Populism monitoring via computer-based methods for data extraction, cleaning and analysis.

George Makris

Aristotle University of Thessaloniki

DEFINING POPULISM

During the last decades there has been an intensification on the study of populism. Studies of the populist phenomenon take place both on policy level (e.g. rise of populist parties and movements, effects of those actors on the party system etc.) and on discourse level.

The most common scientific definition in the field of populism is given by the "ideational approach" (Kaltwasser et al., 2017; Mudde & Kaltwasser, 2017). According to this definition, the populist ideology considers society to be separated into two antagonistic groups: the people and the elite. More specifically, people is a homogeneous group that is morally good and pure, and elite is a group that is corrupted and opposes the people placing their own interest over the people's. Also, according to the populist ideology, politics should be an expression of the people's general will.

Although, there is still a debate in the scientific community regarding a precise conceptual clarification of the term populism, it seems like the aforementioned definition is the most capable to make up a methodological starting point for empirical studies and research on the populist phenomenon (Mudde, 2004; Mudde & Rovira Kaltwasser, 2018).The reason for that is the fact that with this particular definition it is easier to classify texts as populist (or not) as well as to detect populism, in general (Mudde & Rovira Kaltwasser, 2018). Also, the ideational approach has provided the research community with the capacity to study populism both from the supply and the demand side, using various methods (Hawkins & Rovira Kaltwasser, 2017). Therefore, this will be the approach followed in the present paper for the detection of populism on the internet.

COMPUTER SCIENCE AND POLITICAL SCIENCE

Over the past few decades, the major innovations in the intersection between political science and computer science have been featured in Political Analysis, which is the official journal of the Society for Political Methodology and the Political Methodology Section of the American Political Science Association (APSA). For the purposes of this paper, we will refer to text data collection and analysis and machine learning. In fact, two issues have been published that condense all relevant journal articles on each of these topics (Cranmer, 2017; Roberts, 2016).

Content analysis (newspaper articles, political speeches, party manifestos and so on) is one of the methodologies that has been well established for years in political science. The use of computers, however, has in recent years led to new methodologies for automated content analysis. This innovation has helped (and will help much more in the future) political scientists to analyse large volumes of textual data ('text-as-data'), as it is inherently impossible for a researcher to read very large volumes of text, and very expensive to hire many coders to analyse huge volumes of text. Therefore, automated analysis enables any researcher to analyze and draw reliable results about large volumes of textual data easily and quickly on their computer (Grimmer & Stewart, 2013).

As far as text analysis is concerned, there are many different computational methods. One of them is supervised machine learning methods (Grimmer & Stewart, 2013). In this case, the researcher is faced, for example, with a classification problem, where he/she needs to predict in which category each text belongs to. Thus, he reads a part of the texts himself, manually classifies it and trains an algorithm on the classified data (training set) in order for the algorithm to predict the rest of the texts on its own (test set). As for the evaluation of the model, this can be done in two ways. Either by having the researcher manually code a few more texts and evaluate the model already trained on them (and then apply it to the rest) or by dividing the original data (training set) into k random subsets (k resampling folds) by training the algorithm on k-1 of them and evaluating it on the last one, then doing the same for each fold.

Clearly there are other methods for text analysis, such as dictionary methods (Grimmer & Stewart, 2013) in which there are dictionaries that assign each word a specific tone score and then each word is counted in the text in order to compute an average tone score in the text. The most common application of this method is for sentiment analysis which will be discussed below.

So far in the literature there are not enough scientific articles on the study of populism through computational methods. However, some of the most interesting contributions are two. First, by Cocco and Monechi (2021), who use machine learning on 268 election manifestos from 99 parties to classify sentences as populist or not and based on this classification calculate a populism score for different parties at the international level. Second, by Hawkins and Silva (Hawkins & Silva, 2018) who attempt to detect populist discourse in the election manifestos of 144 parties in 27 countries using two methods, one using human factor and one automated method using machine learning, concluding that the latter can be very useful if and when there is enough data (in this case, enough manifestos).

OBJECTIVE

The aim of this paper is to test, find and propose computational methods that will facilitate those who wish to detect and analyse populism in large volumes of text, without having to read and code all the texts, but only a part of them. For this purpose,

methodologies from the field of automated content analysis, as described above, were used.

DATA

Data collection was carried out using the web scraping method. The data was then cleaned and four methods were applied to the analysis to detect populism: sentiment analysis, detection of "enemies" of the people, a combination of the previous two and machine learning.

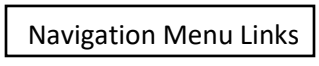
The data of the present study are articles from 6 well-known news websites. kathimerini.gr, efsyn.gr, avgi.gr, tovima.gr, tanea.gr, protothema.gr. The websites were chosen because of their high user traffic. The time period during which the articles were collected was 1-27 February 2022. The articles were collected using the web scraping method and were collected three times a day with a time interval of 6 hours between them (10:00, 16:00, 22:00). The total number of articles collected was 4751.

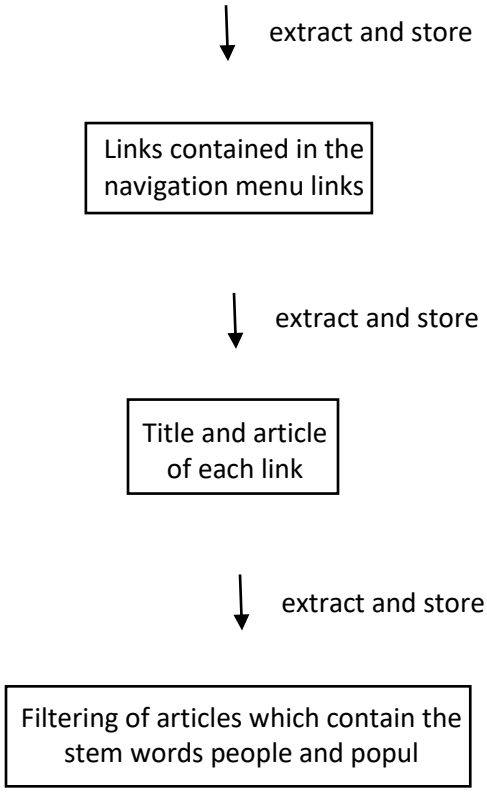
Web Scraping

Web scraping is a method for extracting data from web pages. In political science it is not a widespread method, although it has been used (Jackman, 2006). All web pages consist of three elements. HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets) and Javascript. These elements make up the front-end portion of web pages, i.e. what the user sees. The server of a web page sends these three elements to the user's browser and the browser in turn processes the code it receives to create the web page for the user to view. HTML is used to create the basic structure of the web page, CSS is used to style it, and Javascript is used to create the interactive elements of the web page.

