Journal of Artificial Intelligence Research 73 (2022) 633-672 Submitted 06/2021; published 02/2022

Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning

Tiziano Fagni TIZIANO.FAGNI@IIT.CNR.IT
Stefano Cresci STEFANO.CRESCI@IIT.CNR.IT
Institute of Informatics and Telematics (IIT)
National Research Council (CNR)
via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

Predicting the political leaning of social media users is an increasingly popular task, given its usefulness for electoral forecasts, opinion dynamics models and for studying the political dimension of polarization and disinformation.

Here, we propose a novel unsupervised technique for learning fine-grained political leaning from the textual content of social media posts. Our technique leverages a deep neural network for learning latent political ideologies in a representation learning task. Then, users are projected in a low-dimensional ideology space where they are subsequently clustered. The political leaning of a user is automatically derived from the cluster to which the user is assigned. We evaluated our technique in two challenging classification tasks and we compared it to baselines and other state-of-the-art approaches. Our technique obtains the best results among all unsupervised techniques, with *micro F1* = 0.426 in the 8-class task and *micro F1* = 0.772 in the 3-class task. Other than being interesting on their own, our results also pave the way for the development of new and better unsupervised approaches for the detection of fine-grained political leaning.

1. Introduction

Since the advent of Facebook and Twitter, politicians have had an increasing online presence in order to reach out to as many potential electors as possible. As of today, digital campaigning (including social media) has become mandatory, as people are massively consuming political content from social platforms[1]. Recently, 20% of interviewed social media users admitted to have changed their minds about a political issue because of something they read on social media[2]. Political activity on social media is also positively correlated to offline political activism (e.g., attending offline political events) (Vaccari et al., 2015). Politically-interested users are keen to know the stance of their friends, to read about candidates and campaigns, and to discuss pressing issues and election results (Grčar et al., 2017; Tucker et al., 2018). In spite of the relatively small readership of online platforms compared to that of traditional media (e.g., TVs, newspapers, and radio channels), the sociopolitical relevance of social media is still massive. In fact, second-order effects – typical of complex systems – allow for significant portions of the political social media content to be discussed also on traditional media, thus somehow still making it into the minds of people who don't even use social media at all (Benkler et al., 2017).

1. www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
2. www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Given this picture, it comes with little surprise that the task of learning the *political leaning of social media users* recently received a surge of attention. In literature, this task is also referred to as political *stance*, *ideology*, *polarity* or *alignment* prediction. Firstly, it represents a natural extension to the early efforts by social and political scientists at this task. In fact, ideology lies at the core of many theories in political science and has long been used to investigate individual behavior and preferences, governmental relations, and links between them (Bond & Messing, 2015). Traditional estimates are based on explicit preferences, such as roll-call votes, co-sponsorship records, and records of financial contributions to political campaigns. However, these data are typically available only for a few political figures (e.g., roll-call votes) or for a limited number of ordinary individuals, they are hard to acquire, and they are made available or updated infrequently. These limitations make fine-grained, continuous, large-scale analyses of political preferences challenging, if not outright infeasible. Conversely, social media represent a trove of both explicit and structured (e.g., likes and social relationships), as well as implicit and unstructured (e.g., text), data about the habits and preferences, including political ones, of millions of users. As such, many social and political scientists recently turned their attention to political analyses on social media – e.g., by estimating political leaning from social media data and by comparing such estimates with more traditional ones (Tucker et al., 2018). Meanwhile, also computer scientists found value in learning users political leaning, for a myriads of goals, such as: to forecast the outcome of elections (Tumasjan et al., 2010; Ahmed et al., 2016); to estimate accurate priors for models of opinion diffusion (Dandekar et al., 2013; Mäs & Flache, 2013); to measure and mitigate online polarization (Wong et al., 2016; Garimella et al., 2017; Nizzoli et al., 2021); to measure the effects of information operations, disinformation campaigns and propaganda (Nikolov et al., 2021; Tardelli et al., 2022; Cinelli et al., 2020; Ferrara et al., 2020); to explore the political dimension of bad actors, such as social bots and trolls (Hegelich & Janetzko, 2016; Rizoiu et al., 2018; Luceri et al., 2019; Yan et al., 2021; Cresci, 2020).

