200K+ Crowdsourced Political Arguments for a New Chilean Constitution

Constanza Fierro **Jorge Pérez** **Mauricio Quezada**
Department of Computer Science, Universidad de Chile
{cfierro,jperez,mquezada}@dcc.uchile.cl

Claudio Fuentes-Bravo
Center for Argumentation and Reasoning Studies, Universidad Diego Portales
claudio.fuentes@udp.cl

Abstract

In this paper we present the dataset of 200,000+ political arguments produced in the local phase of the 2016 Chilean constitutional process. We describe the human processing of this data by the government officials, and the manual tagging of arguments performed by members of our research group. Afterwards we focus on classification tasks that mimic the human processes, comparing linear methods with neural network architectures. The experiments show that some of the manual tasks are suitable for automatization. In particular, the best methods achieve a 90% top-5 accuracy in a multi-class classification of arguments, and 65% macro-averaged F1-score for tagging arguments according to a three-part argumentation model.

1 Introduction

The current constitution of Chile was written during Pinochet's dictatorship (Political Constitution of the Republic of Chile, 1980). Since the return to democracy in 1990, there has been an increasing pressure to make changes to this constitution. During 2016, the Chilean government finally decided to begin with a participative process to delineate what a new constitution should consider (Jordán et al., 2016). Several aspects of the Chilean process diverged from a classical way of producing a new constitution. The first phase of the process included small assemblies across the country, big group discussions at the regional level, on-line individual surveys, and so on. All the generated data was uploaded by the participants using a dedicated Web-site: http://unaconstitucionparachile.cl.

One of the most interesting parts of the process was the local participative phase in which small groups join together in a half-day meeting. During the meeting the participants had to agree on which are the most important *constitutional concepts*, writing an argument about why each of these concepts is relevant. The process produced a dataset of 200,000+ political arguments that was openly published in a raw and anonymous form (General Secretariat, Presidency of Chile, 2016).

In this paper we present the curated and tagged dataset of political arguments produced in the local phase of the 2016 Chilean constitutional process, and we analyze it to understand what type of automated reasoning is necessary to classify and tag the components of these arguments. We describe the manual processing and tagging performed by the government officials and then by members of our research group. We consider a three-part argumentation model dividing arguments into *policies* (e.g., "The state should provide free education for all"), *facts* (e.g., "Global warming will threaten food security by the middle of the 21st century"), and *values* (e.g., "The pursuit of economic prosperity is less important than the preservation of environmental quality"). This tagging included the manual parsing, normalization and classification of every single argument in the dataset, and was used as input for the official report of the 2016 process (Baranda et al., 2017).

The effort and resources spent in the manual classification and tagging of arguments was considerable, taking months of work. This motivates us to look for ways to automatize at least parts of these tasks. In particular, one of our motivations is the possibility of adding more arguments from new participant groups, but without the burden of relying on such an expensive and time consuming manual post processing.

We present several baselines on tasks that mimic the human classification and tagging pro-

Proceedings of the 4th Workshop on Argument Mining, pages 1–10

cessing. We consider two tasks. The first task is a multiclass classification problem of arguments according to the constitutional concept that they are referring to. The second task is an automatic tagging of arguments according to our three-part argumentation model. For these tasks, we compare standard methods, in particular, Logistic Regression, Random Forests and Support Vector Machines, with modern neural-network architectures tailored for natural language processing. Our baseline methods achieve a good performance thus showing that some of the manual tasks are suitable for automatization. In particular, our best methods achieve more than 90% top-5 accuracy in the multiclass classification on the first task. Regarding the second task, we obtain a performance of over 65% macro-averaged F1-score.

The data presented in this paper is one of the biggest datasets of arguments written in the Spanish language. Moreover, to the best of our knowledge, it is the only dataset of its characteristics in the Chilean Spanish dialect. We expect that this dataset plus our baselines can be useful for analyzing political arguments in Spanish beyond the specific tasks that we consider in this paper. The full dataset is available at https://github.com/uchile-nlp.

Related work

One particular work that is similar to ours presents a dataset of ideological debates in the English language and the specific task of classifying the stance (e.g. in favor or against) (Somasundaran and Wiebe, 2010). This work deals with controversial topics and the corresponding stances, but not on how relevant the topics are to propose public policies. Another similar corpus is the one regarding suggestions of the future use of a former German airport (Liebeck et al., 2016). This corpus is similar to ours in the sense of having informal arguments about public policies, but differs considerably in size (about 1% of our dataset).

