.3	0.0
April 2011	Lack of information	96.2	3.8	0.0
May 2011	Lack of information	100.0	0.0	0.0
August 2011	Lack of information	100.0	0.0	0.0
May 2012	Lack of information	89.5	10.5	0.0
August 2012	Lack of information	94.1	5.9	0.0
September 2013	Lack of information	80.6	19.4	0.0
January 2011	Other	18.2	81.8	0.0
April 2011	Other	31.3	68.7	0.0
May 2011	Other	33.6	66.4	0.0
August 2011	Other	34.1	65.9	0.0
May 2012	Other	16.3	83.7	0.0
August 2012	Other	15.5	84.5	0.0
September 2013	Other	6.3	93.7	0.0
January 2011	Partisanship	49.0	51.0	0.0
April 2011	Partisanship	42.3	57.7	0.0
May 2011	Partisanship	79.2	20.8	0.0
August 2011	Partisanship	100.0	0.0	0.0
May 2012	Partisanship	30.2	69.8	0.0
August 2012	Partisanship	42.9	57.1	0.0
September 2013	Partisanship	76.0	24.0	0.0
January 2011	Personal economic impact	100.0	0.0	0.0
April 2011	Personal economic impact	90.5	9.5	0.0
May 2011	Personal economic impact	91.1	8.9	0.0
August 2011	Personal economic impact	83.3	16.7	0.0
May 2012	Personal economic impact	100.0	0.0	0.0
August 2012	Personal economic impact	100.0	0.0	0.0
September 2013	Personal economic impact	85.7	14.3	0.0

Table 12: The classification results for the sentiment of all the tweets by category. Source: Author's calculations.

We can compare these figures to Tables 6 and 7, to get a sense of change between the insights given by the *tf-idf* measure versus using a more sophisticated classifier. In what follows, we focus on the positive coefficients. Given the use of the OVA classification strategy, the negative coefficients indicate that those words tend to show up together in unrelated tweets. So the negative coefficients in one class are just the positive coefficients for all the other classes.

Figure 2 gives a sense of what those tweets that are classified as distrusting of institutions contain. First and foremost, they address or mention the government. The second most mentioned institution is that of the gas distributors. The stem *form* indicates that people are distrustful of this new way of receiving the subsidy. Similarly, *quit* indicates that people think the government is no longer offering the subsidy. Unsurprisingly, this token also has a high weight in the “Lack of information” category seen in Figure 3. Furthermore, the terms *leña* and *cocinar* indicate a concern with the use of firewood for fuel, increasing in the presence of fuel shortages, and those who use wood stoves.

Coefficient Plot: Distrust of Institutions

Figure 2: The top 100 coefficients in absolute value associated with the “distrust of institutions” throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

Figure 3 is characterized by words that indicate doubt (*si*), questions *cuand*, *cuant*, *dond*, or speculation on what is going to happen *van*. An increase in the price of *pupusas* is one particular concern. As noted above, there is some overlap with the “Distrust” category as to be expected. This, again, illustrates the concept of co-occurrence. These words in isolation may indicate a distrust in institutions but seen in the context of tweets demonstrating uncertainty, we are able to identify these tweets as expressing a lack of

information about the reform.

Coefficient Plot: Lack of Information

Figure 3: The top 100 coefficients in absolute value associated with the lack of information throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

Coefficient Plot: Partisanship

Figure 4: The top 100 coefficients in absolute value associated with partisanship throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

The “partisanship” category in Figure 4 contains mainly the names of political parties and politicians as expected. We will not say much further about this category.

Figure 5 contains the terms which indicate the “personal economic impact” category. It is evident that people are greatly concerned (*neg_emoticon*) about changes to their electric bill (*luz*, *kw*, *energ*, *electr*, *dollar_amount*, *mas*). The tweets also lament an increase in the price of tortillas, specifically, and price increases (*altos*) in general that are often mentioned along with the subsidy changes. Many of the tweets that contain *acab* report their experiences with the reform – from reporting to have lost the subsidy entirely to having just paid higher prices for gas.

Coefficient Plot: Personal Economic Impact

Figure 5: The top 100 coefficients in absolute value associated with “personal economic impact” throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

The “other” category contains mainly informational or news tweets, that use more formal language and often contain a link to a news story. Some highlight the millions spent on the subsidy or the million who benefit from it *millon*, *beneficiari*. Some report a change in the mechanism of delivery *nuev*, *mecan*, which is in contrast to the less formal *forma* used above. Similarly, *licuado* (*licu*) is often used to refer to the propane gas itself.