More specifically, what we are interested in in terms of web scraping is mainly the HTML and CSS code of a web page. There is a tool as a google chrome browser plug-in, the selector gadget, which allows the user to interactively navigate with his mouse to the web page he wants, select the information he wants to extract and quickly and easily find the CSS selector that leads to that information. In this paper, the selector gadget was used to locate the CSS selectors and HTML elements that lead to the articles and article titles. Then, using these elements and the rvest library in R an algorithm was created which consists of four steps. The logic of the algorithm is visualized in the figure below (figure1).

Figure 1: Web Scraping Algorithm





ANALYSIS

Prior to the analysis, the data was cleaned. Cleaning as a process involves the removal of all 'noise', such as the most frequent stopwords (e.g. articles, pronouns, etc.) and less frequent words, punctuation, capital letters, tones etc. All of the above were carried out for the data in this study. For the removal of stopwords a list containing all Greek stopwords was used.

The analysis was carried out using four methods: sentiment analysis, identifying words referring to the "enemies" of the people, combining the above two, and finally machine learning. The main objective was to find the best method that gives accurate results (detects populist articles correctly) and limits the volume of articles the researcher needs to read. However, no method can completely replace reading for various reasons (see Grimmer & Stewart, 2013).

SENTIMENT ANALYSIS

As most articles containing the words "people" and "popul" are not necessarily populist (based on the scheme of populism described earlier), it is necessary to

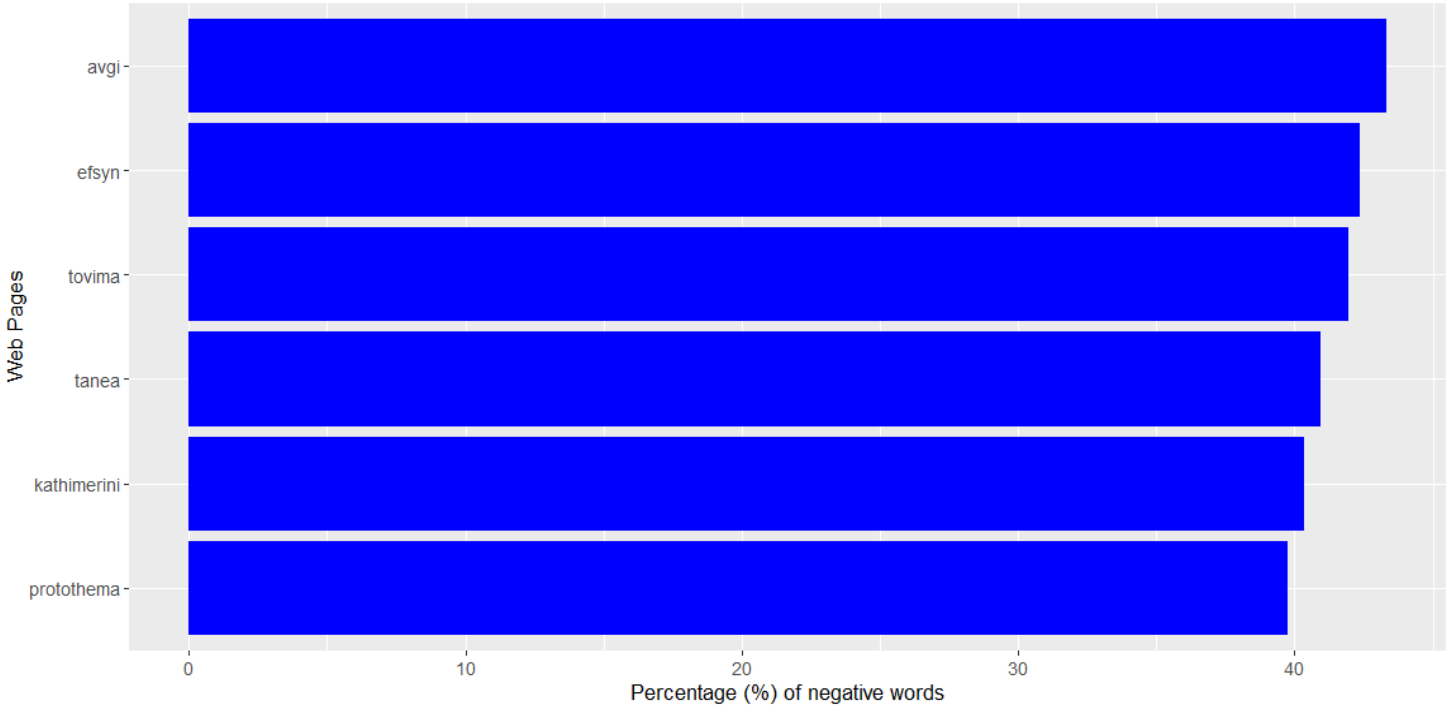
investigate which of them are actually populist and which are not. Therefore, the first method applied to detect populism was sentiment analysis.

Sentiment analysis is a technique for detecting emotions in volumes of text (e.g. anger, fear, joy, etc.) and also polarity. Polarity is usually measured on a scale of -1 (negative emotion) to 1 (positive emotion) and is used to show whether the overall emotion of a text (or set of texts) is positive or negative. Sentiment analysis belongs to the so-called dictionary methods (Grimmer & Stewart, 2013). There are dictionaries where each word is assigned a sentiment score (polarity score), usually from -1 to 1. The words in the dictionary are then counted to the text whose sentiment the researcher wants to measure and an average sentiment score is calculated based on each word's score.