A dataset of political arguments in the English language is presented with the corresponding annotation of sentiment, agreement, assertiveness, etc., obtained from an online forum (Walker et al., 2012; Abbott et al., 2016). The dataset consists of pairs question-answer about different topics. The main differences lay in the informal nature of an online forum and that the opinions are made by individuals. In our corpus, the arguments are made from collective meetings in a semi-formal setting.

In the Spanish language, a corpus consisting of 468 theses and undergrad proposals was made public in 2015 (González-López and López-López, 2015). The main difference is the formal tone of its contents and the homogeneity of the individuals that produced the texts. Gorrostieta and López-López (2016) perform classification techniques for argument mining on that dataset. Regarding the size of the data, it is roughly 10% of the dataset presented in this work. We did not find any more related datasets in the Spanish language.

2 Background of the 2016 Chilean Constitutional Process

Here we discuss the background of the constitutional process in Chile, and we describe how the data was generated. The process was divided in several steps (Jordán et al., 2016). First, citizens interested in discussing a new constitution were invited to organize themselves in small groups called *Self-convened Local Meetings* (SLMs). Every SLM was composed of 10 to 30 people that had to meet between April and June 2016. From June to August 2016 there were meetings at the municipality level and finally at the regional level, in which bigger groups discussed the output of the previous phases. The whole process was supervised by a *Citizen Council* of 15 members politically independent from the government. In January 2017, considering the output of all the previous phases, the Citizen Council produced the *Citizen Foundations for a New Constitution* (Baranda et al., 2017) in a set of documents that were given to the Chilean president of the time, Michelle Bachelet. The presidency is, at the time of this paper, preparing a bill to be sent to the Congress during late 2017. The decision about the mechanism to produce the constitution is to be decided by the 2018-2022 Congress (Jordán et al., 2016).

The 2016 phase of the process was a success in terms of the number of participating citizens, especially the SLMs phase. The government expected to have at most 3,000 SLMs, but more than 8,000 were successfully completed across the whole country with 106,412 total participants (General Secretariat, Presidency of Chile, 2017). This was 5 times the number of participants in the regional phase. In this paper we focus on the data produced by the SLMs.

SLMs and raw data

SLMs were guided by a form provided by the government (Jordán et al., 2016) which was replicated in the Web site used upload the info after the SLM. The form proposes four **topics** for discussion: Values (**V**), Rights (**R**), Duties (**D**), and Institutions (**I**). Among every topic, the participants should select seven constitutional **concepts**. For example, for the **V** topic they can select concepts such as "Dignity", "Gender Equality" or "Justice", and for the **R** topic they can select concepts such as "Privacy", "Non discrimination", "Right to education" and so on. The form included example concepts for every topic (37 example concepts for **V**, 44 for **R**, 12 for **D**, and 21 for **I**), but the participants can also include their own concepts. In that case they should select the concept "Other" and then write the new concept and its argument. We call them *open concepts*. For every selected constitutional concept, the participants should write an **argument** (in natural language) explaining why this concept should be considered in an eventual future constitution. Table 1 shows examples of (real) **concept**-**argument** pairs for the **R** topic.

The complete raw dataset of SLMs is composed of 205,357 arguments organized in the four mentioned topics. The total number of words (concept plus arguments) in the complete corpus is 4,653,518, which gives an average of 22.6 words per argument ($\sigma = 13$). Most of the arguments were given for concepts proposed in the SLM form, and only 10.7% were given for open concepts. Nevertheless, since open concepts are freely written by the participants, the data contains an important number of (syntactically) different constitutional concepts (11,568). Table 2 shows a summary of these numbers organized by topics. Table 3 shows the portion of arguments from the total that were given for open concepts.

It should be noticed that SLMs participants were diverse in terms of age, educational level, professional background and so on. As expected, arguments have different styles, and some of them partially lack proper grammatical constructions or correct punctuation (Table 1).

3 Tagging and processing of the corpus

3.1 Concept classification

The initial analysis by the Government officials was a frequency count of constitutional concepts. The main difficulty was that although concepts may be syntactically different, they can represent the same abstract idea (e.g. "Gender equality" and "Equality of rights for men and women"). To cope with this problem, they first tried to classify all the open concepts as one of the 114 initial constitutional concepts proposed in the SLM form. They proceed by classifying inside every topic (e.g., an open concept in topic **V** should only be classified as one of the 37 original **V** concepts). In the classification, every open concept-argument pair was independently classified by two annotators, and discrepancies were solved by the inclusion of a third one. In the published data there are 22,015 arguments with an open concept. Of them, 10,263 were successfully classified as one of the 114 initial concepts, 3,001 were considered as unclassifiable and the remaining 8,751 were clustered to form 47 new constitutional concepts with few arguments each (213 in average).