The negative and neutral sentiment tweets in Figures 7 and 8, respectively, are mirror images of one another since this ends up as a binary classifier. The neutral sentiment words tend to be those words associated with the “other” category. This is not unexpected since this category contains mostly informational tweets. The negative tweets themselves do not contain many negative sentiment words, themselves, other than *no* and *neg_emoticon*. This is a consequence of pretty much every non-informational tweet being tagged as a negative tweet. We discuss potential strategies for mitigating this problem in Section 7. It is noteworthy that there are some tokens that express positive

Coefficient Plot: Other

Figure 6: The top 100 coefficients in absolute value associated with the “other” category throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

sentiment in Figure 9 such as *bien* and *pos_emoticon*. The signal is just too low for them to carry enough weight to overwhelm the rest of the language used in the corpus.

Coefficient Plot: Negative Sentiment

Figure 7: The top 100 coefficients in absolute value associated with negative sentiment throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author’s calculations.

We now turn our attention to the change in these use of language over time and

Coefficient Plot: Neutral Sentiment

Figure 8: The top 100 coefficients in absolute value associated with neutral sentiment throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author's calculations.

Coefficient Plot: Positive Sentiment

Figure 9: The top 100 coefficients in absolute value associated with positive sentiment throughout the entire sample. The blue indicates positive coefficients. The red indicates negative coefficients. Source: Author's calculations.

examining how it reflects a change in concerns.

6.3 Examining Topic Drift

Figures 10-14 illustrate the potential for obtaining a deeper understanding of topic drift over time by breaking down the coefficient plots of the above section by time period during which the tweets appeared. We would like to draw your attention to two aspects of these graphs. First, August 2011 and August 2012 demonstrate sparse features for each category except for the “other” category. This could mean one of two things – people simply were not talking about the subsidy reform during these months on Twitter, or we are no longer capturing the tenor of the conversation expressed in our search taxonomy. Given that we observe similarly dense features for each month across the categories, however, it seems that the former is more likely to be true. For example, our taxonomy still captures partisan discussions taking place in May 2012 in the aftermath of the March 2012 congressional elections in El Salvador, but we do not see many partisan tweets during August 2011 or 2012. Furthermore, the “other” category does not demonstrate this same sparsity. This suggests that there is still news being reported about the subsidy, but that people are no longer discussing it. We may tentatively take this as evidence that people do not have anything further about which to complain, so they have ceased to tweet about it.

The second aspect to which, we would draw attention is the rise of the term *distribuidor* in the May 2011 and August 2011 tweets. It seems that though it may be the case that the gas distributor strikes of May 2011 were inconsequential [Calvo et al., 2014], perhaps they influenced public perception in a very negative way. In the next section we will offer some avenues for further improvements on this study and then conclude.

7 Future Work and Improvements

We begin this section with some caveats, then offer some next steps for making potential improvements to the accuracy of our classifier and gaining additional insights in to the subject matter of the tweets about the reforms. For the first caveat, it is worth noting that such novel sources as Twitter data are not yet fully understood. In particular baseline information must be sought and representativity more fully investigated [Tufekci, 2014]. However, we believe the effect of the bias of Twitter towards more affluent demographics is somewhat mitigated since the reform, according to an incidence analysis, was particularly “pro-poor” and benefited the great majority of the population [Tornarolli and Vazquez, 2012].

In addition to the potential sampling bias, the subjectivity bias of the labeling exercise

Coefficient Plot: Partisanship by Date

Figure 10: The top 100 or fewer coefficients associated with partisanship over time. The blue indicates positive coefficients and the red negative. This graph demonstrates topic drift during the sample period. The scales of each graph are not constant over time. Source: Author’s calculations.

is of some concern. As mentioned above, we see *pupusa* as an influential token in the “lack of information” category. However, *tortilla* is a strong indicator of “personal economic impact.” This difference may simply be due to the fact that one or both taggers differed subjectively in their tagging of the category for these tweets or that the difference was truly in the context of the tweet. It would be, however, difficult to argue that generally *pupusa* is a distinguishing token of “lack of information” while *tortilla* indicates personal economic impact. The co-occurrence property of the classifier should help us distinguish when one category is appropriate, but the classifier is only as good as its input. Typically, it is common to have several taggers see the same tweets and to keep those with high interrater agreement to mitigate the potential effects of subjectivity. That said, we argue that having domain experts tag these tweets is already a potential mitigating factor.

