RUDN Journal of Political Science
Вестник РУДН. Серия: ПОЛИТОЛОГИЯ
2020 Vol. 22 No. 3 517–532
http://journals.rudn.ru/political-science

DOI: 10.22363/2313-1438-2020-22-3-517-532

Research article / Научная статья

Natural Language Processing for the Analysis of the Political Characterisation of Migration in the Croatian Political Discourse

Gabriele De Luca, Marko Beck

Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

Abstract. This paper tackles the issue of analyst bias in performance of comparative political analyses on political discourse, by leveraging data and machine-learning over human prior knowledge. The case studied is characterization of the issue of migration in the Croatian political discourse, which was chosen arbitrarily. We developed a machine-learning system that identifies most prominent features in the Croatian political discourse, with regards to migration and were interested solo in comparative political analysis in political science. This system does not rely on human judgement on the part of the researchers, and can be thus considered to be "objective", short of possible sampling or selection bias. It is replicable. If provided, the same dataset and algorithm used, same conclusions should be reached by any scientist. This result was achieved by creating a text corpus from news items and press releases extracted from the websites of Croatian political parties currently represented in the Parliament. Available and collected data consist of public announcements mainly from IDS (Istarski Demokratski Sabor / Istrian Democratic Assambly), SDSS (Samostalna Demokratska Srpska Stranka / Independed Democratic Serb Party) and HSLS (Hrvatska Socijalno Liberalna Stranka / Croatian Social Liberal Party). Data analyzed suggests three dominant phrases of the research process. All political parties had similar political stand towards pointed out issues. Three most significant phrases were determined. First phrase is related to words "Demography" and "Reduction" and finding suggest that most analyzed articles relates towards migration of Croatian citizens in connection to economic hardships of some kind. Phrase two is related to words "Border" and "Croatia-Serbia" which strongly indicates relation to migration and is related towards inter-Balkan migration, mostly connected with consequences of the Croatian War of Independence from 1990's, and is of most interest to SDSS, a Serb minority party in Croatia. Phrase three is related towards Marrakesh Agreement (Global Compact for Safe, Orderly and Regular Migration), where most of analyzed data shows that parties have a constructive but ambivalent stance towards migration from the third countries. Research conducted on available data, shows that wide spread international migration is not in the focus of most Croatian political parties, while topics and interest for inter-Balkan and Croatian economic/political migration dominates Croatian political spectre

Keywords: political discourse, public information campaign, machine learning, information retrieval, natural language processing, migration

Article history: Submitted on 01.04.2020. Accepted on 10.05.2020.

For citation: Gabriele De Luca, Marko Beck. Natural Language Processing for the Analysis of the Political Characterisation of Migration in the Croatian Political Discourse. *RUDN Journal of Political Science*. 2020; 22 (3): 517–532. DOI: 10.22363/2313-1438-2020-22-3-517-532

Обработка естественного языка для анализа политического определения миграции в хорватском политическом дискурсе

Де Лука Г., Бек М.

Российский университет дружбы народов
Российская Федерация, 117198, Москва, ул. Миклухо-Маклая, 6

Аннотация. Статья посвящена решению проблемы предвзятости аналитиков при проведении сравнительного анализа политического дискурса. Предлагаемое решение строится на анализе данных и использовании машинного обучения для обработки естественного языка. Кейс, который мы изучаем в связи с этой проблемой, относится к определению проблемы миграции в хорватском политическом дискурсе. Была разработана система машинного обучения, которая выявляет наиболее характерные черты хорватского политического дискурса в отношении миграции: эта система свободна от исследовательской субъективности. Исследование воспроизводимо, и при условии, что используется тот же набор данных и алгоритм, любой ученый должен прийти к тем же выводам. Этот результат был достигнут на основе сбора корпус-текстов из новостных материалов и пресс-релизов с веб-сайтов хорватских политических партий, представленных в парламенте, а также группу алгоритмов классификации машинного обучения для матриц Bag-of-Words, вычисленных из корпуса. Мы определили наиболее точную модель, классификатор дерева решений, которая была выбрана для дальнейшего анализа из-за ее точности и интерпретируемости. Нами также проанализированы правила принятия решений, определенные этим классификатором, которые затем были интерпретированы людьми, чтобы определить политические особенности текста, которые лучше всего предсказывают связь этого текста с темой миграции. В итоге подробно раскрыты три правила, идентифицированные с этой процедурой, которые мы считаем особенно интересными.

Ключевые слова: политический дискурс, кампания общественной информации, машинное обучение, поиск информации, обработка естественного языка, миграция

История статьи: Поступила в редакцию 01.04.2020. Принята к публикации 10.05.2020.

Для цитирования: *De Luca G., Beck M.* Natural Language Processing for the Analysis of the Political Characterisation of Migration in the Croatian Political Discourse // Вестник Российского университета дружбы народов. Серия: Политология. 2020. Т. 22. № 3. С. 517–532. DOI: 10.22363/2313-1438-2020-22-3-517-532

Introduction and task definition

We aimed to develop a system for comparative analysis of an issue of migration, as well as of the way in which it is characterised in public information campaign of Croatian political parties. The underlying objective is to test whether it is possible to conduct comparative assessments of the party system of any given country, with regards to an arbitrarily selected policy issue, with minimal or no background knowledge of the political system of observed country, or of the way in which observed policy issue is treated by the local national parties.

Comparative politics is believed to be particularly affected by the problem of selection bias [1], in sense that results obtained tend to reflect more the prejudices of human analyst than the complexity of underlying political reality [2]. Machine learning can help escape the intellectual pitfall, by tackling quantitative method problems, such as analysis of political discourse [3], which have originally been treated through qualitative methods [4].