A total of 18 annotators plus 4 managers participated in the classification; they had a professional background in sociology and completed one day of training. The annotation achieved 87% total agreement and a Cohen's Kappa score of 0.85 (Cortés, 2017). The process was performed by the United Nations Development Program and the Department of Psychology of one of the main universities in the country, and is briefly described in a report prepared by the Constitutional Systematization Committee (2017). The technical details reported here were provided via personal communication (Cortés, 2017).

3.2 Argumentation model and tagging

The model used for the manual analysis of the arguments of the corpus is an adaptation of the criteria of Informal Logic for the detection and analysis of arguments (Hitchcock, 2007), the theory of collective intentionality of Searle (2014) and Tuomela (2013), and the classification of controversial topics in the American academic debate of Snider and Schnurer (2002) and Branham (1991).

Theoretical background

Hitchcock's (2007) account of argument subsumes the possibility that premises and conclusions may be speech acts of different sorts. In particular, it allows a premise to be any communication act which asserts a proposition (such as suggesting, hypothesizing, insulting and boasting), and allows a conclusion to be a request for information, a request

concept	argument	(argument mode)
Equality before the law	There should exist equality before the law for regular people businessmen politicians businessmen and politicians relatives without privileges or benefits.	(policy)
Right to a fair salary	The worst of all inequalities is the salary of the congressmen, Ministers and others with respect to the salary of (CLP)$250,000 of the workers.	(value)
Right to education	It is a fundamental social right, the basis of equality that democratizes access to the construction of thought to develop the potential of the participative citizen.	(fact)

Table 1: Examples of constitutional concepts and arguments for the topic "Rights" produced during a SLM. (Arguments were translated from Spanish trying to preserve their original draft and punctuation.) The final column is the annotation according to the argumentation model.

Topic	words	arguments	open conc.	gov. conc.
V	1,202,629	53,780	1,876	37
R	1,253,300	53,060	3,712	44
D	1,156,644	48,758	2,860	12
I	1,040,945	49,759	3,120	21
Total	4,653,518	205,357	11,568	114

Table 2: Statistics for SLMs raw data with open and government concepts.

	V	**R**	**D**	**I**	Total
#	4,625	6,173	4,596	6,621	22,015
%	8.6%	11.6%	9.4%	13.3%	10.7%

Table 3: Arguments with open concepts.

to do something, a commissive, an expressive, or a declarative. This broadening of the notion of argument is essential to recognize and distinguish the diverse roles that argument and inference play in real-life contexts.

From a pragmatic point of view, we can determine, based on the ideas of Searle (2014) and Tuomela (2013), that the opinions formulated on the arguments that we analyzed reflect different purposes. If the expression analyzed is identified as an assertive speech act (a report of facts, rules or states), then we can reconstruct such reasoning as a factual one ("The production of genetically modified foods is a political problem for Latin America"). If the expression can be identified as a directive speech act, then it is a reasoning of politics ("Chile must be incorporated into the OECD").

Factual and political reasonings follow a propositional pattern that allows one to reconstruct a partial or fragmented enunciative structure. This happened frequently in the arguments of SLMs. Once the arguments were reconstructed, we used the classification proposed by Snider and Schnurer (2002) and Branham (1991) for controversial topics. This classification gathers 150 years of categorization of statements in the tradition of academic debating in the United States which made it a fairly robust strategy for categorizing natural language. With this strategy we classified all the arguments in the corpus using three kinds of propositions: facts, values and policies.

Facts, values and policies

Factual propositions speak of what it "is", "was" or "will be". They are composed of a subject ("the house", "capitalism"), the verbal formula "is", which entails the idea of identity or subduction, and finally, a set of conditions.

Value propositions represent evaluation statements that use abstract binary concepts (such as beautiful vs. ugly, relevant vs. irrelevant, equity vs. inequity), regarding people, places, things or events (Snider and Schnurer, 2002, pp. 88–89). The value propositions are composed, in similar terms of the factual thesis, of a subject (or study case), a verbal form "is", and a set of conditions. Value propositions differ from facts in the presence of a qualificative, consisting of an adjective whose semantic function is to evaluate either positively or negatively. Pragmatical or instrumental qualifications such as "efficient", "useful", and "convenient", are usually considered as value propositions. Nevertheless, it is preferable to treat them as facts if their value depends exclusively on factual situations, e.g., if we say "S is efficient" meaning that it spends the lesser possible resources.