These two potential pitfalls aside, there is still more that could be done with this data to potentially improve these results. There are certainly more avenues that could be explored with respect to feature engineering. We might reduce each political party or politician mention to a POLITICS concept. We might retain some punctuation and replace ellipses or exclamation points with some placeholder to indicate IMPLICATION or EXCITEMENT. The presence of an ELLIPSIS and a URL_LINK would almost certainly

Coefficient Plot: Personal Economic Impact by Date

Figure 11: The top 100 or fewer coefficients associated with “personal economic impact” over time. The blue indicates positive coefficients and the red negative. This graph demonstrates topic drift during the sample period. The scales of each graph are not constant over time. Source: Author’s calculations.

indicate an information tweets – what we captured in the “other” category. Instances of some manifestation of laughing (e.g., “Jajaja” or “Jaja”) could be replaced by a LAUGHING concept. This may help capture instances of positive emotion or sarcasm better.

With respect to the sentiment classification, we could use untagged tweets that contain positive emoticons to help identify the features that indicate positive sentiment. Similarly, we are exploring the use of a Spanish language sentiment lexicon, in which psychologists or linguists have identified the psychological valence of a list of emotionally charged words.

The use of bigrams could be further improved. They did not turn out to be very important informationally, beyond marginally improving the cross-validation results. We might instead identify collocations, words that are juxtaposed frequently and have their own meaning, and keep those instead of all bigrams. Similarly, there are still some purely functional stop words that appear in our results that could be removed so that we rely less on the regularization of the algorithm. We could include time fixed effects¹¹, time

¹¹Or perhaps seasonal effects given our observations for August.

Coefficient Plot: Distrust of Institutions by Date

Figure 12: The top 100 or fewer coefficients associated with the “distrust of institutions” over time. The blue indicates positive coefficients and the red negative. This graph demonstrates topic drift during the sample period. The scales of each graph are not constant over time. Source: Author’s calculations.

interactions, or an indicator for a retweet.

Finally, we might take a deeper look in the “other” category and in the August 2011 and 2012 tweets to see if it would be beneficial to further refine our taxonomy and search the Twitter firehose archive again. Given even the small number of “Irrelevant” tweets here, this is unlikely and may simply reflect a secular drop in twitter use during August in El Salvador. Collecting tweets over the entire timeline of interest rather than restricting the tweets to the time of the *La Prensa Grafica* surveys may also help smooth out these statistical aberrations.

8 Conclusion

In this study, we were able to confirm that Twitter can be a valuable complement to existing household survey data. We found that the decrease in negative sentiment tweets concerning several issues surrounding the propane gas subsidy reform in El Salvador coincided with the increase in positive sentiment found by household surveys conducted by

Coefficient Plot: Lack of Information by Date

Figure 13: The top 100 or fewer coefficients associated with “lack of information” over time. The blue indicates positive coefficients and the red negative. This graph demonstrates topic drift during the sample period. The scales of each graph are not constant over time. Source: Author’s calculations.

La Prensa Grafica. Furthermore, we were able to provide deeper insights in to prominent content of these subjects with our results suggesting that the short-lived distributor strikes in May 2011 and the public’s views of the distributors, in general, may have influenced the negative public perception of the reform more than previously acknowledged.

We also provided some methodological suggestions for those researchers wishing to undertake similar studies, noting the difficulties in obtaining a representative, a baseline of tagged tweets, and in making inferences from small samples. We also gave methodological suggestions for improving results by avoiding over-fitting, performing feature engineering, and iterating from preliminary results back to the taxonomy stage. Overall, we were able to provide a more nuanced understanding of the public debate on the El Salvador propane gas subsidy reform.

Coefficient Plot: Other by Date

Figure 14: The top 100 or fewer coefficients associated with “other” over time. The blue indicates positive coefficients and the red negative. This graph demonstrates topic drift during the sample period. The scales of each graph are not constant over time. Source: Author’s calculations.