Existing approaches to the prediction of political leaning mainly focus on analyzing only the social or interaction networks (Garimella et al., 2016; Wong et al., 2016), or only the content of shared messages (Pla & Hurtado, 2014; Di Giovanni et al., 2018; Yan et al., 2019; Preoţiuc-Pietro et al., 2017), with few exceptions where content and networks are simultaneously considered (Lahoti et al., 2018; Aldayel & Magdy, 2019). Network-based approaches are grounded on the assumption that ideologically-similar users are likely to interact with, or to follow, each other. A first limitation arises when this assumption is violated – namely, in all those cases where like-minded users never interact, or in those equally-frequent cases where opposing users interact (e.g., to argue or to convince each other). There also exist users that do not follow others, or that follow a very limited number of accounts, which inevitably complicates network-based approaches. Notable examples of this kind are *@POTUS* in the US and media outlets/journalists that do not follow other accounts, for neutrality reasons, but that represent interesting subjects of political leaning analyses. Another limitation involves the large amounts of data needed for the analysis (e.g., the social or interaction graph), which are seldom promptly available. Content-based approaches are instead mainly limited by the intrinsic difficulty of processing natural language, and by the need for large corpora of manually-annotated messages and language-specific resources. Moreover, the majority of existing solutions adopt supervised approaches, which have been shown to lack

generalizability and to suffer from the limited availability of comprehensive and reliable ground-truth datasets (Cohen & Ruths, 2013).

1.1 Our Approach

Our goal in this work is that of developing an unsupervised content-based technique for predicting the political leaning of social media users. We will focus on two different tasks: (i) the prediction of the preferred political party of a user (fine-grained task), which in political science literature is typically referred to as party identification; and (ii) the prediction of its political pole (coarse-grained task). For bipolar systems, the latter task simply involves the prediction of left-right ideology, which for US data is typically measured in a continuous, one-dimensional space, with techniques such as the well-known DW-NOMINATE (Poole & Rosenthal, 1985). Notably, labels obtained for the two tasks represent different user traits and should not be equated or used interchangeably. For instance, the difference between the preferred party and the ideological position of a user in the left-right scale is straightforward when considering the shifts that parties exhibit between different elections (Busch, 2016). This is particularly true for the application and evaluation scenario of our work: the tripolar Italian political system (Pasquino, 2019). Nonetheless, we are interested in evaluating the efficacy of our proposed method in solving each of the two tasks separately.

In contrast with previous work, where political *ideology* and *leaning* were considered as synonyms, here we make an important distinction. By drawing upon definitions from the Oxford English Dictionary, we define *ideology*[3] as a latent set of concepts that forms the basis of a user's political preferences. Instead, we define *leaning*[4] as the practical political preferences of a user (e.g., its preferred party). Our approach for predicting political leaning, independently on the task (i.e., the desired fine or coarse prediction granularity), directly stems from the previous definitions. In fact, we first adopt an unsupervised approach to learn informative political representations of social media users. We then project users into a lower-dimensional space derived from their latent representations, which corresponds to the political ideology space. Finally, we leverage the topology of the political ideology space to infer the political leaning of each user. As such, our predicted leanings strictly depend on the latent ideologies learned for every user.

1.2 Contributions

Operationally, we propose a novel unsupervised solution for estimating the political leaning of social media users that is able to overcome the main limitations of previous approaches. Our method follows the scheme shown in Figure 1. Our solution initially leverages a deep neural network for learning latent users representations. Then, we feed these representations to a UMAP model in order to project and position users in a latent political ideology space. Finally, we leverage properties of the ideology space to infer the political leaning of every user, via clustering. We evaluate our proposed method and those used for comparisons on two challenging tasks. Specifically, we learn both fine-grained (i.e., party-level) and coarse-grained (i.e., pole-level) political leaning of Twitter users. Our solution achieves state-of-the-art results in both tasks, compared to existing unsupervised techniques. Specifically,

3. https://www.lexico.com/en/definition/ideology
4. https://www.lexico.com/en/definition/leaning