In this paper, the reason why sentiment analysis was chosen to detect populism is because it was hypothesized that an article in which there is a pattern of populism (a "pure" people versus a "hostile" elite) would possibly be more likely to have negativity because of the hostility towards the elites who are considered to treat people in a hostile way. This stems both from the tendency of populist discourse to see the world in a Manichean way, divided into two camps, and from the tendency to moralize politics. It should also be noted that sentiment analysis was not performed to identify individual sentiments, but to calculate the general polarity of the texts.

Below in Graph 1 we see the percentage of negative words on each website (as a percentage of the total words on each).

Graph 1: Relative frequency distribution of negative words, per website



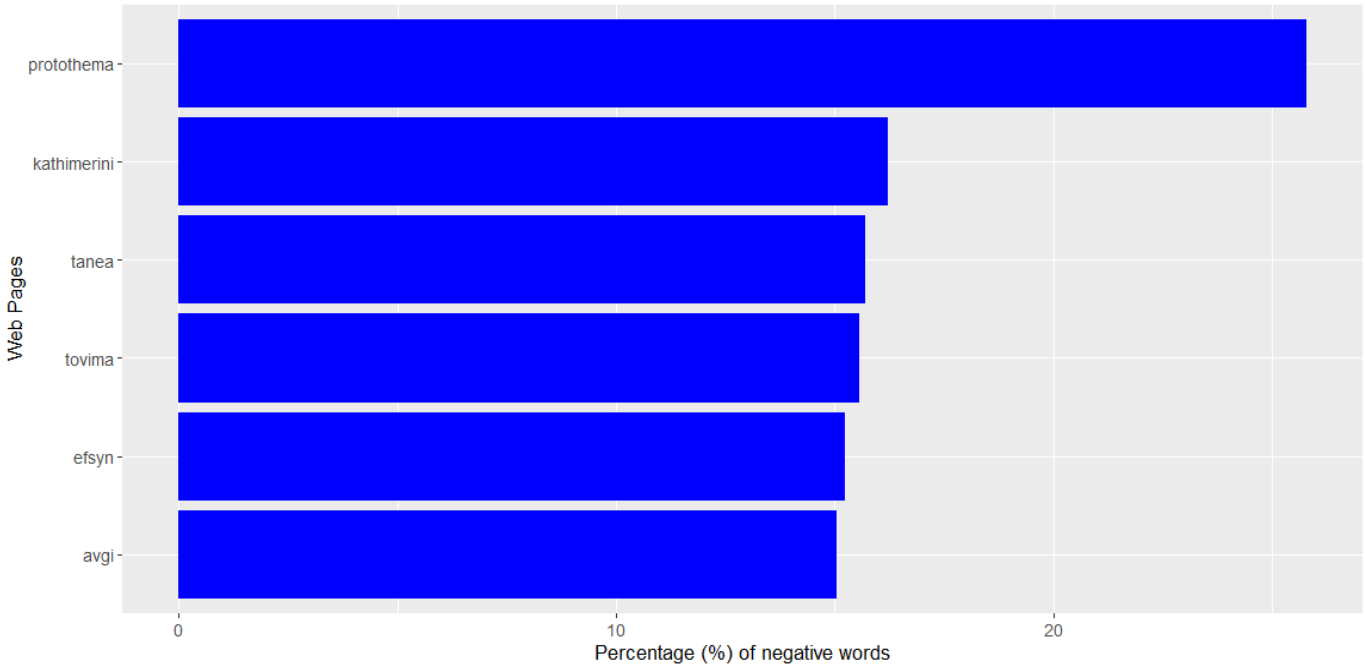
Based on this frequency distribution, we see that the most negative words in comparison are found in the newspapers "Avgi" and "Efsyn", and the least in "Proto Thema" and "Kathimerini", with "News" and "Vima" being somewhere in between. This distribution strongly resembles a left-right ideological distribution. Thus, we see on the left side (the axes are rotated so I mean the top) websites that are indeed leaning to the left.

Avgi is the official newspaper of the leftist and populist party Syriza. “Efsyn” on the other hand is an independent newspaper which is equally characterized by leftist and populist views, drawing articles and columnists often from various leftist parties and movements.

In the middle of the axis we see the two newspapers of the Lambrakis news group, "Nea" and "Vima", which are traditionally placed in the centre-centre-left area. Finally, on the right we see "Kathimerini", which traditionally belongs to the centre-right area, and lastly "Proto Thema", for which we cannot easily give an ideological placement. But it certainly belongs to the broader centrist area in that it often advocates liberal democracy in its opinion articles. This particular distribution is quite similar to a populist-anti-populist one.

At one end we have newspapers which articulate strongly populist and left-populist discourse, while at the other end we have newspapers which are often strongly anti-populist, such as Kathimerini in various opinion articles it publishes from time to time. Therefore, Figure 1 is a first finding that suggests that sentiment analysis may be an effective method for identifying populist articles (or part of them), and it seems to confirm the assumption we made at the beginning of this chapter that populism is more likely to be accompanied by negative sentiment.

Graph 2: Relative frequency distribution of positive words, per website



Graph 2 shows the percentage of positive words on each website. We observe exactly the opposite picture compared to the previous graph. The websites have been ranked in the reverse order, with "Proto Thema" being by far the website with the most, proportionally, positive words. Therefore, we confirm that positive sentiments characterize the less populist (or even anti- populist) newspapers, while negative ones characterize the more populist ones.

And indeed, if you look at the main text of the negative articles, that's exactly what you find. Some extracts from these articles are as follows: "Class unions, not government/employer unions", "Hands off unions, popular struggles and strikes". "But the Greek people have learned them well this time and at the ballot box they are preparing their own, this time, 'epic'." As long as the Right will not forget what the Right means, the Greek people will not forget what the Right means, what Novartis means, what the scandal of the “cover-up” that is in constant progress in this country means", "elections are the only democratic way out of this state of siege".