References

- Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Analytica chimica acta*, 760:25–33, 2013.
- Christopher R Bilder and Thomas M Loughin. Strategies for modeling two categorical variables with multiple category choices. In *American Statistical Association Proceedings of the Section on Survey Research Methods*, pages 560–567, 2003.
- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
- Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
- Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

- O. Calvo, B. Cunha, and R. Trezzi. When winners feel like losers. Technical Report Mimeo, The World Bank, 2014.
- Raquel Fernandez and Dani Rodrik. Resistance to reform: Status quo bias in the presence of individual-specific uncertainty. *The American economic review*, pages 1146–1155, 1991.
- Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12(1):8, 2012.
- Manuel Garcia-Herranz, Esteban Moro, Manuel Cebrian, Nicholas A Christakis, and James H Fowler. Using friends as sensors to detect global-scale contagious outbreaks. *PloS one*, 9(4):e92413, 2014.
- Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.
- Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1021–1024. International World Wide Web Conferences Steering Committee, 2013.
- Maxime Lenormand, Miguel Picornell, Oliva G Cantu-Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frias-Martinez, and Jose J Ramasco. Cross-checking different sources of mobility information. *arXiv preprint arXiv:1404.0333*, 2014.
- Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.
- Andrés Monroy-Hernández, Emre Kiciman, Munmun De Choudhury, Scott Counts, et al. The new war correspondents: The rise of civic media curation in urban warfare. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1443–1452. ACM, 2013.
- Anthony J Onwuegbuzie and Kathleen MT Collins. A typology of mixed methods sampling designs in social science research. *Qualitative Report*, 12(2):281–316, 2007.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *ICWSM*, 11:281–288, 2011.
- Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. Twitter user gender inference using combined analysis of text and image processing. *VEL Net 2014*, page 54, 2014.
- Mark A Stoové and Alisa E Pedrana. Making the most of a brave new world: Opportunities and considerations for using twitter as a public health monitoring tool. *Preventive medicine*, 63:109–111, 2014.
- L. Tornarolli and E. Vazquez. Incidencia distributiva de los subsidio en el salvador. Technical report, Technical report, Interamerican Development Bank, 2012.
- Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*, 2014.
- UNICEF. Tracking anti-vaccination sentiment in eastern european social media networks. Working paper, UNICEF, 2013.
- Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 439–444. International World Wide Web Conferences Steering Committee, 2014.
- Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.

A Appendix

This appendix contains the English and Spanish language instructions that were used to create tasks on Mechanical Turk and eventually given to the domain experts.

Instructions for Categorizing Tweets

Project Name: Propane Gas Subsidy in El Salvador

Category Names:

1. Lack of Information
2. Partisanship
3. Distrust Institutions
4. Personal economic impact
5. Other
6. Irrelevant

General Instructions:

We are assessing the perception of twitter users in El Salvador towards the government's propane gas subsidy program. In April 2011, the government of El Salvador implemented a substantial reform of the subsidy for gas. Before the reform, consumers paid a fixed subsidized price to buy gas bottles (\$5.10), After the reform, the price of the bottles in the shops increased to \$13.60 and, as a compensation, individual households started receiving a transfer of \$8.50 per month in their electricity bill.

We want you to put each tweet into a certain category as described below. Please do not follow any links or @ mentions in the tweet to obtain more context.

Selection Criteria:

Category: Lack of information

Includes: Tweets in which the user expresses confusion over the propane gas subsidy. For example, they do not know why the price of propane gas increased or how to take advantage of the subsidy.

Excludes: Tweets in which the user expresses uncertainty over how the subsidy will affect their lives. These tweets should be marked "personal economic impact."

Category: Partisanship

Includes: Tweets in which the user mentions a specific political party or political ideology (right vs. left) with respect to the propane gas subsidy.

Excludes: Tweets in which a political party is mentioned but does not concern the gas

subsidy. These should be marked “irrelevant.”

Category: Distrust of Institutions

Includes: Tweets in which the user expresses a lack of trust in institutions to carry out the subsidy. Institutions might include the government, the propane distributors, or the businesses who sell propane.

Excludes: Tweets in which a particular political party or politician is mentioned. These should be marked “Partisanship”

Category: Personal Economic Impact

Includes: Tweets in which the user mentions how the propane gas subsidy will impact their household or their livelihood directly.

Excludes: Tweets which may fall under any other category.

Category: Other

Includes: Tweets which concern the propane gas subsidy, but do not fall under any of the other categories.

Excludes: Tweets that do not concern the propane gas subsidy. These should be marked “irrelevant.”

Category: Irrelevant

Includes: Tweets that do not concern the propane gas subsidy.

Excludes: Any tweet that concerns the propane gas subsidy.

Instructions for Sentiment Tagging:

Strongly positive: Select this if the tweet embodies emotion that was extremely happy or excited toward the topic.

Positive: Select this if the tweet embodies emotion that was generally happy, pleased, or satisfied, but the emotion wasn’t extreme.