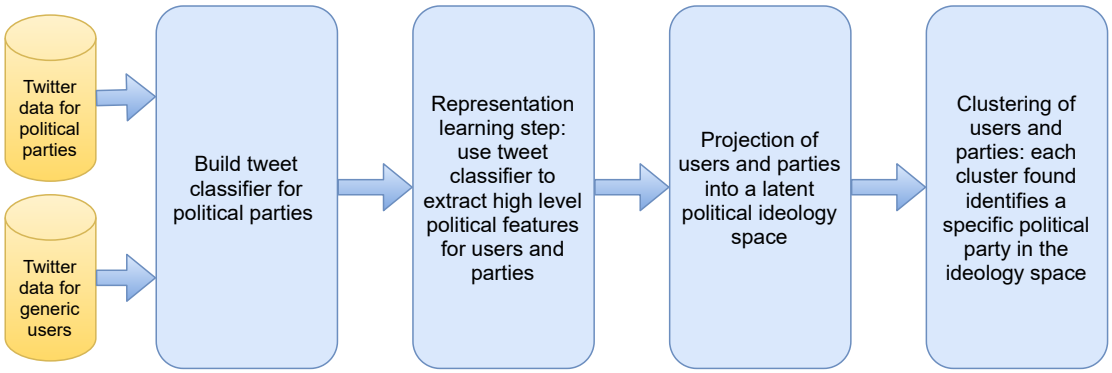


Figure 1: Outline of our proposal.

it achieves $F1 = 0.43$ and $F1 = 0.77$ when predicting fine- and coarse-grained leanings, whereas other unsupervised techniques and baselines achieve $F1 \leq 0.35$ and $F1 \leq 0.71$, respectively. Our technique is exclusively based on the textual content of user-generated social media posts. However, despite exploiting solely this noisy data source, it achieves performances that are comparable or even better than techniques that leverage cleaner signals (e.g., social relationships and interactions such as retweets and likes). This makes our technique particularly valuable since it obtains state-of-the-art performance without the need for gathering explicit user preferences or data-demanding network representations. In addition, by adopting an unsupervised deep learning approach, we are also language-independent, we avoid the need for manually-annotated corpora and linguistic resources, and we improve the generalizability of our results with respect to the traditional supervised approaches that are intrinsically limited by the availability of accurate and extensive ground-truth datasets (Cohen & Ruths, 2013; Cresci, 2020).

Our main contributions can be summarized as in the following:

- We provide a state-of-the-art unsupervised method for learning both fine-grained and coarse-grained political leaning of social media users.
- Our nuanced solution disentangles the sub-tasks of learning latent political ideologies from that of inferring political leanings, which were mixed and overlapping in previous works.
- We demonstrate the usefulness of unsupervised deep learning and projection with UMAP, to accurately position users within a latent ideology space.
- We show the profitability of leveraging the topology of the learned ideology space to infer political leaning via clustering.

1.3 Reproducibility

Our data are publicly available for scientific purposes[5].

5. https://doi.org/10.5281/zenodo.5793346

1.4 Roadmap

The remainder of the paper is organized as follows. In Section 2 we discuss previous works on the prediction of political leaning from social media. Then, before presenting our solution, in Section 3 we outline the political context in which our study is positioned and we provide details about our dataset. Section 4 describes our deep learning approach for learning latent political ideologies of social media users. In Section 5 we discuss our approach for positioning users in a latent political ideology space, and for inferring their political leaning. Experiments[6] and results are presented in Section 6, while Section 7 draws conclusions and highlights promising directions for future work.

2. Related Work

In this section we briefly survey extant literature for the prediction of political leaning. We split previous works based on the information used to make predictions.

2.1 Content-Based Approaches

Among the first approaches at this task are those solely based on the analysis of the textual content of messages. Pla and Hurtado (2014) investigated the use of sentiment analysis features. They trained a supervised classification model capable of labeling users based on their coarse-grained leaning – namely, as either left-leaning, right-leaning, center-leaning or undefined. Similarly, Di Giovanni et al. (2018) leveraged a set of linguistic syntactic features in a supervised classification task. The goal of their system was that of learning the political preference of Twitter accounts towards the 4 main parties in Italy. Despite focusing on fine-grained (i.e., party-level) predictions, Di Giovanni et al. worked with only 4 parties, instead of the 8 considered in our present work.

The previous works are representatives of a rather large body of work based on supervised content classification. Results obtained by these systems are however disputed by Cohen and Ruths (2013), since they tend to overestimate performances by focusing on politically active users (instead of *normal* or politically *inactive* users) and since their classification performances rapidly plummet when applied outside of the narrow range of examples used for training the systems. Similar results were also recently obtained by Yan et al. (2019), who evaluated the generalizability of text-based supervised systems for classifying partisanship and political ideology. Specifically, the authors built 3 datasets derived from the US Congressional Record, polarized media websites, and political wikis. Then, they trained a set of supervised classifiers on a dataset and they evaluated their performance in classifying texts from the other datasets. Among the supervised algorithms used for text classification are logistic regression as well as deep learning-based classifiers such as Marginalized Stacked Denoising Autoencoders and Semi-Supervised Recursive Autoencoders. Results show the difficulty of supervised and semi-supervised systems in generalizing from one dataset to another, thus motivating research and experimentation with unsupervised approaches.