If a method found is to achieve the aforementioned task, as an outcome of a formalized procedure it could, in principle, be replicated by any interested scientist, in order to systematically produce the same predictable outcome. Assessments of this type would be devoid of human bias which tends to characterise comparative assessments nowadays. The analysis of political discourse is, unfortunately, largely based on poorly defined concepts. Some scholars suggest that the very notion of analysis of political discourse is ambiguous, and its conclusions rather subjective and non-formalised [5].

We follow data-driven approach taken from sector of machine-learning, specifically the branch of natural language processing [6]. Approach has been applied to a case arbitrarily chosen, and specifically the characterisation of migration in the political discourse of Croatian parties. As suggested by previous literature [7], up to sometimes after 2007, the Croatian political parties did not systematically use their internet pages as tools for public information campaigns. The situation, however, changed from then, and now there is enough data to be used as input data for the procedure developed.

Political attitudes in Croatia towards migration

Migration is a hot-topic in Croatia, because of the country's geographical position on the "Balkan migration route", which made it one of critical spots during 2015 European migration crisis [8]. Earlier, specifically after 1945, the waves of migration through Croatia were characterised by political reasons [9], as political dissidents decided to flee the country in order to avoid punishment by the political leadership [10]. More recent wave of migration, which took place in the '90s, can also be identified as an emergent consequence of the Serbo-Croatian war [11]. Contemporary Croatia is though primarily defined not as a source of emigration for local population, but as a country of transit for migration flows directed at Europe [12]. Digital media has played important role in shaping public's attitudes towards a phenomenon which only partially was observable in day to day life. Images retrieved managed to successfully enter the political construction of the world as seen by Croatian population [13]. Consequentially, features of political world seen by Croatian population can be effectively studied by studying messages on the topic of migration and political discourse transmitted over digital channels [14].

Some *a priori* predictions on the content of these features can be made, on the basis of theoretical understanding of specialised literature on subject. Those predictions can be used to test validity of model we will further develop. The largest Croatian political party, the HDZ (Hrvatska demokratska zajednica / Croatian

Democratic Union), has historically been in favour of the idea that historical diaspora should constitute an integrated component of the political system [15]. Theory would thus suggest that migration can be considered as systemic component of Croatian politics, insofar as it promotes nationalistic tendencies of the population [16]. Discussion of immigration to Croatia, as opposed to emigration from it, has however entered the political discourse only recently, starting from migration crisis of 2015 [17]. Nationalistic parties tended to be against it, while the idea that immigration is systemic has been promoted by the leftist political parties [18]. Croatian political system therefore seems to respect well-known division between conservatism of the right-wing parties, which are generally against immigration, and liberalism of the left-wing parties, who are generally in favour [19].

Within the context of theoretical predictions regarding analysis of Croatian political discourses on migration, we therefore expect the following:

1) Political discourses before 2015 should focus primarily on the subject of the Croatian War for Independence.

2) Political discourses after 2015 should primarily focus on immigration from outside of Europe.

3) Political discourses after 2015 should show a split in the attitude towards migration, with right-wing political parties being generally against it, and left-wing political parties being generally in favour of it.

The model is to be set forth and develop in order to test collected and retrieved political texts against these theoretical expectations.

Natural language processing for political analysis

Large collection of texts, called *corpus*, had to be collected in order to perform data mining [20]. It was determined that as many news and press releases from websites of all Croatian political parties as possible, would be suitable source of data needed. All of 20 political parties currently seated in the National Assembly in Zagreb, as of December 2018, has been acknowledged as a relevant political party. After manually inspecting all 20 parties websites we have concluded 14 websites was suitable for automatic information retrieval and extraction, so 14 individual crawlers was build with purpose to retrieve and extract all suitable texts. Texts were then parsed to extract their features of interest: date, title, and main body of the article.

In this manner, a dataset comprising of 9185 texts has been created. Texts were then preprocessed by removing stopwords and stemming individual words, in order to decrease dimensionality of the corpus, whilst minimising the loss of meaningful content.

Chosen texts were automatically labeled on whether or not they contained keywords unequivocally related to the policy issue of migration. In our opinion, the only part of methodology requiring subjective judgement was deciding what keywords were relevant. Data labeled as relevant was in the end inspected for internal consistency.

The best performing classification algorithm is the Decision Tree. While being overfit for dataset and deprived of generalisation capability, it provided best explanatory power and allowed us to extract rules about the policy issue of migration. This is why it has been deemed acceptable, even desirable. For the purpose of this research, we were interested solo in performing comparative analysis. Partial representation in corpus is likely going to develop some selection bias in formulation of results. Some important absences among the political parties represented in our dataset can be identified: the HDZ (Hrvatska demokratska zajednica / Croatian Democratic Union), the party with majority of seats in Croatian Parliament, is not represented in the dataset due to technical reasons. Some other parties are also absent, as described in more detail later (Table 2). Due to latest, we cannot affirm full representativeness of our conclusions. They are, however, the best approximation of all available data. If and when more data becomes available, conclusions may have to be updated.

All code developed was written by us in Python, with the usage of open-source libraries such as *Requests*[1], *NLTK*[2], and *Sklearn*[3]. Additional open-source libraries were also used for some specific tasks during preprocessing, and they are cited in the body of this text accordingly to the step of procedure in which they were first employed. No pre-made or proprietary program was employed at any step.

Data collection

As in all scientific experiments, our research started with identification and collection of data. Procedure formalised without accounting for possible human bias was followed for selection of data. First step was to list all political parties (Table 1) represented in the Parliament at the time of collection.

The list of active parliamentary parties contained 20 names stated in alphabetical order.