Finally, policies, or political propositions, are formulated according to a question of the type "What should be done?". The political propositions are composed of a deontological modal indicator "it should" (or an equivalent). In general,

Argument mode	Amount	Percentage
policy	135,489	66.0%
fact	37,397	18.2%
value	11,912	5.8%
undefined	11,238	5.5%
blank	9,321	4.5%

Table 4: Distribution of argument modes resulting from human annotators.

the object will be composed of a verbal form that aims towards an illocutive intention (e.g. allow, prohibit, approve, made), and a subject or object. Political debates can be referred as potential actions of the local or national government ("Chile should have free education at all levels").

Every fact, policy and value proposition was normalized to follow the structural pattern *subject–verb–direct object*, having in some cases a complement that comprises indirect objects or other kinds of syntactic complements. This choice of reconstruction allows us to go deeper in a morphosyntactical analysis without forcing the more elemental claims to have a complex construction. With all these ingredients, we consider a tagging scheme in which every argument is normalized identifying the following essential parts: (1) subject, (2) verbal syntagm, (3) nominal syntagm, (4) prepositional syntagm, and (5) argumentative mode (either fact, value or policy). As an example, consider the following sentence in Spanish: "Se debe aceptar el matrimonio homosexual en Chile". Its verbal syntagm is "Se debe aceptar" ("it should be allowed"), the nominal syntagm is "el matrimonio homosexual" ("gay marriage") and the propositional syntagm is "en Chile" ("in Chile"). In this case the subject is implicit, which is a typical form to state policies in Spanish (starting with the form "Se debe"). Given this component identification it is clear that the sentence has a policy mode.

Normalization tagging process

We considered candidate annotators from local undergrad students and professionals in sociology, psychology, political science, linguistics, etc. They were given a 90-minute orientation and then tested in a normalization and tagging task of 50 arguments. Those candidates that achieved at least 80% accuracy (compared to a gold standard of examples previously annotated and corrected by the team) were invited to continue as annotators on an on-site work alongside with research assistants from our group. Every annotator was closely followed by one manager during the first five working days. The manager corrected the annotations along with the annotator and, if needed, re-trained him or her. After those first days, the annotators that achieved a proper standard in the evaluation of the team, processed arguments independently of the manager, but every annotation was inspected for correctness by the manager. Those annotations considered as incorrect were sent back to the pool of unprocessed arguments, to be processed again by a different annotator. More than 120 annotators participated in the process, receiving 0.15 USD per correctly annotated argument. After completing the process, we performed a validation step, by sampling a random set of annotations, which were corrected again by the team. The error estimated by using that procedure was less than 15%. It should be noticed that the quality control procedure used here was a compromise between academic methodologies and the requirements made by the contracting party, which stressed the short time available to complete the analysis of the 200,000+ cases. Table 4 shows the number of arguments tagged in every mode of our argumentation model. As the numbers show, most of the arguments (66%) were tagged as policies.

4 Classification tasks

We consider three main tasks. Task A and Task B are associated to the classification of concepts (Section 3.1) and Task C to the tagging process of arguments according to our argumentation model.

One of our main motivations is to mimic the classification of open concepts described in Section 3.1. Towards this goal, we first define a task that tries to predict to which concept a given argument is referring to. Formally, let C_G be the set of 114 constitutional concepts provided by the Government in the SLMs. Recall that SLMs were divided in four topics, thus C_G can be partitioned in four disjoint sets of concepts, $C_G^{\mathbf{V}}$, $C_G^{\mathbf{R}}$, $C_G^{\mathbf{D}}$, and $C_G^{\mathbf{I}}$, one for each topic. Let D_G be the set of concept-argument pairs (c, a) such that $c \in C_G$ (that is, concept-argument pairs that were explicitly written as one of the 114 government concepts by the SLM participants), and let A_G be the set of all arguments associated to concepts in C_G. Similarly as for C_G, we can partition D_G and A_G into sets D_G^T and A_G^T with $T \in \{\mathbf{V}, \mathbf{R}, \mathbf{D}, \mathbf{I}\}$. We have all the necessary notation to formalize our first task.

Task A. *Fix a topic $T \in \{\mathbf{V}, \mathbf{R}, \mathbf{D}, \mathbf{I}\}$. Given an argument $a^\star \in A_G^T$, predict the concept $c \in C_G^T$ such that $(c, a^\star) \in D_G^T$.*

Notice that Task A is essentially defining four independent classification problems, one for each different topic. We show in the next sections that finding models for Task A proves to be useful in solving a classification problem for open concepts that we next formalize.