6. Throughout the manuscript we use the term "experiment" with its conventional meaning in computer science – that is, an analysis, measurement, or evaluation campaign. This is different from its meaning in other disciplines (i.e., the social sciences) where experimental approaches involve treatments or interventions and are opposed to observational ones.

Another major drawback of supervised classification is that political leaning is typically provided as a discrete (e.g., binary) variable. A first improvement over these works was done by Preoţiuc-Pietro et al. (2017), who predicted the political orientation of Twitter users on a 7 point scale ranging from "very conservative" to "very liberal", with several points reserved for moderate users. They leveraged several features extracted from the tweets posted by the analyzed users, including features derived from LIWC, sentiment, topics, named entities and word2vec, and the prediction was performed with simple supervised classification algorithms (e.g., logistic regression). Conversely, more recent works moved towards unsupervised approaches. Kulshrestha et al. (2017) proposed a system where the leaning is obtained by measuring the similarity between the topic vectors of users, with those of known seed democrats and seed republicans. The political leaning was provided for each user in the $[0, 1]$ continuous range. This system is unsupervised, however it requires known sets of seed users, raising the question as to how to obtain such sets. Moreover, an additional challenge to face when developing systems for predicting continuous (rather than binary or crisp) leanings, is the lack of ground-truth values for training or evaluating the system.

2.2 Network-Based Approaches

Approaches purely focused on network characteristics currently represent only a minority of existing works. Barberá (2015) built a Bayesian spatial model of the Twitter social network that is based on homophilic network properties. The political leaning of each user is determined via Ideal Point Estimation. Similarly, Bond and Messing (2015) exploited user likes to Facebook pages to obtain estimates of political ideology for both parties, politicians, and ordinary users. Estimates are computed via Singular Value Decomposition (SVD) of an agreement matrix, which corresponds to a normalized adjacency matrix derived by projecting the bipartite matrix of user likes to parties onto the set of parties. Instead, Wong et al. (2016) computed political leaning by solving a convex optimization problem. By leveraging Twitter data, the objective function embeds signals derived from both the analysis of retweeting behaviors and features of the retweet networks. These previous works are unsupervised and provide leaning estimates in the $[0, 1]$ continuous range. Notably, these works, as well as all others that output one-dimensional scores, can only be applied to bipolar systems (e.g., to binary prediction tasks). This means that they are not suitable for application to the detection of fine-grained political leaning, a task that demands the prediction of multiple classes (i.e., the possible political parties), nor to the detection of coarse-grained political leaning in those systems that have more than two poles. An example of the latter is the current Italian political system (Pasquino, 2019), to which we apply our proposed methodology. The usefulness of the so-called left-right scale, operationalized as the $[-1, 1]$ or $[0, 1]$ continuous range, is also questioned by Bauer et al. (2017), who found that different individuals assign different meanings to the "left" and "right" concepts. As such, estimates based on a unique left-right scale for all individuals risk being biased and inaccurate. More broadly, Bauer et al. also raised the issue of self-reports, such as those obtained from survey respondents, as a ground-truth for training automated systems. In fact, many recent studies uncovered severe biases in self-reports, which motivates research on alternative means of obtaining ground-truth measurements (Bastick, 2021; Verbeij et al., 2021).

A recent interactions-based state-of-the-art unsupervised approach is presented by Darwish et al. (2020). Authors built user representations based on the users they retweeted. Then, they experimented with several projection and dimensionality reduction techniques, such as t-SNE and UMAP. Finally, they clustered projected users and labelled clusters via manual inspection. As a result of this process, each user is assigned to the label of the cluster to which it belongs. The system presented by Darwish et al. (2020) has been employed also for predicting the political bias of media outlets and famous public characters (Stefanov et al., 2019), and to estimate the polarization of Twitter users with respect to certain debated topics and political issues (Darwish, 2018).