Website of each individual party has been accessed in order to collect relevant texts. Their "News" or "Press releases" section were often most relevant for our research. Parties (Table 2) added to the list of targets for developing crawlers and parsers were those whose websites were suitable for automatic scraping, and which also have published a non-irrelevant number of news articles or press-releases[4]. Table that follows contains full indication of parties, and their websites, which were selected as fit for automatic information retrieval and extraction, and an explanation as to why the others were not included[5]. Index of each row corresponds to index used in the previous table.

[1] URL: http://docs.python-requests.org/en/master/
[2] URL: http://www.nltk.org/
[3] URL: https://scikit-learn.org/
[4] We deem relevant a text collection of at least a dozen news items.
[5] Information contained in this table is accurate as of December 2018.

Table 1

Political parties represented in the Croatian Parliament

№	Party name
1	Bandić Milan 365 – Stranka rada i solidarnosti
2	Bruna Esih – Zlatko Hasanbegović: Neovisni za Hrvatsku
3	Građansko-liberalni savez
4	HRAST – Pokret za uspješnu Hrvatsku
5	Hrvatska demokratska zajednica
6	Hrvatska demokršćanska stranka
7	Hrvatska narodna stranka – liberalni demokrati
8	Hrvatska seljačka stranka
9	Hrvatska socijalno-liberalna stranka
10	Hrvatska stranka umirovljenika
11	Hrvatski demokratski savez Slavonije i Baranje
12	Istarski demokratski sabor
13	Most nezavisnih lista
14	Narodna stranka – Reformisti
15	Nezavisna lista mladih
16	Promijenimo Hrvatsku
17	Samostalna demokratska srpska stranka
18	SNAGA – Stranka narodnog i građanskog aktivizma
19	Socijaldemokratska partija Hrvatske
20	živi zid

Source: Created by the authors on the basis of information available on the website of the Croatian Parliament. URL: http://www.sabor.hr/hr/zastupnici/parlamentarne-stranke (accessed on 21 December 2018).

Table 2

Political parties whose news items were included in the dataset

№	URL	Included	If No, why	If Yes, tag
1	http://www.365ris.hr	Yes		365
2	http://www.neovisni.hr	Yes		NZH
3	http://glas.com.hr	Yes		GLAS
4	http://www.h-rast.hr	Yes		HRAST
5	http://www.hdz.hr	No	Website uses Cloudflare and ReCaptcha	
6	http://www.demokrscanihds.hr	Yes		HDS
7	https://www.hns.hr	Yes		HNS
8	http://www.hss.hr	No	Only 6 news items are present	
9	http://www.hsls.hr	Yes		HSLS
10	http://www.hsu.hr	Yes		HSU
11	http://www.hdssb.hr	Yes		HDSSB
12	http://www.ids-ddi.com	Yes		IDS
13	https://most-nl.com	Yes		MOST
14	https://reformisti.hr	Yes		REFORM
15	http://nlm-vrgorac.com	No	Site unresponsive	
16	http://promijenimohrvatsku.hr	No	Only 10 news items available	
17	http://sdss.hr	Yes		SDSS
18	https://snaga.hr	No	Only 12 news items available	
19	http://www.sdp.hr	Yes		SDP
20	https://www.zivizid.hr	No	A minified JS function displays the news	

Source: Made by the authors on the basis of the elements of Table 1, above.

At this stage collection of raw html pages consisted of 9677 files. Our parsers then extracted following features from each of available pages: date of publication, title, and main body of the article. These features, along with the party affiliation of

each text, were used to populate the columns of dataset. As some articles comprised exclusively of images or embedded videos, texts extracted from such articles were null, and thus were dropped from the dataset. Similarly, duplicated texts were also removed. This process left us with 9185 non-null rows in dataset, corresponding to as many unique observations. At this stage the dataset looked like this (Table 3).

Table 3

Head of the Corpus of Political Texts Contained in the Dataset

index	date	link	party	title	text
0	2018-12-18 00:00:00	http://www.sdp.hr/aktualno/arsen-bauk-predlaze...	SDP	Bauk predlaže da se djeci s teškoćama omogući ...	Predsjednik Kluba zastupnika SDP-a Arsen Bauk ...
1	2018-12-16 00:00:00	http://www.sdp.hr/aktualno/bernardic-najavio-s...	SDP	Bernardić najavio SDP-ov akcijski plan za refo...	Predsjednik SDP-a Davor Bernardić u nedjelju j...
2	2018-12-15 00:00:00	http://www.sdp.hr/aktualno/bernardic-smisao-sd...	SDP	Bernardić: "Smisao SDP-a je smanjivanje nejedn...	Predsjednik SDP-a Davor Bernardić na zadnjoj o...

Source: Dataset created by the authors, on the basis of texts parsed from the websites included in Table 2.

Consequentially, the corpus developed was deemed fit for conduct of natural language processing tasks, such as the analysis of the discursive features related to the policy issue of migration.

Data analysis

Exploratory data analysis was performed on the data collected. Strong disbalance within the dataset has been identified accordingly both to party affiliation and to the year of publication. We believe that unbalance in the data extracted is representative of non-uniform behaviour across parties and across time, with regards to the usage of party websites as tools for public information activities. Figure 1 contains the breakdown of texts in our dataset, grouped by political party.

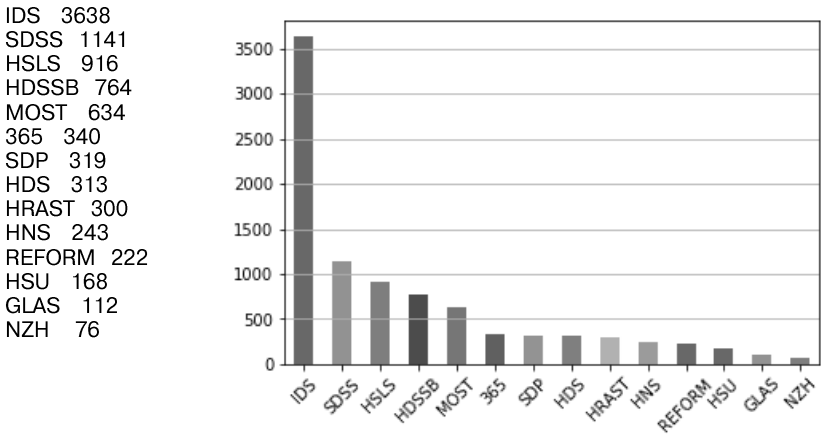


Fig. 1
Distribution of texts grouped by party
Source: Authors, on the basis of information contained in the dataset.