Let C_O be the set of open concepts, that is, the set of concepts $c^\star$ such that $c^\star \notin C_G$. Similarly as for the previous task, one can define D_O (the set of pairs with open concepts) and A_O (the set of arguments for open concepts) and their partitions by topics C_O^T, D_O^T and A_O^T with $T \in \{\mathbf{V}, \mathbf{R}, \mathbf{D}, \mathbf{I}\}$.

Task B. *Fix a topic $T \in \{\mathbf{V}, \mathbf{R}, \mathbf{D}, \mathbf{I}\}$. Given a pair $(c^\star, a^\star) \in D_O^T$, determine a concept $c \in C_G^T$ to which $(c^\star, a^\star)$ is most probably referring to.*

Our final task is a prediction of the argumentation mode and is formalized as follows.

Task C. *Given an argument $a^\star \in A_G \cup A_O$, predict the most suitable tag for $a^\star$ according to our argumentation model (policy, fact, value).*

Notice that in our final task we do not make any distinction by topic or whether the argument was given for an open concept or not.

5 Methods

We consider two types of methods to compute (non-trivial) baselines for the above-mentioned tasks: standard linear classifiers, and simple neural-network based methods tailored for natural language processing. We begin by describing the standard classifiers and the features that we consider.

5.1 Standard classifiers

We consider three baseline standard classifiers: Logistic Regression (LR), Random Forests (RF) (Breiman, 2001), and Support Vector Machines (SVM) (Cortes and Vapnik, 1995). The setting involves several combinations of feature sets and normalizations. Feature sets comprise (1) the extraction of *unigrams*, *bigrams*, and unigrams plus bigrams (denoted as *ngram*), and (2) raw tokens (denoted as *raw*) and Part of Speech tagged tokens (denoted as *POS*). Normalizations comprises (1) raw term counts (denoted as *count*), (2) term counts normalized by term frequency (denoted as *tf*), and (3) normalized by term frequency and inverse document frequency (denoted as *tf-idf*). For all combinations we use the lemma of a token instead of the original token, and stopwords are removed. This ends up in 18 combinations for every one of the three classifiers, resulting in 54 baselines.

5.2 Neural networks classifiers

We consider two methods, fastText (Joulin et al., 2016) and Deep Averaging Networks (Iyyer et al., 2015), that have been proposed as simple yet efficient baselines for text classification. We also consider the use of *word embeddings*.

FastText Joulin et al. (2016) propose a simple two-layer architecture for text classification called fastText. The input for the classifier is a text represented as a bag of words. In the first layer the classifier transforms those words into real-valued vectors that are averaged to produce a hidden-variable vector representation of the text. This representation is fed to a softmax output layer. The model is then trained with stochastic gradient descent. Joulin et al. (2016) show that fastText outperforms competing methods by one order of magnitude in training time, having superior accuracy in a tag prediction task over 300,000+ tags.

Deep averaging networks Iyyer et al. (2015) propose what can be considered as a generalization of the above method; after the first hidden averaging layer, the average is passed trough one or more feed forward layers. The final output layer is also a softmax layer. As in the case of fastText, the authors show a significant performance gain in training time while having a high accuracy in a sentiment analysis task. The resulting family of models is called Deep Averaging Networks (DAN) (Iyyer et al., 2015).

Word embeddings Word embeddings are vector representations for words learned from the contexts in which words appear in large corpora of text (and idea that can be traced back to the distributional semantics hypothesis in linguistics (Harris, 1954)). There are several methods to learn word representations from unlabelled data (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016), and usually, training over more data produces vectors with better semantic characteristics. Word embeddings can be used to check the similarity of two texts by simply averaging the vector representation of the words

of each text and then computing a vector similarity measure (such as the cosine similarity).

It has been shown that pre-trained vectors can help when using neural networks for text classification (Kim, 2014). In our experiments we also consider versions of fastText and DAN with pre-trained word-embedding vectors in the first layer.

5.3 Implementation details

The standard classifiers are implemented with Scikit-learn (Pedregosa et al., 2011). For tokenization, lemmatization and Part of Speech tagging, we use FreeLing (Carreras et al., 2004), which supports the Spanish language. For fastText we use the C++ implementation provided by Grave et al. (2016). We implemented DANs using the Keras framework (Chollet et al., 2015). We use pre-trained word embeddings computed from the Spanish Wikipedia by using the method proposed by Bojanowsky et al. (2016).