Therefore, to conclude whether sentiment analysis is indeed effective in terms of possibly identifying populist articles, or, more precisely, limiting the set of articles to those most likely to be populist, a sample of 20 articles was drawn by random sampling from the set of most negative articles (<-1.5), read and coded as populist or not. The 'ideational approach' to populism described above was used as the coding method. That is, articles were coded as populist if the scheme of the 'pure' people vs the 'corrupt' elite was present in them, and if politics was viewed as the expression of the general will of the people.

In this way, 9 out of 20 were identified as populist. In other words, half of the negative articles were indeed populist and the rest were not. This suggests that sentiment

analysis is capable of identifying a fair proportion of populist articles (identifying those that are negative), but certainly not all or most of them.

DETECTION OF "ENEMIES" OF THE PEOPLE

The second method for detecting populism is to detect words in articles that refer to "enemies" of the people. This is a modification of the lexicometric method. The lexicometric method is a relatively simple method that counts the frequency of occurrence of the words "people" and "popul". This method has been applied by Stavrakakis and Katsambekis to detect populism in Alexis Tsipras' speeches (2014).

In the present study, this method was applied with a slight modification. Instead of calculating the frequency of the words "people" and "popul" in the articles, the frequency of words denoting the "enemies" of the people was calculated. As shown in the previous chapter, since the negative sentiment articles are likely to be populist, there is a possibility that these articles reflect an "animosity" towards the elite(s) who negatively patronize the people.

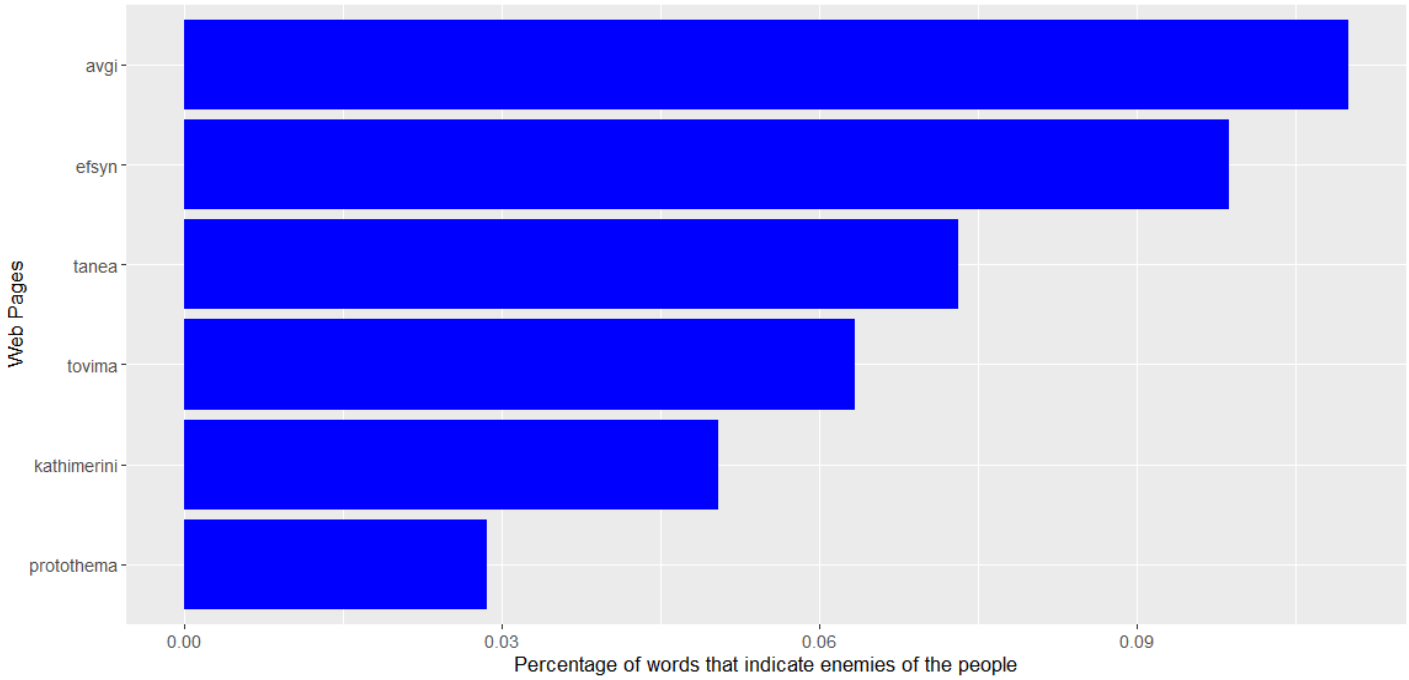
Therefore, if negative articles can lead to the detection of populism, so too is the detection of "enemies" of the people likely to lead to the same result. The words that somehow capture the "enemies" of the people, which were detected in the articles, are as follows: "elite, establishment, corrupt etc."[1] Graph 3 shows the frequency distribution of the words indicating "enemies" of the people, per website.

In this graph we observe a frequency distribution identical to that of graph 1. In this case too, we observe that "Avgi" comes first, then "Efsyn", then the newspapers of the Lambrakis group, and lastly "Kathimerini" and "Proto Thema". This finding suggests three things.

First, that the newspapers with the most negative sentiment are also the newspapers with the most references to the "enemies" of the people and vice versa. Second, the websites follow a distribution that also in this case resembles the left-right axis and the populist-anti-populist axis. That is, newspapers such as "Avgi" and "Efsyn" which have more references to the "enemies" of the people are also more left-wing ideologically, while the newspapers "Kathimerini" and "Proto Thema" which have fewer references are more in the centre-right area. Similarly, the former two usually have more populist content than the latter. Third, it seems to confirm the assumption we made that more negative sentiments may be accompanied by references to "enemies" of the people.

Graph 3: Relative Frequency distribution of words denoting "enemies" of the people, per website

[1] Those words were drawn from the DataPopEu project.