It can be easily noticed that distribution of texts is skewed. Five most-verbose parties, alone, produce 77% of total texts present in our dataset. Thus expected that they would contribute more in determining the political features associated to the issue of migration; after labelling the data, we can see this may not necessarily be the case. Unbalance noted, in our judgement, is a reflection of different natural behaviour of political parties studied. Based on this, assumption made is – more a given party publishes, regardless the topic, it will have higher influence on the public political discourse. All texts present in the dataset will be treated as equal during the machine learning phase of this research.

Figure 2 shows that all of texts are sufficiently recent, which becomes second characteristic of the research.

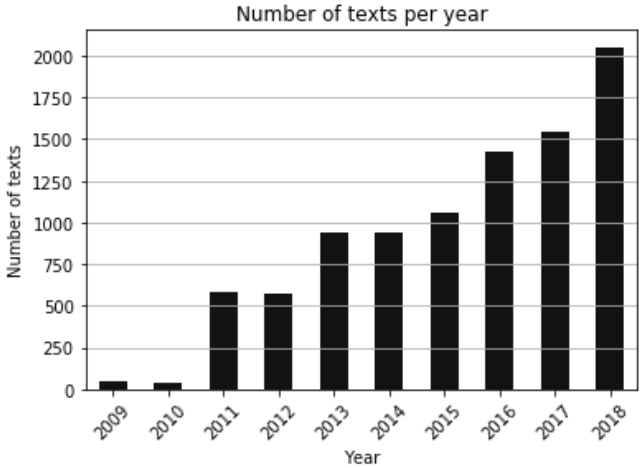


Fig. 2
Distribution of texts grouped by year of publication
Source: Authors, on the basis of information contained in the dataset.

Most texts have been published in the last few years. Only a handful has been published before 2011. Having all of the retrieved texts being published in the period of interest, we did not deem it necessary to further subset the dataset.

Preprocessing of data retrieved

This step includes removing of stopwords, tokenization and stemming of the whole corpus.

Stopwords, the most frequent words in any given language, such as conjunctions and personal pronouns with little semantic value, were removed first. List of stopwords used is slightly modified version of the ones retrieved from GitHub[6], since Croatian stopwords are not currently included in NLTK, the

[6] Specifically, we used Gene Diaz's list of stopwords retrieved from: https://github.com/stopwords-iso/stopwords-hr/ We have also used the stopwords which are contained in the code for the stemmer we selected (see next footnote), and have finally added some more stopwords which were missing in the original two lists that we used.

standard Python package for NLP. After removing stopwords from texts, we tokenized remaining characters accordingly to the regular expression "\w+", which returns all groups of alphanumeric characters present in a string. Each token was additionally converted to lowercase as necessary. Next step was to stem each token, by using an open-source rule-based stemmer which was developed by Nikola Ljubešić and others [21][7]. The collection of stemmed tokens was then used to compute the Bag-of-Words matrix associated with the corpus of texts. The BoW matrices were computed by excluding all tokens containing one of the keywords used for labelling the data, as described later in this paragraph and than computing the absolute frequencies of occurrence of unigrams, unigrams and bigrams together, and bigrams alone. Three BoW matrices, which could be fed to our classificators were obtained.

Last step in the preprocessing of data was to label it. To do so we employed an automatic method for labelling. An arbitrary list of keywords, unequivocally associated with the policy issue of migration has been made. Same list was tested against the dataset and was progressively reduced until it contained the minimal number of keywords that would provide the highest marginal gains. Keywords which passed the procedure, or rather their stems, are enumerated in the table 4.

Table 4

Keywords used for the automatic labelling of the texts

Keyword	Meaning
'migrac' and 'migran'	Migration, migrant, and compound words
'izbegl'	Refugee
'^azil'	Asylum. The caret marks the beginning of a token
'raseljen'	Deportation
'useljavanj'	Immigration
'iseljavanj'	Emigration

Source: Authors, on the basis of the apriori knowledge of the researchers.

Selection of these particular keywords is largely arbitrary and ultimately derives from the *a priori* knowledge of the researchers on what "migration" means. There was only 403 out of 9185 texts that contained at least one of the keywords. It is only 4.38% of the whole corpus that was labeled positively for a binary classification task. Both, automatic and manual inspection of results has been carried out. Manual inspection verified machine learning findings. Automatic inspection in order to check for particularly unbalanced distribution of texts was also carried out. Findings are graphically envisaged below.

As shown, relative distribution of positives across parties is sufficiently homogenous, albeit a bit skewed. It can be additionally noted that the parties which produce more texts do not necessarily produce higher quotas of texts related to the

[7] The stemmer itself can be found on: http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/ (accessed: 21 December 2018). Minor modifications to the code were implemented by us so that it could work from memory rather than hard-drive, in order to include the stemmer into the machine learning pipeline.

issue of migration. Correlation coefficient between the distribution of relative frequencies of positives per party, and the overall number of texts, positive and negative, produced per party, is -0.27. This shows that there is no significant relation between the number of texts produced and importance of the issue of migration. Parties which publish more, in general are not necessarily more concerned about migration, similarly, parties that publish less, are not necessarily less concerned about migration (Figure 3 and 4).