The aforementioned work is the most similar existing solution with respect to our present contribution. However, contrarily to (Darwish et al., 2020), we do not explicitly exploit retweets between users, but we rather leverage the noisy textual content of their tweets. Consequently, a crucial component in our solution is the deep learning network used to learn latent user representations from tweets. In addition, we make different choices with respect to the techniques used for dimensionality reduction, projection and clustering. Finally, we automatically label clusters based on the labels of the pivots contained in each cluster, rather then with manual intervention. In our work, we also evaluate systems on a more challenging task than that tackled by Darwish et al. (2020) (e.g., binary classification), demonstrating and discussing the advantages of our solution.

2.3 Mixed Approaches

Another large body of work is based on a combination of content and network analysis. The advantage of simultaneously exploiting both textual content and network representations, such as those resulting from user interactions, was recently motivated and quantified by Aldayel and Magdy (2019). Specifically, they found that several different dimensions of online profiles and activities can provide useful signals to predict stance and leaning. Among them, some of the most informative signals can be extracted from user posts, user interactions with other users, websites visited, and user likes to other content on the platform.

Among the first works to jointly exploit content and interaction networks is (Conover et al., 2011). Authors exploited features derived from hashtags and from the retweet network, in a supervised binary classification task. Similarly, also Pennacchiotti and Popescu (2011) focused on supervised binary political classification. Their system is fed with features encompassing profile, tweeting behavior, linguistic, social, and interaction network information. Being based on supervised classification, both previous works still suffer from the limitations outlined by Cohen and Ruths (2013) and can only provide a dichotomic estimate of polarity. The work by Lahoti et al. (2018) instead provided interesting advances on this task. It is a state-of-the-art unsupervised framework based on non-negative matrix factorization, which learns a shared latent space between users and content. Similarly to other already-surveyed works, political leaning is considered as a one-dimensional continuous variable in the $[0, 1]$ range. However, the framework can be used to model more than one variable at a time (e.g., ideology, popularity), which represents an interesting improvement over previous works.

One of the limitations of network-based and mixed approaches is the need for explicit social relationships or user preferences (e.g., likes, retweets). In fact, it has been demonstrated

that extracting these information is a data- and time-demanding task and that such information is not always available (e.g., due to platform data-access restrictions) (Cresci et al., 2015). In turn, this decreases the applicability of such techniques and hinders large-scale social media analyses. Contrarily, our proposed technique achieves comparable or better performances while only exploiting the textual content of user posts, which are readily available.

3. Preliminaries and Data

This section provides preliminary information on the political landscape in which our analyses take place. Furthermore, it provides details on our dataset and its labeling.

3.1 The Italian Political Landscape

We focus our study on politically-active Italian Twitter users. Thus, our aim for this work is predicting the leaning of Italian Twitter users, within the current Italian political spectrum. Before delving into the details of our methodology, we first outline the Italian political landscape as of November 2020.

The last Italian general elections were held in March 2018, and resulted in the populist party Five-star Movement (M5S) winning the election with 32.7% votes, followed by the center-left Democratic Party (PD) with 18.7% votes and the far-right League (LE) party that obtained 17.4% votes. Despite receiving slightly more votes than LE, PD is considered one of the losers of the election, since it dropped from 40.8% votes received at the 2014 European elections, to 18.7% in 2018. The last major Italian party is the center-right Forward Italy (FI) that obtained 14.0% votes. Based on this outcome, a coalition government was formed in May 2018 by M5S and LE. This lasted until August 2019, when a government crisis initiated by LE led to the formation of a new coalition government in early September. This government, which is still in charge at the time of writing, is led by M5S and PD, together with other minor parties[7]. The peculiarity of the current Italian political landscape is represented by the populist and anti-establishment M5S, whose members refuse to position in the traditional left-right bipolar paradigm since they regard M5S as a *non-party*[8]. As a consequence, the coarse-grained Italian political landscape is a tripolar system consisting of right-leaning parties, left-leaning parties, and the M5S (Pasquino, 2019). Notably, carrying out predictions of political leaning in a tripolar system has implications on the techniques used for the analysis, since some of the existing ones have been specifically designed for bipolar systems (e.g., left *vs* right, liberals *vs* conservatives, in favor *vs* against a given topic).