In order to complete the analysis and to confirm whether the method of detecting the "enemies" of the people is indeed effective in detecting populism, a sample of 20 articles was randomly drawn from the articles containing references to the "enemies" of the people. These articles were then read and coded using the 'ideational' approach described above. Of these, 9/20 were identified as truly populist. This suggests to us that this method has similar results to sentiment analysis (in which equally 9 articles were found to be populist from a random sample of 20 negative articles).

To make this more obvious, I quote some extracts from the main text of the above articles: "You have turned from a utility company into a utility company for the blue golden boys and private investors, who will profit at the expense of households, farmers, shops, who see the bills and have a stroke", "While the wage earner, the pensioner, the professional is groaning, you are giving bonuses to yourselves and your children", "by voting against (the government) in their 'face' the policy that sacrifices life, health, income, labour rights, workers' - people's needs as a whole, in order to secure the profits of the few, the choices of the present rotten state and system".

Therefore, in summary, it seems that identifying words in the articles that refer to "enemies" of the people has very similar results to sentiment analysis. Therefore, the two methods may be used interchangeably or complementarily to identify populism or to limit the set of articles to articles that are likely to be populist.

COMBINATION OF SENTIMENT ANALYSIS AND DETECTION OF PEOPLE'S "ENEMIES"

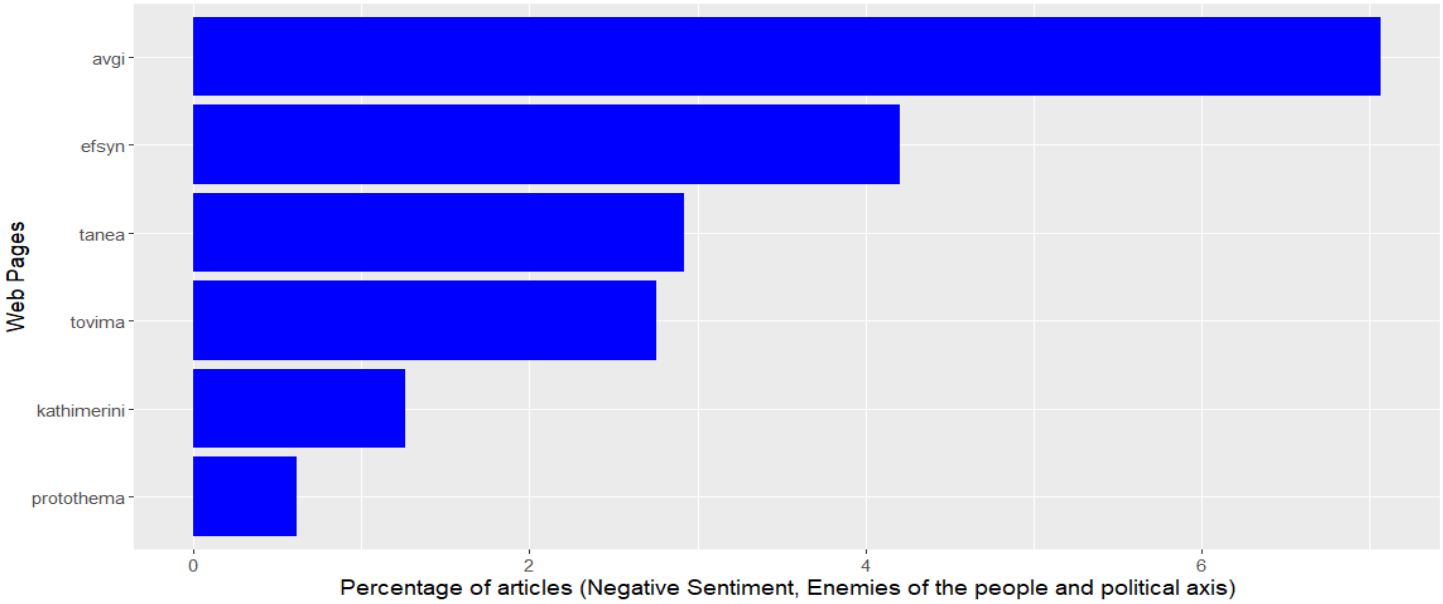
Finally, I tried to combine the two methods mentioned above. Thus, I identified those articles which were simultaneously sentimentally negative, mentioned the

"enemies" of the people and also referred to the political axis-topic as derived from various applications of LDA (Latent Dirichlet Allocation)[2]. From now on, those articles combining all three of these features will be called NEP (from Negative Sentiment, Enemies of the People, Political Axis).

Some short excerpts from the main text of these articles are listed: "Swallowing as a... retaliation to the Greek Prime Minister, their sponsor, any oppositional criticism, denunciation or accusatory reaction, covering with their shameless silence every prime ministerial blunder, every governmental scandal, every rape of the rule of law by Mitsotakis' "executive state"", "One is the list, Mitsotakis' list! Full of sponsors, friends, friends, partners, blue children, "day laborers" and gas-guzzlers who voraciously consume the property of the Greek people without any control and without any shyness...".

It is therefore clear that a combination of the two methods we examined is quite effective in identifying populist articles. This can be seen from the frequency distribution of NEP articles, per website (Graph 4) which is similar to graphs 1 and 3.

Graph 4: Relative Frequency distribution of NEP articles, per website



Finally, a random sample of 20 articles was then drawn from the NEP articles, read and coded as populist or not. Of the 20, 10 were found to be populist.

In conclusion, we can note that both sentiment analysis and the detection of "enemies" of the people in articles can be effective methods for detecting populist articles

[2] LDA results are included in the whole text which is my MSc thesis.

and can be used interchangeably. In fact, when used together they can lead to equally good results (and slightly better)

However, neither method is capable of detecting all of the populist articles, but only of narrowing down the volume and targeting the researcher to articles that are most likely to be populist (either negative articles, articles referring to the "enemies" of the people, or both). The most effective method, which is able to detect most populist articles, seems to be the third method, machine learning, which will be discussed below.