365 14
GLAS 13
HDS 9
HDSSB 29
HNS 15
HRAST 13
HSLS 13
HSU 5
IDS 113
MOST 46
NZH 9
REFORM 4
SDP 34
SDSS 86

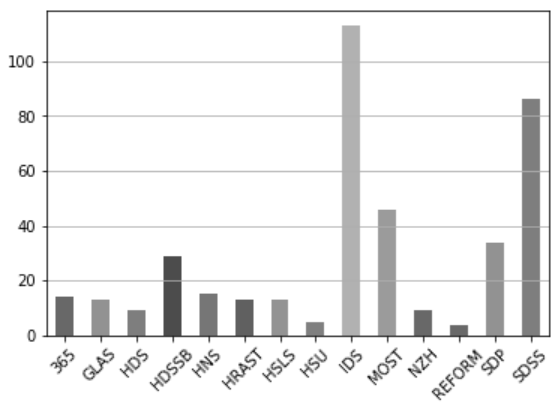
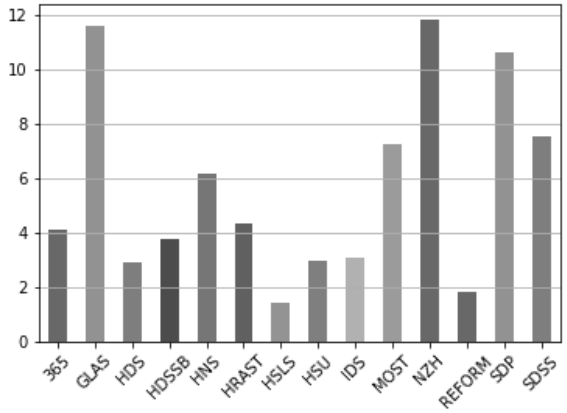


Fig. 3

Percentage of positives over the total texts for each party

365 4.11
GLAS 11.60
HDS 2.88
HDSSB 3.79
HNS 6.17
HRAST 4.33
HSLS 1.41
HSU 2.97
IDS 3.10
MOST 7.25
NZH 11.84
REFORM 1.80
SDP 10.65
SDSS 7.53



Min: 1.41%, Max: 11.84%, Mean: 5.67%, STD: 3.58%

Fig. 4
Absolute frequencies of positives on migration per party
Source: Authors, on the basis of information contained in the dataset.

Development of the machine learning model

The Bernoulli Naive Bayesian Classifier, the Support Vector Machine and the Decision Tree machine learning models were tested for accuracy. Hyperparameters of models were fine-tuned with grid search. As accuracy measure we used the F1 score of the models' predictions, which is a metric suitable for binary classification

tasks such as ours, when the two labels are unbalanced [22]. We used the Bag-of-Words matrices computed on unigrams, unigrams and bigrams, and bigrams alone, as input data, while the input labels were the ones calculated accordingly to the procedure described in the paragraph above. It is important to remind, as stated above, that the three matrices were calculated by explicitly excluding any and all tokens which contained stems of words used to label the data. As a consequence, our classificators would not be able to learn the rule we used to automatically label the data, which would result in a trivial and predictable output. Instead, by blinding the classificators to the words used to label the data, we could train them to find what other predicting features are present in the text themselves, and study them afterwards. Keeping this clarification in mind: the classificators did not see the keywords we used to label the data.

Next step was to train each of the three models on each of the three types of Bag-of-Words matrices, and measure the F1 score for each model for each matrix after fitting the models. Result of this experiment is reported in Table 5. The F1 score is truncated to the second decimal digit.

Table 5

F1 scores of the tested Machine Learning algorithms

Classifier	Input matrix	F1 score
Bernoulli Naive Bayesian	Unigrams	0.24
	Unigrams and bigrams	0.18
	Bigrams	0.16
Support Vector Machine	Unigrams	0.31
	Unigrams and bigrams	0.58
	Bigrams	0.26
Decision Tree	Unigrams	0.95
	Unigrams and bigrams	0.98
	Bigrams	0.97

Source: Authors, on the basis of the output of the program.

Training and scoring was repeated multiple times with different random seeds to account for randomness. The results were all similar. On the basis of the F1 scores calculated, decision tree classifier trained on unigrams and bigrams was selected as the best performing algorithm, and was analysed further.

Output of the model and assessment of results

Model structure for the best performing Decision Tree, the one trained on unigrams and bigrams was computed and displayed. Many rules have been identified, too many to be discussed thoroughly. Few of the rules identified have been selected, in order to discuss and interpret them here in more detail. All these rules are either located by the root or the tree, or in close proximity to it, as indicated case by case.

Rule 1. If (demographical) and (reduction), then "migration"[8] (Figure 5).

[8] The texts that result from this rule are accessible at the following links, as of December 2018.

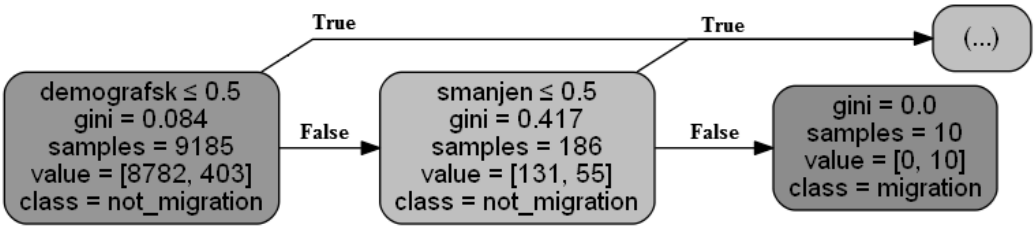


Fig. 5
Demography and Reduction are words which characterise migration
Source: Authors, on the basis of the output of the program.