In addition to the aforementioned parties, in this study we also consider 4 minor parties that together accounted for 8% votes in the 2018 general elections, thus covering the whole extent of the Italian political spectrum and including both major and minor parties. Table 1 summarizes the main information, name and color conventions for all considered parties and their leaders. Henceforth, we refer to the party Twitter accounts as our *pivots*, since they

7. https://en.wikipedia.org/wiki/Conte_II_Cabinet
8. https://en.wikipedia.org/wiki/Five_Star_Movement

leaning	party name	party handle	leader handle	label	color	#users
RIGHT	CasaPound Italy	*@casapounditalia*	*@distefanoTW*	CPI	●	2,997
	Brothers of Italy	*@FratellidItaIia*	*@GiorgiaMeloni*	FdI	●	2,507
	League	*@legasalvini*	*@matteosalvinimi*	LE	●	2,705
	Forward Italy	*@forza_italia*	*@berlusconi*	FI	●	746
M5S	Five-star Movement	*@Mov5Stelle*	*@luigidimaio*	M5S	●	3,206
LEFT	Democratic Party	*@pdnetwork*	*@nzingaretti*	PD	●	2,377
	+Europe	*@piu_europa*	*@bendellavedova*	+E	●	4,335
	Communist Ref.	*@direzioneprc*	*@maurizioacerbo*	PRC	●	1,326

Table 1: Information about the 8 Italian parties, and their leaders, considered in this study. Rows are grouped according to the coarse-grained political leaning, representing the tripolar Italian political system.

play an important role in the estimation of political leaning. Notably, the only preliminary data needed by our framework are (i) the pivots, and (ii) their coarse-grained leaning.

3.2 Twitter Dataset

Our aim for this work is to develop a framework for estimating political leaning in an unsupervised fashion (i.e., with no manual labeling involved). To combine the strengths of labeled datasets (e.g., rich, high-quality data) with those of unsupervised approaches (e.g., generalizability, no bias or errors due to manual labeling), our desiderata is to acquire a dataset that is *implicitly* labeled, with respect to political alignment. We met our desiderata by leveraging favorited (i.e., liked) tweets, and by considering political likes as proxies for political leaning. Other options, also adopted in some previous works, could have involved the exploitation of retweets or follower relationships to political parties. However, we consider likes to be stronger indicators of political preference (Aldayel & Magdy, 2019).

Operationally, we first crawled the Twitter timelines of our pivots. Then, for each collected tweet, we obtained a list of users that liked that tweet. At the end of this process we obtained a bipartite graph linking 20,199 users to our 8 considered parties, based on explicit user likes to party tweets. The number of users that liked at least one party tweet is reported in the last column of Table 1, for every party. We completed our data collection by crawling the most recent 200 tweets from the timelines of all 20,199 users, which resulted in more than 3.6M tweets, in total. When building the dataset, we only included users whose timeline contained at least 25 tweets. For each user, we collected at most up to 200 tweets. This data collection process roughly covered the months of August to early October 2019. On average, user timelines include 179.3 tweets, evenly distributed during our data collection period. Finally, we performed a stratified sampling to split our dataset into a training (90% – 18,169 users), a validation (3% – 604 users) and a test (7% – 1,426 users) partition. As a result of our splitting strategy, the distributions of parties and poles across the 3 data partitions are comparable.

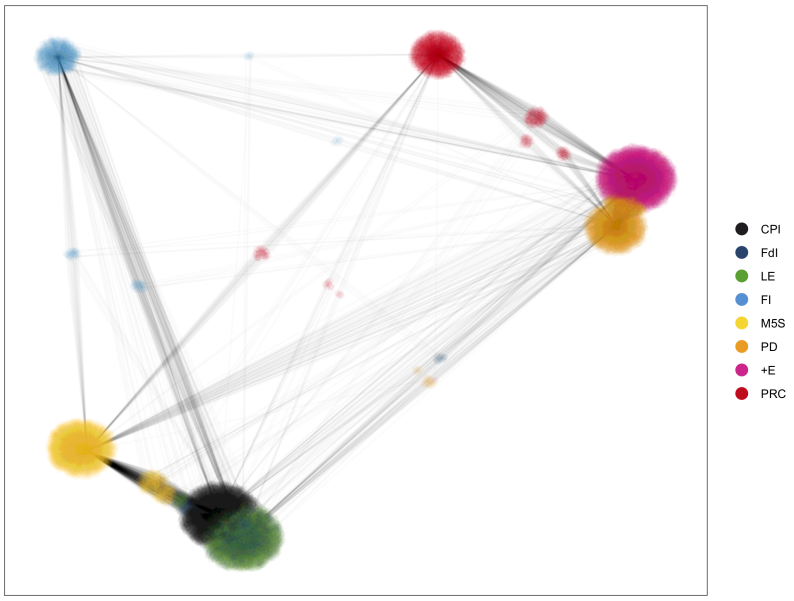


Figure 2: Louvain clustering of the weighted user-similarity network. Edge weights are based on user likes to party tweets. Clusters are color-coded and each cluster is associated to a political party. User labels resulting from this clustering are used as ground-truth for evaluating predictions of user political leaning.