MACHINE LEARNING

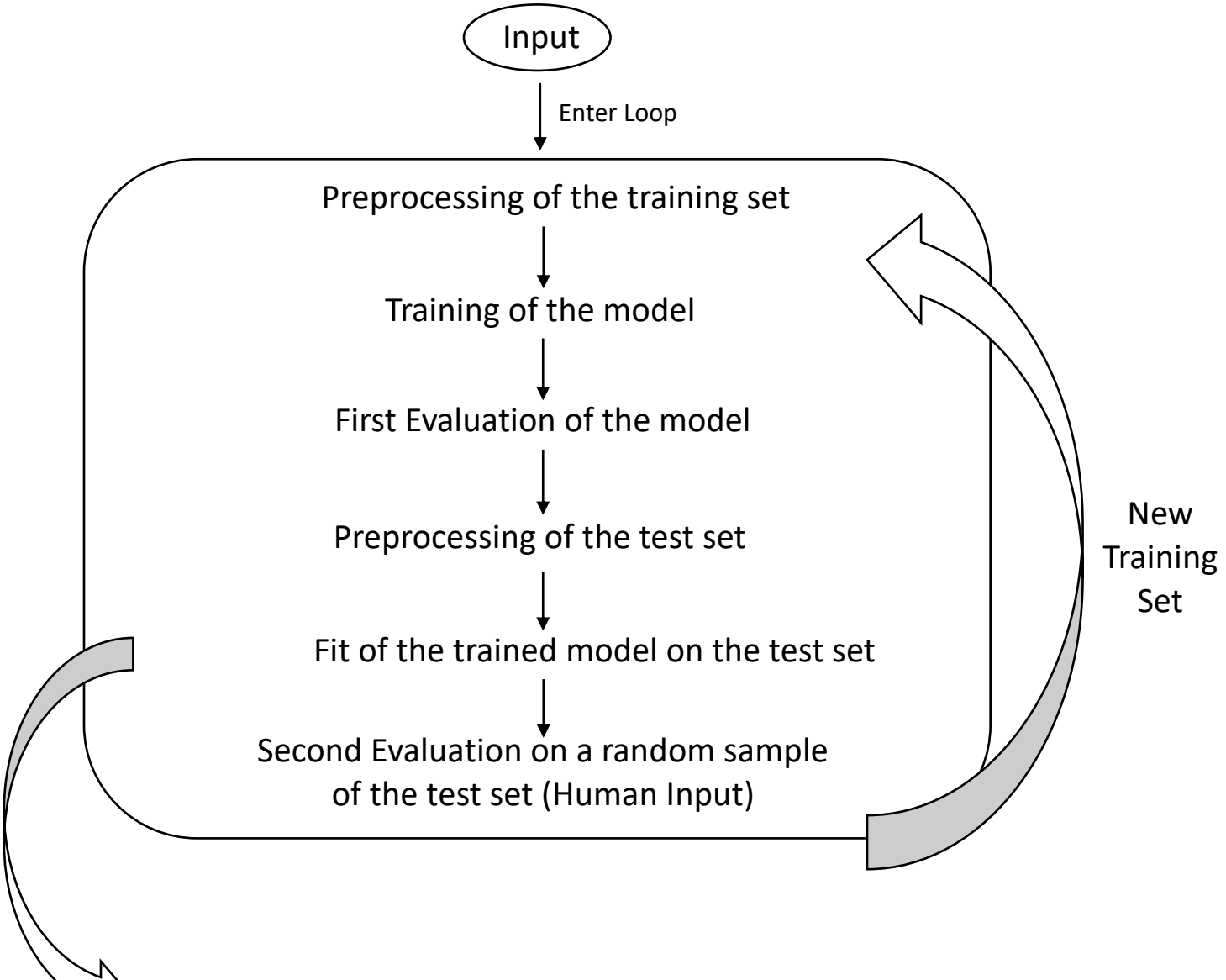
For the detection of populism in this paper, supervised machine learning was used, where a sample of 572 articles coded as populist (1) or non-populist (0) was used as a training set.

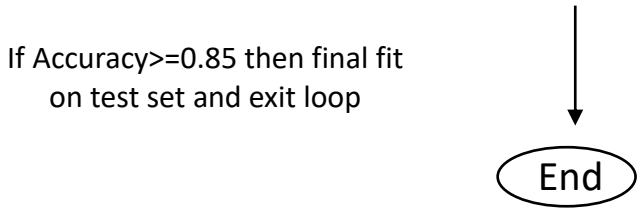
The aim was to train a statistical model on the computer using train data and then classify the rest on its own. The statistical model used was logistic regression.

As the training set in this case is very small, a machine learning model alone cannot be used, as such models need a large amount of data (observations) to be able to give accurate results. Thus, an algorithm of the HITL (Human-In-The-Loop) class was created (Zanzotto, 2019). HITL algorithms are based on human-machine interaction, i.e., the human creates a program so that the computer repeats (loops) an operation, and when it completes this operation it "asks" for a correction from the human to repeat the same operation again in a better way etc. This cycle continues in this interactive way until the computer achieves the desired degree of accuracy. Such algorithms are suitable when there are not enough data to train the computer, and help to achieve the desired accuracy in training without having to spend a lot of time (or years) creating a large enough training set.

THE ALGORITHM FOR THE CLASSIFICATION OF DATA

Figure 2: HITL Machine Learning Algorithm.





ANALYSIS OF THE POPULIST ARTICLES

Below are some analyses performed on the articles predicted as populist by the algorithm. It should be noted here that it is possible that the results may be biased, as the initial coding of the articles was based solely on the subjective opinion and perception of the author. This is obviously one of the limitations of machine learning. Since all machine learning algorithms learn only from training data, it is easy to replicate social biases or prejudices of the researcher using them (Cranmer, 2017). Especially in terms of text analysis, the goal of supervised machine learning in particular is to replicate human coding in an automated way (Grimmer & Stewart, 2013).