6 Experiments and results

In the following sections we describe the experiment settings and results for the tasks defined in Section 4.

For Task A and Task B, we compare our methods using accuracy and top-5 accuracy (percentage of cases in which the correct class belongs to the top-5 predictions). Accuracy is useful in our case, given that there are several classes (12 to 44) and the biggest is around 10% of the total instances. The use of top-5 accuracy allows us to evaluate our models in the scenario of helping humans to quickly determine the class an argument is referring to. For Task C we use macro-averaged precision, recall and F1-score as metrics due to the class imbalance.

6.1 Task A

For Task A we consider pairs (c, a) with $c \in C_G$ and such that a was not marked as *blank* in the manual classification process (Section 3.2). This gives us a total of 169,242 pairs. We divide this set into four sets, one for each topic $(\mathbf{V}, \mathbf{R}, \mathbf{D}, \mathbf{I})$, that we use as data for the four instantiations of Task A. In every case we randomly divide the data into 80% train, 10% dev and 10% test sets with a stratified sampling. For the standard models, we use 90% for training (train plus dev), as we do not use the dev set to tune model parameters.

Table 5 reports our results for Task A for each topic. The first row shows a majority baseline as comparison and the last column reports the average over the four topics as an overview of the performance. For the standard classifiers we report only the best-performing configuration for each strategy. All reported results are over the test set.

In almost all topics, fastText with pre-trained word embeddings is the best performing model for (top-1) accuracy, with Logistic Regression being behind by a little margin. For the case of fastText, the use of pre-trained vectors and bigrams gives an average of 2% in gain over plain fastText. For top-5 accuracy, fastText is again the best performing model, however, in contrast to the previous case, the use of bigrams can harm the performance. The best methods achieve over 90% top-5 accuracy for all topics.

In the case of standard models, both Logistic Regression and SVM have competitive performance compared to more complex models. We found that the use of bigrams actually hurts the performance of the linear models, although using them in conjunction with unigrams improve the accuracy in some cases. We believe that this is due to the typical sparsity that the use of bigrams introduce in the models. Using only unigrams and tf-idf gives the best performance at top-5 accuracy in the Logistic Regression.

6.2 Task B

For Task B we consider as test set the 10,263 pairs (c, a) with open concepts that were manually classified as one of the 114 concepts in C_G (as described in Section 3.1). We perform experiments considering as input the string of the concept and also the concatenation of the concept and argument strings, and we feed this input to the same models computed for Task A. That is, we do not re-train our models, instead we use the same trained models for the previous task to solve this new task with a different test set. We consider two additional simple baselines that only compares the strings of the concepts:

- **Edit-distance**: given $(c^\star, a^\star) \in D_O^T$ we compute the edit distance between $c^\star$ and all the elements $c \in C_G^T$, and rank the results.
- **Word-embedding**: given an input $(c^\star, a^\star) \in D_O^T$ we compute the cosine distance between the average word-embedding of (the words in) $c^\star$ and the average word-embeddings of every $c \in C_G^T$, and rank the results.

	Values (37 classes)		**R**ights (44 classes)		**D**uties (12 classes)		**I**nstitutions (21 classes)		Average	
	Acc	Top-5	Acc	Top-5	Acc	Top-5	Acc	Top-5	Acc	Top-5
Majority	8.5	39.5	12.2	40.7	14.1	60.1	12.6	43.8	11.8	46.0
RF+unigram+raw (tf)	56.1	79.5	62.3	83.0	68.1	90.5	61.7	84.4	62.0	84.3
LR+unigram+raw (tf-idf)	66.3	**91.0**	70.3	91.7	75.5	96.2	69.6	91.6	70.4	92.6
LR+ngram+raw (count)	67.5	90.8	70.7	91.6	76.6	96.1	**70.2**	91.5	71.3	92.5
SVM+ngram+POS (tf-idf)	67.9	-	70.7	-	76.2	-	69.8	-	71.2	-
fastText	65.9	89.4	68.6	90.6	75.1	95.8	68.4	91.1	69.5	91.7
fastText+bigram	64.9	88.2	67.1	89.1	75.9	95.4	68.5	91.0	69.1	90.9
fastText+pre	67.1	90.7	70.8	**92.3**	75.7	**96.4**	69.3	92.5	70.7	**93.0**
fastText+pre+bigram	**68.0**	90.2	**71.1**	91.8	**76.9**	95.8	69.4	**92.7**	**71.4**	92.6
DAN+pre	64.5	89.4	68.2	91.7	73.6	96.2	66.4	91.8	68.2	92.3

Table 5: (Task A) Classification results. Top-1 and top-5 accuracy is reported for each baseline and topic.