The root of the tree as well as the the location of the first split in the dataset is shown above. If both words "demographical" and "reduction" are simultaneously present in a given text, then the text is a text about migration. This is an important rule, because it is at the root of the tree and it is thus very easy to interpret, being comprised of just two chained logical propositions. While the word "demographical" is, indeed associated with migration in the sense that migration is a demographical phenomenon. Studying decision rules isn't giving us a clear explanation on whether demography was mentioned in the sense of a demographical increase, a demographical decrease, a demographical variation of the Croatian population, or if some other demographical phenomenon was cited. Individual inspection of texts has been carried out to determine context in which "demography" was mentioned. Articles retrieved accordingly, seem to uniquely refer to the emigration of Croatians from the country and especially of the young and unemployed. The word "reduction" is mentioned in a variety of contexts. It relates at times to the reduction in the level of public expenditures, to reduction of unemployment, to the reduction of taxation, and also, sometimes, to the reduction of the Croatian population as a consequence of emigration. It thus appears that, if a text is about demographics and reductions, then it is about the emigration of Croatians in connection to economic hardships of some kind.

1) URL: http://www.sdp.hr/press/ministar-mrsic-za-jutarnji-list-u-2015-planiramo-povecanje-javnih-radova-koje-ce-financirati-drzava/
2) URL: https://most-nl.com/2018/09/07/planirate-uvesti-red-godine-nereda-jedne-opcije-njihovih-partnera-onda-vam-prvo-kazu-da-politicki-montirano/
3) URL: https://most-nl.com/2018/07/30/most-nezavisnih-lista-zakon-subvencioniranju-stambenih-kredita-jedino-doveo-do-rasta-cijena-nekretnina/
4) URL: https://most-nl.com/2018/06/10/ministarstvo-demografije-institucija-bez-stvarnog-smisla/
5) URL: https://most-nl.com/2017/09/04/hrvatska-treba-stambenu-politiku-a-ne-zastitare-na-ulazu-apn-a/
6) URL: http://www.ids-ddi.com/vijesti/aktualno/6063/demetlika-porazavajuca-demografska-slika-hrvatske-nije-uzrok-nego-posljedica-problema/
7) URL: http://www.ids-ddi.com/vijesti/aktualno/5180/demetlika-decentralizacija-ostaje-mrtvo-slovo-na-papiru/
8) URL: http://www.365ris.hr/mjere-demografske-politike-gradu-zagrebu-mjere-podrske-djeci-mladima-obiteljima/
9) URL: http://sdss.hr/klub-sdss-a-podrzava-prijedlog-zakona-o-poljoprivredi/
10) URL: http://sdss.hr/prvo-citanje-prijedloga-zakona-o-potpomognutim-podrucjima/

Rule 2. If (border) and (Croatia – Serbia), then very likely "migration"[9].

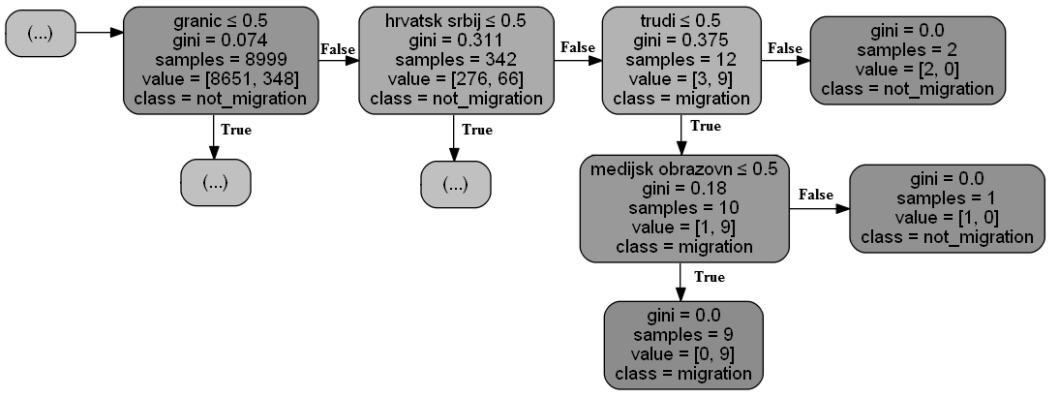


Fig. 6
Borders and Croatia-Serbia are words which characterise migration
Source: Authors, on the basis of the output of the program.

We believe that the first two terms of the IF clause stated above are the most important ones, among the ones represented in the selected branch. Remaining two predictors, "trudi"[10] and "medijsk obrazovn" do not deem to have great explanatory power, not for a human, at least[11]. First two conditions alone, identify a subset of 12 documents, 9 of which correctly relate to the issue being studied[12]. We remind the reader that a randomly selected text from the corpus has a 4.38% chance of being related to migration, with comparison to a 75% chance if the rule indicated above is respected when performing the non-random selection. We have manually inspected the positives retrieved. All of those texts contain a regional dimension. Both Croatia – Serbia, appear in all texts. Texts retrieved accordingly are all about