Nevertheless, there has been an attempt on the part of the author to be as objective as possible in reading the articles and to identify populism using the 'ideational approach' methodology described above. However, it would be more appropriate for the researcher to establish specific coding rules while reading the articles, and to classify them according to these before proceeding with the machine learning process.

That said, I proceed below to present some analyses of the populist articles predicted by the algorithm.

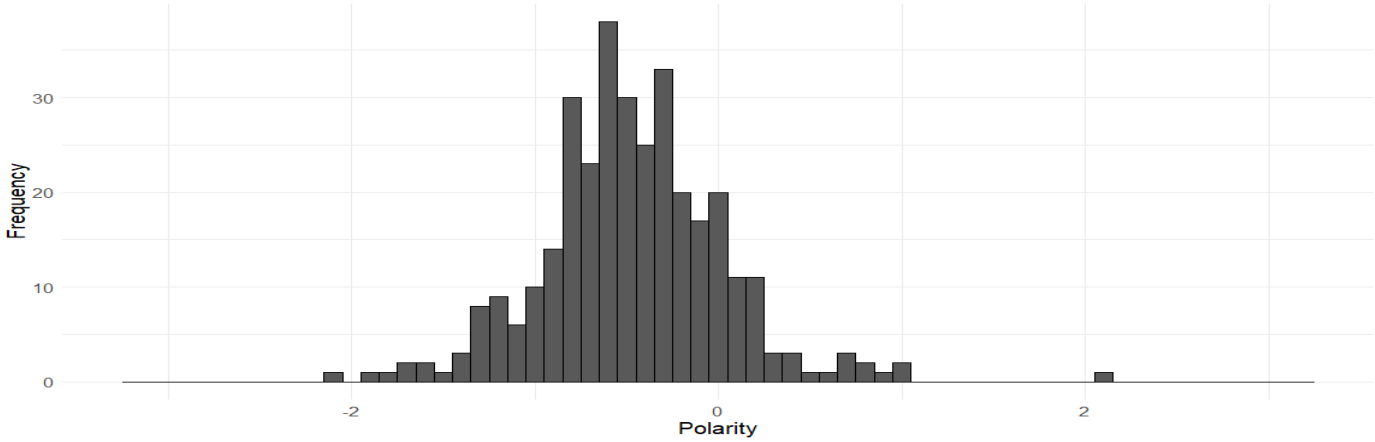
First, out of a total of 4751 articles, 333 were found to be populist, i.e. a percentage of 7%. The following analysis was performed on these 333.

Next, a sample of 20 articles was drawn, by random sampling, from the articles that were classified as populist by the algorithm, in order to manually verify the accuracy of the algorithm. Of the 20 articles, 17 were found to be populist. Therefore, we see that the machine learning algorithm is the most effective of the methods tested for identifying populist articles.

Indeed, if we look at some extracts from the main text of the articles, we see the existence of populism. For example: "he called on the prime minister to stop "the regime policy that tries to hide the voice of people who are struggling and demanding. It is no coincidence that recently one will not see the rural mobilizations on TV channels". "Finally, a conclusion so ridiculous and so provocative was written at Mitsotakis' order. The guilty parties assure the Greek people 'beyond any doubt' that they are innocent!"

We then proceed to a brief quantitative analysis and exploration of the articles classified as populist by the algorithm. First, in Graph 5 we see the distribution of the polarity of the populist articles. We see that the distribution is skewed to the right, suggesting that the majority of populist articles have a negative polarity score. This seems to confirm the claim that populism is more likely to be detected in articles characterized by negativity, which supports that sentiment analysis may be an effective method for identifying a portion of populist articles.

Graph 5: Polarity Distribution of Populist Articles



Finally, Table 1 shows the frequency distribution of populist articles per website. We see that the frequency distribution of populist articles per website closely follows the frequency distribution of all the other methods (Graph 1,3,4). The order in which the websites are ranked is almost the same. The only difference is the position of "Kathimerini" where the algorithm puts it in the middle of the distribution while in the other graphs it appears to be second to last. The other websites are exactly in the same order (from "Proto Thema" the algorithm did not detect populist articles therefore it is considered as being in the last position, just as in the other graphs). Therefore, we see another indication, that machine learning yields similar results to the other methods for the detection of populism, the only difference being that it is more accurate in classification.

Table 1: Frequency and relative frequency distribution of populist articles, per website

Web Pages	f	Rel. f (%)
avgi	307	6
efsyn	12	0.3
kathimerini	6	0.1
tanea	4	0.08
tovima	4	0.08

Regarding the frequency distribution of Table 1, although the algorithm seems to correctly represent the order of the web pages, there seems to be a large variance in the articles between the web pages. “Avgi” seems to be overrepresented.

This problem is the most common problem in machine learning and is called overfitting. Overfitting results from the inability of a machine learning algorithm to use the computed parameters to make inferences based on unknown, new data, even though its performance on the original data (training set) can be very satisfactory (88% accuracy in our case).

Therefore, the possible solutions to combat the overfitting bias would be two in this particular case. The first is to introduce more predictor variables in the model (features as they are called in the language of machine learning) to control the variance of the dependent variable (i.e. train the classifiers in large volumes of data/words). If those predictors (words) are not present then new predictors should be invented, like sentiment scores, word ratios etc.

However, the most important way to solve this particular overfitting problem would be to use Bayesian models that use Bayesian priors. Bayesian priors are distributions whose shape we assume in advance and train the models we want through assumptions we make about the shape of these distributions, their mean, standard deviation, etc. In this case, we can now, after the analyses we have performed, claim to have enough evidence of the shape of the prior distribution of populism by website (Table 1, Graph 4 etc.), which we did not have before (we could only hypothesize that it was something like this), as the distribution obtained by the algorithm and the distribution obtained by the other methods are very similar, so perhaps this is a contribution of the paper towards the use of more effective Bayesian models.

Other models used for the overfitting problem are complex generalizations of linear regression models to which parameters (e.g. λ) are added to remove variables and reduce the dimension of the data space such as ridge regression, elastic net regression and so on. However, in this case it did not make sense to use such a model as the words in the training set were too few in the first place.