	Values (37 classes)		**R**ights (44 classes)		**D**uties (12 classes)		**I**nstitutions (21 classes)		Average	
	Acc	Top-5	Acc	Top-5	Acc	Top-5	Acc	Top-5	Acc	Top-5
Majority	03.0	22.2	02.8	20.2	10.1	48.6	07.1	25.4	05.8	29.1
Edit-distance (c)	41.2	60.6	30.9	46.7	41.6	64.9	22.7	38.6	34.1	52.7
Word-embeddings (c)	60.2	86.3	58.8	79.1	60.4	80.8	45.5	86.1	56.2	83.1
RF+unigram+POS (tf) (c, a)	50.5	76.8	63.1	84.1	69.9	94.7	46.7	68.1	57.5	80.9
LR+ngram+POS (count) (c, a)	60.4	89.9	71.9	**92.7**	77.6	95.0	56.1	84.8	66.5	**90.6**
SVM+ngram+POS (tf-idf) (c, a)	61.4	-	71.9	-	78.4	-	55.3	-	66.7	-
fastText+pre (c)	61.4	89.0	**73.3**	91.8	79.0	95.3	55.3	86.4	67.2	**90.6**
fastText+pre (c, a)	60.7	89.9	70.6	92.3	75.5	95.5	52.7	83.2	64.9	90.2
fastText+pre+bigram (c)	**62.9**	87.4	72.4	91.0	**79.2**	94.7	**60.2**	**86.7**	**68.7**	90.0
fastText+pre+bigram (c, a)	60.9	89.9	71.1	92.1	76.3	95.4	53.8	81.2	65.5	89.6
DAN+pre (c)	61.6	87.2	70.4	92.6	77.9	**96.3**	55.6	82.3	66.4	89.6
DAN+pre (c, a)	60.4	**91.1**	69.6	92.4	75.0	95.2	51.4	80.8	64.1	89.9

Table 6: (Task B) Classification results. Top-1 and top-5 accuracy are reported for each baseline and topic. After each baseline, (c) indicates that only the concept is used as a test instance, and (c, a) indicates that both the concept and the argument are used.

We report accuracy and top-5 accuracy per topic in Table 6. Regarding (top-1) accuracy, fastText and DAN perform best when only the string of the concept is given as input, a trend that changes for top-5 accuracy in which having the concept plus the argument actually helps to make better predictions (except for topic **I**). In our experiments we observed that the gap in top-k accuracy between using and non-using the argument consistently increases as k increases. On the other hand, we found that the use of the concept plus the argument improves the performance of the linear models. As a final comment, the estimated human accuracy for this task was 87% (Cortés, 2017), and our best method achieves 68.7% in average. This gives an important space for improvement.

	Prec.	Recall	F1
Majority	24.4	33.3	28.2
RF+unigram+POS (tf)	64.1	50.0	53.0
LR+ngram+POS (count)	65.1	54.7	57.9
SVM+ngram+POS (tf-idf)	66.5	55.1	58.3
fastText+pre	69.6	59.7	63.3
fastText+bigram	68.9	62.0	64.8
fastText+pre+bigram	**69.9**	**62.4**	**65.4**
DAN+pre	67.1	59.0	62.1

Table 7: (Task C) Classification results. Values correspond to macro-averaged metrics.

6.3 Task C

For this task we consider the set of all arguments that have been tagged as either policy, fact, or value by the process described in Section 3.2. That is, we do not consider blank or undefined arguments. Thus the dataset is composed of 184,798

arguments from which 73.3% are policies, 20.2% facts and 6.5% values. We split our set into 80% train, 10% dev and 10% test sets. Since our dataset contains clearly unbalanced classes we consider macro-averaged precision, recall and F1-score as our performance metric. Results on the test set are reported in Table 7. FastText with pre-trained vectors and bigrams is the best performing model with 65.4% F1. This model achieves a performance of 81.1% accuracy which is close to the estimated human accuracy of the process (85%).

7 Conclusions

In this paper we have presented the corpus of political arguments produced in the 2016 Chilean Constitutional Process together with several baselines for classification tasks. This corpus is one of the largest tagged datasets of arguments in the Chilean Spanish language.

Our defined tasks and baselines can be useful in applications beyond the ones we analyzed in this paper. In particular, the classification of arguments into concepts could be useful to identify political subject matters in open text in the Spanish language.