Neutral: Select this if the tweet does not embody much of a positive or negative emotion. This includes sentiment statements like “I guess it’s ok” and statements that do not express any sentiment like statements of fact.

Negative: Select this if the tweet embodies emotion that is perceived to be angry, disappointed, or upset with the subject of the tweet, but not to the extreme.

Strongly negative: Select this if the tweet embodies negative emotion toward the topic that can be perceived as extreme.

Instrucciones para categorizar Tweets

1. Nombre del proyecto: Subsidio al Gas Propano en El Salvador.

2. Nombre de las categorías:

- (a) Falta de Información
- (b) Opinión sesgada o parcializada
- (c) Instituciones que inspiran desconfianza
- (d) Impacto Personal Económico
- (e) Otros
- (f) Irrelevante

3. Instrucciones Generales:

Estamos evaluando la percepción que tienen los usuarios de twitter en El Salvador sobre el programa de subsidio al gas propano del gobierno. En abril de 2011, el gobierno de El Salvador implementó una reforma sustancial del subsidio al gas. Antes de la reforma, los consumidores pagaban un precio fijo subvencionado para comprar botellas de gas (\$5.10), Después de la reforma, el precio de las botellas en las tiendas aumentó a \$13.60 y, como una compensación, las familias comenzaron a recibir una transferencia de \$8.50 por mes en su factura de electricidad. Gentilmente solicitamos ponga cada tweet en la categoría a la que corresponda de acuerdo a las mismas descritas abajo. Por favor no seguir ningún enlace o menciones “@” que el tweet ofrezca para obtener más contexto.

4. Criterios de Selección:

(a) Categoría: Falta de información

Incluye: Tweets en los cuales el usuario expresa confusión sobre el subsidio de gas de propano. Por ejemplo, los usuarios no saben por qué el precio de gas de propano aumentó o como tomar ventaja del subsidio.

Excluye: Tweets en los cuales el usuario expresa incertidumbre sobre como el subsidio afectará sus vidas. Estos tweets deberían pertenecer a la categoría el “Impacto Personal Económico.”

(b) Categoría: Opinión sesgada o parcializada

Incluye: Tweets en los cuales el usuario menciona un partido político específico o la ideología política (la derecha vs. la izquierda) con respecto al subsidio de gas de propano.

Excluye: Tweets en los cuales un partido político es mencionado, pero no concierne el subsidio de gas. Estos deberían ser marcados como “Irrelevantes”.

- (c) Categoría: Instituciones que inspiran desconfianza

Incluye: Tweets en los cuales el usuario expresa una falta de confianza en instituciones para llevar a cabo el subsidio. Las instituciones podrían incluir al gobierno, a los distribuidores de propano, o a los negocios que venden propano. Excluye: Tweets en los cuales un partido político o un político en particular es mencionado. Estos deberían pertenecer a la categoría “Opinión sesgada o parcializada”

- (d) Categoría: Impacto Personal Económico Incluye: Tweets en los cuales el usuario menciona como el subsidio al gas propano afectará su hogar o su sustento directamente. Excluye: Tweets que puedan caer bajo cualquier otra categoría.

- (e) Categoría: Otros Incluye: Tweets que conciernen al subsidio al gas propano, pero no pertenece a ninguna de las otras categorías. Excluye: Tweets que no conciernen al subsidio al gas propano. Estos deberían ser marcados como “irrelevantes.”

- (f) Categoría: Irrelevante Incluye: Tweets que no conciernen al subsidio al gas propano. Excluye: Cualquier tweet que concierne al subsidio al gas propano.

Instrucciones para el Etiquetado o la Acogida al Producto:

Enfáticamente positivo: Seleccione esto si el tweet expresa una emoción extremadamente feliz o entusiasmada sobre el tema.

Positivo: Seleccione esto si el tweet expresa una emoción feliz, contenta, o satisfecha en términos generales, pero la emoción no era extrema.

Neutro: Seleccione esto si el tweet no expresa una emoción verdaderamente positiva o negativa. Esto incluye declaraciones como “supongo que está bien” y las declaraciones que no expresan ningún sentimiento como las declaraciones que relatan los hechos.

Negativo: Seleccione esto si el tweet expresa una emoción que es percibida como enfado decepción, o de molestia con el tema del tweet, pero no al extremo.

Enfáticamente negativo: Seleccione esto si el tweet expresa una emoción negativa hacia el tema que pueda ser percibida como extrema.