Despite the limitations mentioned above, we can say that in general the model gave correct and reliable (as shown by the comparison with the other methods) results and also gave an accurate picture of the distribution of populist articles per website, at least in terms of the order of the websites. The accuracy rate which was achieved with the Human-In-The-Loop approach is 88%, in terms of the fitting of the algorithm to the training set. So what remains in future analyses is to apply the methods described above to avoid overfitting.

A solution to the overfitting problem would probably approximate the actual distribution somewhat more closely by identifying more populist articles from "efsyn" which are probably more in reality and also identifying the populist articles from the "Proto Thema" website which, although much less, certainly exist.

Having made the above methodological observations, we now move on to some final observations. After applying the machine learning algorithm we saw that in general the predicted populist articles are sentimentally negative. After reading a sample of the predicted populist articles, it was found that populism seems to be present in political news and political commentary, mainly from left-wing parties and politicians. Thematically, it concerns populist criticisms of government policies by opposition parties, usually in a showy tone ("they", "have", etc.) on issues such as the war in Ukraine, the economic crisis, etc. And the usual cause of criticism is the "deception" of the people in order to serve economic interests, such as friendly media to the government, etc. Similarly, there is also a criticism at the international level of "imperialism" of the USA, NATO, etc. Finally, the newspapers where most populist articles are concentrated are "Avgi" and "Efsyn". This seems a reasonable result, as Avgi is a party newspaper of the left-wing populist party SYRIZA, while Efsyn is an independent newspaper which draws articles and columnists from various left-wing populist movements and parties.

CONCLUSION

This paper presented four computational methods for detecting populism in a set of 4751 articles from a sample of well-known news websites. The data collection was done by the web scraping method.

The aim was to make an attempt to facilitate through the use of the computer those who wish to study populism in a large volume of texts, without having to read and code all the texts, but only a part of them. These methods were first, sentiment analysis; second, detecting words referring to the "enemies" of the people; third, combining the first and second method; and fourth, machine learning.

The first two seemed equally effective mainly in their ability to identify potential populist texts and, more importantly, to narrow down a large volume of texts by effectively targeting the researcher on potential populist articles. Their use can be done alternately or simultaneously. When the above two methods are combined they seem to be more effective in detecting populism. This also seems to confirm our original premise that when negative sentiment is combined with references to the "enemies" of the people, populist content is quite likely to emerge. Therefore, it is proposed to combine the two methods to detect populism.

However, if the researcher's goal is to identify all, or at least most, populist articles, then neither of the above methods (or their combination) is fully effective in this regard.

In this case, the third method developed and applied, an HITL machine learning algorithm, is recommended, which appeared to be the best of the four, as it is able to correctly identify more populist articles, always replicating the researchers' logic of what is populist and what is not, as previously mentioned. This algorithm is recommended in order to achieve high accuracy while training the algorithm when this is done using very small training sets, as it involves human-computer interaction and dual evaluation, one by the computer and one by human. In this particular application of the algorithm to the data that were collected, the maximum accuracy achieved in training was 88%. However, those who wish, can set this percentage higher and therefore read and correct more articles if they want to achieve higher accuracy.

The final application of the trained algorithm to the test set gave accurate and reliable results. On the other hand, of course, the problem of overfitting, common in machine learning, occurred. However, some methods were presented to combat this problem effectively in future studies. These are: 1) using large amounts of data and 2) using the evidence in this paper about the possible shape of the prior frequency distribution of populism among websites, it may be possible in the future to use a different model for prediction, rather than logistic regression, such as a naive Bayesian model or any other Bayesian model that assumes the existence of a prior distribution whose characteristics we now know or, at least, have enough evidence to assume.

Citations

Cocco, J. D., & Monechi, B. (2021). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. *Political Analysis*, 1–17. https://doi.org/10.1017/pan.2021.29

Cranmer, S. J. (2017). *Introduction to the Virtual Issue: Machine Learning in Political Science*. 9.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hawkins, K. A., & Rovira Kaltwasser, C. (2017). What the (Ideational) Study of Populism Can Teach Us, and What It Can't. *Swiss Political Science Review*, *23*(4), 526–542. https://doi.org/10.1111/spsr.12281

Hawkins, K. A., & Silva, B. C. (2018). *A Head-to-Head Comparison of Human-Based and Automated Text Analysis for Measuring Populism in 27 Countries*. 42.

Jackman, S. (2006). *Data from Web into R.* The Political Methodologist 14(2). https://thepoliticalmethodologist.files.wordpress.com/2013/09/tpm_v14_n21.pdf

Kaltwasser, C. R., Taggart, P. A., Espejo, P. O., & Ostiguy, P. (2017). *The Oxford Handbook of Populism*. Oxford University Press.

Mudde, C. (2004). The Populist Zeitgeist. *Government and Opposition*, *39*(4), 541–563. https://doi.org/10.1111/j.1477-7053.2004.00135.x

Mudde, C., & Kaltwasser, C. R. (2017). *Populism: A Very Short Introduction*. Oxford University Press.

Mudde, C., & Rovira Kaltwasser, C. (2018). Studying Populism in Comparative Perspective: Reflections on the Contemporary and Future Research Agenda. *Comparative Political Studies*, *51*(13), 1667–1693. https://doi.org/10.1177/0010414018789490

Roberts, M. E. (2016). Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science. *Political Analysis*, *24*(V10), 1–5. https://doi.org/10.1017/S1047198700014418

Stavrakakis, Y., & Katsambekis, G. (2014). Left-wing populism in the European periphery: The case of SYRIZA. *Journal of Political Ideologies*, *19*(2), 119–142. https://doi.org/10.1080/13569317.2014.909266

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

This paper was produced at the DataPopEu project. This research project was supported by the Hellenic Foundation for Research and Innovation (EL.ID.E.K.) in the framework of the Action "1st Call for Research Projects EL.ID.E.K. for the support of faculty members and researchers and the supply of high value research equipment" (Project Number: 3572)