[9] The texts that result from this rule are accessible at the following links, as of December 2018.
1) URL: http://sdss.hr/kolektivizacija-krivice-osim-sto-je-nepravedna-sjeme-je-zla/
2) URL: http://sdss.hr/aleksandar-vucic-srbima-u-rh-hvala-vam-sto-cuvate-srpsko-ognjiste-ime-i-prezime/
3) URL: http://sdss.hr/nerazumni-ljudi-smatraju-da-je-izvinjenje-uslov-da-se-razgovara-razgova-rajmo-i-stvorimo-pretpostavke-za-ozbiljnu-gestu-izmirenja/
4) URL: http://sdss.hr/pupovac-za-n1-vazno-je-da-susret-dvoje-predsjednika-bude-pragmatican-i-konkretan/
5) URL: http://sdss.hr/u-beogradu-promivisana-knjigavreme-sporta-i-razonode-titina-hrvatska-i-njeni-srbi-1951-1971-autora-cedomira-visnjica/
6) URL: http://sdss.hr/predsednik-pupovac-za-tanjug-interes-hrvatske-i-srbije-je-otvoren-dijalog-o-svim-pitanjima/
7) URL: http://sdss.hr/program-samostalne-demokratske-srpske-stranke/
8) URL: http://sdss.hr/%d0%bf%d1%80%d0%be%d0%b3%d1%80%d0%b0%d0%bc-%d1%81%d0%b0%d0%bc%d0%be%d1%81%d1%82%d0%b0%d0%bb%d0%bd%d0%b5-%d0%b4%d0%b5%d0%bc%d0%be%d0%ba%d1%80%d0%b0%d1%82%d1%81%d0%ba%d0%b5-%d1%81%d1%80%d0%bf%d1%81/
9) URL: https://www.hns.hr/vijesti/politicka-akademija/hrvoje-koscec-na-konferenciji-european-week-of-regions-cities/
[10] To put effort, verb.
[11] Media, adjective; education, noun.
[12] It should be added that, immediately to the left of the first decision rule represented in the graph, is present the root of the graph itself. Specifically, the condition "demographical" ≥ 1 must be valid, so that this branch of the tree is activated.

migration in historical terms, as a consequence of the Croatian war of independence in the 1990's. As an outcome of the war, Serbian minority has been relocated or displaced. This is the context in which those texts relate to migration. It is important to mention that the texts classified accordingly to this rule belong for the most part to the political party SDSS (8 texts out of 9), and only residually to HNS (1 out of 9). No other party is represented in this subset of texts. It appears that the SDSS, the Autonomous Serbian Democratic Party, is the most concerned of all parties about the regional, intra-Balkan dimension of the phenomenon of migration.

The presence of Rule 2 can however be considered as a confirmation of the theoretical prediction performed earlier, regarding the expectation for the political discourse in Croatia on migration to be discussing the Serbo-Croatian conflict and its consequences.

Rule 3. If (Marrakesh agreement), then "migration"[13].

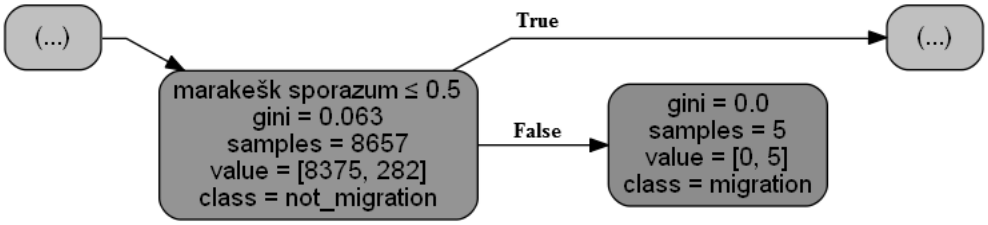


Fig. 7
The words Marrakesh Agreement characterise migration
Source: Authors, on the basis of the output of the program.

The Marrakesh agreement is a name, commonly used by local political parties, which refers to an international agreement formally known as *Global Compact for Safe, Orderly and Regular Migration*. This is an intergovernmental agreement signed in Marrakesh on December 2018 [23]. The words "Marrakesh Agreement", in International Law and in English language, commonly identify the international treaty by the same name on which the WTO was established[14], which is not an agreement about migration. We set to manually inspect this peculiar characteristic of the retrieved texts, which seemed to systematically misuse a term in place of another. Texts retrieved accordingly talked, indeed, about the Global Compact for migration, due to the fact that its incumbent signature on the part of the Croatian government was at the time an important topic for heated political discussion.

[13] The texts that result from this rule are accessible at the following links, as of December 2018.
1) URL: http://www.sdp.hr/aktualno/bernardic-najavio-sdp-ov-akcijski-plan-reformu-pravosuda/
2) URL: http://www.neovisni.hr/kresimir-kartelo-marakeski-sporazum-odbacili-su-svi-s-nacionalnim-mozgom/
3) URL: http://www.neovisni.hr/sto-je-skriveno-u-marakeskom-sporazumu/
4) URL: https://glas.com.hr/2018/11/11/nikad-si-necu-oprostiti-sto-nisam-probila-blokadu-u-koloni-sjecanja-u-vukovaru/
5) URL: https://glas.com.hr/2018/11/06/problem-migracija-tek-je-poceo/

[14] Marrakesh Agreement establishing the World Trade Organization (with final act, annexes and protocol). Concluded at Marrakesh on 15 April 1994. Full text available at: https://treaties.un.org/doc/publication/unts/volume%201867/volume-1867-i-31874-english.pdf

Conclusion

It is possible to build a system that allows determination of what political features characterise the issue of migration in the public information campaign of Croatian political parties. The system requires very little *a priori* knowledge on the part of the researchers on the structure of the political party system in Croatia and also of the issue of migration itself. This system does not rely on human judgement on the part of the researchers, and can be thus considered to be "objective", short of possible sampling or selection bias. It is replicable. If provided, the same dataset and algorithm used, same conclusions should be reached by any scientist.

The dataset was developed by identifying political parties of interest, on the basis of the list of parties currently represented in the Croatian Parliament. Their websites were searched, crawled and parsed as much as technically possible. Dataset was created containing a few thousand news items. Texts were than labelled on the basis of whether or not they contained keywords unequivocally associated with the policy issue being studied. Determination of those keywords was done through human judgement, and it is the only part of this methodology which is not clear how to automate. Machine learning algorithms were tested and the decision tree classifier was deemed the most suitable. By analysing decision rules we identified several political features which characterise the issue of migration in the Croatian political discourse. Three of which were found specifically interesting and due to that were further analysed, forming the body of this analysis.