Chile is going through an important political discussion. Our natural next step is to use our tools to help in the analysis of new opinions, emphasize the transparency, and foster the repeatability of the process to draw new conclusions.

Acknowledgements

We thank the anonymous reviewers, Camilo Garrido, and Miguel Campusano for their helpful comments. We also thank Pamela Figueroa Rubio from the Ministry General Secretariat of the Presidency of Chile and Rodrigo Marquez from the United Nations Development Program for their help in the analysis process. Fierro, Pérez and Quezada are supported by the Millennium Nucleus Center for Semantic Web Research, Grant NC120004. Quezada is also supported by CONICYT under grant PCHA/Doctorado Nacional 2015/21151445.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *LREC*.

Benito Baranda, Jean Beausejour, Roberto Fantuzzi, Arturo Fermandois, Francisco Fernandez, Patricio Fernandez, Gaston Gomez, Hernan Larrain, Hector Mery, Salvador Millaleo, Ruth Olate, Juanita Parra, Lucas Sierra, Francisco Soto, and Patricio Zapata. 2017. Final Report on the Participative Phase of the Chilean Constituent Process (in Spanish). Ministry General Secretariat of the Presidency of Chile https://unaconstitucionparachile.cl/Informe-Final-CCO-16-de-enero-de-2017.pdf.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Robert James Branham. 1991. *Debate and critical analysis: The harmony of conflict*. Routledge.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Xavier Carreras, Isaac Chao, Llus Padr, and Muntsa Padr. 2004. Freeling: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

Constitutional Systematization Committee. 2017. Self-convened Local Meetings: Quantitative Results (in Spanish). Ministry General Secretariat of the Presidency of Chile http://www.sistematizacionconstitucional.cl/app/themes/cs/dist/docs/ela_frecuencias.pdf.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Flavio Cortés. 2017. *Personal Communication*.

General Secretariat, Presidency of Chile. 2016. *Proceso Constituyente Abierto a la Ciudadanía* [Dataset]. http://datos.gob.cl/dataset/proceso-constituyente-abierto-a-la-ciudadania.

General Secretariat, Presidency of Chile. 2017. Quantitative Summary of the 2016 Chilean Constituent Process, Participative Phase (in Spanish). Ministry General Secretariat of the Presidency of Chile https://unaconstitucionparachile.cl/sintesis_de_resultados_etapa_participativa.pdf.

Samuel González-López and Aurelio López-López. 2015. Colección de tesis y propuesta de investigación en TICs: un recurso para su análisis y estudio. In *XIII Congreso Nacional de Investigación Educativa*. page 15.

Jesús Miguel García Gorrostieta and Aurelio López-López. 2016. *Argumentation Identification for Academic Support in Undergraduate Writings*, Springer International Publishing, Cham, pages 98–109.

Edouard Grave, Piotr Bojanowski, Armand Joulin, et al. 2016. fastText. https://github.com/facebookresearch/fastText.

Zellig Harris. 1954. Distributional structure. *Word* 23:146–162.

David Hitchcock. 2007. Informal logic and the concept of argument. In *Philosophy of Logic*. volume 5, Handbook of the Philosophy of Science, pages 101–129.

Mohit Iyyer, Varun Manjunatha, Jordan L. Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. pages 1681–1691.

Tomás Jordán, Pamela Figueroa, Rodrigo Araya, and Carolina Gómez. 2016. Constituent Process open to Citizenship (in Spanish). Ministry General Secretariat of the Presidency of Chile https://unaconstitucionparachile.cl/guia_metodologica_proceso_constituyente_abierto_a_la_ciudadania.pdf.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* .

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1746–1751.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*. pages 3111–3119.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vicent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Èdouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1532–1543.

Political Constitution of the Republic of Chile. 1980. *Decreto 1150: Texto de la Constitución Política de la República de Chile* (in Spanish). National Library of Chile.

John R. Searle. 2014. *Creando el mundo social: la estructura de la civilización humana*. Grupo Planeta Spain.

Alfred Snider and Maxwell Schnurer. 2002. *Many sides: Debate across the curriculum*. International Debate Education Association: New York.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing Stances in Ideological On-line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Stroudsburg, PA, USA, CAAGET '10, pages 116–124. http://dl.acm.org/citation.cfm?id=1860631.1860645.

Raimo Tuomela. 2013. *Social ontology: Collective intentionality and group agents*. Oxford University Press.

Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbot, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *LREC*. pages 812–817.