Political conclusion reached is that Croatian political system confirms the theoretical paradigm stated in literature [19] about traditional division between conservatists, who are against immigration, and liberals, who upbear the process mentioned. Research further highlights the fact that political position of population towards migration is shaped no longer on the exclusive basis of real-world observations and interactions, but increasingly more by messages which are received in the digital sphere and which do not necessarily correspond to real-world events [14]. Alongside developed machine learning system that can be replicable, in political and sense of political science the research shows prevailing regional dimension. Moreover, most retrieved texts have an intra-Balkan dimesion and focus on migration of Croatian citizens in connection to economic hardship and migration in historical term, as a consequence of the Croatian War of Independence in the 1990's in light of the Serbo-Croatian conflict.

References

[1] Geddes B. How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political analysis*. 1990; (2): 131–150.
[2] Pittman J.A., Yang Zh., Yu S. *Political Cycles and Analyst Bias.* 2018. doi: 10.2139/ssrn.3262070
[3] Olsen M., Harvey L.G. Computers in intellectual history: lexical statistics and the analysis of political discourse. *The Journal of Interdisciplinary History*. 1988; 18 (3): 449–464.
[4] Gavrilova M.V. Political discourse as object of linguistic analysis. *Polis. Political Studies*. 2004; 3 (3): 127–139.
[5] Van Dijk T.A. What is political discourse analysis. *Belgian journal of linguistics*. 1997; 11 (1): 11–52.

[6] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. Natural language processing (almost) from scratch. *Journal of machine learning research*. 2011; 12: 2493–2537.
[7] Bebić D. The role of the Internet in political communication and promoting political participation of citizens in Croatia: Internet election campaign 2007. *Media Studies.* 2011; 2: 3–4. (In Croat.).
[8] Ostojic R. *A European Perspective of the Migration Crisis: Russian Experiences*. Zagreb: Friedrich Ebert Foundation; 2016. (In Croat.).
[9] Sharich T. Escape from socialist Yugoslavia-illegal emigration from Croatia since 1945. by the early sixties of the 20th century. *Migration and ethnic themes.* 2015; (2): 195–220. (In Croat.).
[10] Žižić J. What is political emigration in Croatia? *Political analysis.* 2013; 4 (16): 61–64. (In Croat.).
[11] Sundhaussen H. *Forced ethnic migration*. Institut für Europäische Geschichte; 2010.
[12] Felberg T.R., Šarić L. In transit: Representations of migration on the Balkan route. Discourse analysis of Croatian and Serbian public broadcasters (RTS and HRT online). *Journal of Language Aggression and Conflict*. 2017; 5 (2): 227–250.
[13] Vezovnik A., Šarić L. Subjectless images: visualization of migrants in Croatian and Slovenian public broadcasters' online news. *Social Semiotics*. 2020. 30 (2): 168–190.
[14] Šaric L., Felberg T.R. Representations of the 2015/2016 "migrant crisis" on the online portals of Croatian and Serbian public broadcasters. *Migration and Media: Discourses about identities in crisis.* 2019; 81: 203.
[15] Ragazzi F., Balalovska K. *Diaspora politics and post-territorial citizenship in Croatia, Serbia and Macedonia.* CITSEE Working Paper Series. 2011; 18.
[16] Ragazzi F. The Croatian 'diaspora politics' of the 1990s: nationalism unbound? Croatian 'Diaspora Politics' of the 1990s: Nationalism Unbound? In: U. Brunnbauer (ed.). *Transnational Societies, Transterritorial Politics: Migrations in the (Post-) Yugoslav Region, 19th–21st Century*. 2009.
[17] Knezović S., Grošinić M. *Migration trends in Croatia*. Zagreb: Hanns-Seidel-Stviftung, Institute of development and international relations, Kolor Klinika; 2017: 1–39.
[18] Rovny J. The other "other": Party responses to immigration in Eastern Europe. *Comparative European Politics*. 2014; 12 (6): 637–662. doi: 10.1057/cep.2014.25
[19] Gregurović M., Kuti S., Župarić-Iljić D. Attitudes towards immigrant workers and asylum seekers in eastern Croatia: dimensions, determinants and differences. *Migration and ethnic themes*. 2016; 32 (1): 91–122.
[20] Nadkarni P. M., Ohno-Machado L., Chapman W.W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association.* 2011; 18 (5): 544–551.
[21] Ljubešić N., Boras D., Kubelka O. *Retrieving information in Croatian: Building a simple and efficient rule-based stemmer*. 2007.
[22] Lipton Z.C., Elkan C., Naryanaswamy B. Optimal thresholding of classifiers to maximize F1 measure. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Berlin: Springer; 2014: 225–239.
[23] Assembly U.G. Global Compact for Safe, Orderly and Regular Migration. *International Journal of Refugee Law.* 2018; 30 (4): 774–816.

Information about the authors:
Gabriele De Luca – PhD Student of the Department of Comparative Politics of RUDN University (Russia) (ORCID ID 0000-0001-9728-9581 (e-mail: gabriele.deluca@mail.ru).

Marko Beck – PhD Student of the Department of Comparative Politics of RUDN University (Russia) (ORCID ID 0000-0002-9441-2617) (e-mail: beck.marko@gmail.com).

Информация об авторах:
Габриэле Де Лука – аспирант кафедры сравнительной политологии Российского университета дружбы народов (ORCID ID 0000-0001-9728-9581) (e-mail: gabriele.deluca@mail.ru).

Марко Бек – аспирант кафедры сравнительной политологии Российского университета дружбы народов (ORCID ID 0000-0002-9441-2617) (e-mail: beck.marko@gmail.com).