3.3 Ground-Truth Labeling

Since we do not know the preferred party of the users in our dataset, we obtain a ground truth for our task by leveraging user likes to party tweets. Specifically, we first build the bipartite graph of users and party tweets, where links between nodes represent user likes to party tweets. Next, we project the bipartite graph onto the subset of user nodes, obtaining a weighted, undirected user-similarity network. Links in this network represent similarity between users. In order to build this network and to compute the similarity between users, we adopt a simple weighting scheme based on the frequency of common associations in the bipartite graph. In other words, the similarity between two users is measured as the number of tweets liked by both users. Finally, we cluster users in this network with the Louvain community detection algorithm (Blondel et al., 2008). Each user is then labeled with the political party corresponding to the cluster it belongs to. Figure 2 shows the clustered user-similarity network derived from our dataset. Clusters are color-coded and determine the ground-truth label for each user. As shown, in this representation user clusters are sharply defined. The vast majority of users only has edges connecting to other users of the same cluster, with only a few edges connecting users across different clusters. In turn, this implies that user likes to party tweets are a very strong signal of political alignment. In the following, we describe our approach for the challenging task of inferring user political leaning from tweets, which represent a much more noisy signal than likes.

4. Learning Latent Political Ideologies

Determining users political leaning from the analysis of the content posted on social media is a challenging task. One challenge stems from the need to find a clever way to focus the analysis on politically-relevant content only. Indeed, a typical user's timeline is filled with posts related to several different topics (e.g. sport, spare time, work, politics) that embrace all aspects of the user's life. The first difficult step is therefore related to splitting the relevant contents (i.e., those related to politics) from the rest of the messages that, within this context, simply represent noisy and unhelpful data for inferring users political leaning. A second critical aspect, given a post covering political topics, is to properly measure how such message is politically close to the typical ideology of a specific party or pole. This measure can predict how much the post is in agreement toward a specific ideology, so having access to enough messages in a user's timeline that convey this information can contribute to accurately estimate its final political leaning.

We sorted out both these critical issues by proposing a novel unsupervised process organized in seven high-level steps, shown in Figure 3. In step 1 we build an automatic tweet classifier for assessing if a tweet has been produced by a certain political party. Details on how the classifier is trained are given in Section 4.1. In steps 2, 3, and 4 we leverage the classifier to compute vector representations for users and parties. We exploit representations of users and parties to identify a subset of users that are particularly similar to the considered parties. These steps of our methodology are described in Section 4.2. In step 5 we automatically analyze the tweets from the subset of users whose representations are similar to those of the parties. In particular, for each of such users we select a subset of his/her tweets that conveys explicit political opinions. In step 6 we use these user-generated tweets as additional training examples in a second training phase of our tweet classifier, since they represent political tweets with different characteristics than those already seen by the classifier (i.e., those obtained from the official party accounts rather than from ordinary users). The final classifier is used in step 7 to compute the final vectors of users and parties. Each computed vector corresponds to the latent representation of a user. In other words, learned vectors allow to position users in a shared latent political ideology space. Steps 5, 6, and 7 are described in Section 4.3. Finally, in Section 5 we describe how we leverage the relative positions of users in the latent political ideology space to infer the preferred party for each user.

4.1 Predicting the Political Relevance of a Tweet

In this step we are interested in measuring the degree of agreement of a relevant political tweet with the typical political ideology of each party involved in this study. Before describing how this step works, it is helpful to define what a relevant tweet is. We deem a tweet to be *politically relevant* if it expresses a subjective opinion in favour or against a specific party, a party leader, a specific person, or a political position ideologically known to be near to a certain party. In this context, possible examples of relevant tweets are unquoted retweets of tweets posted by political parties or leaders, unquoted retweets of messages of other users where they express a political opinion on something, tweets replaying to political leaders where a user shows its appreciation or a negative attitude toward the author of message.

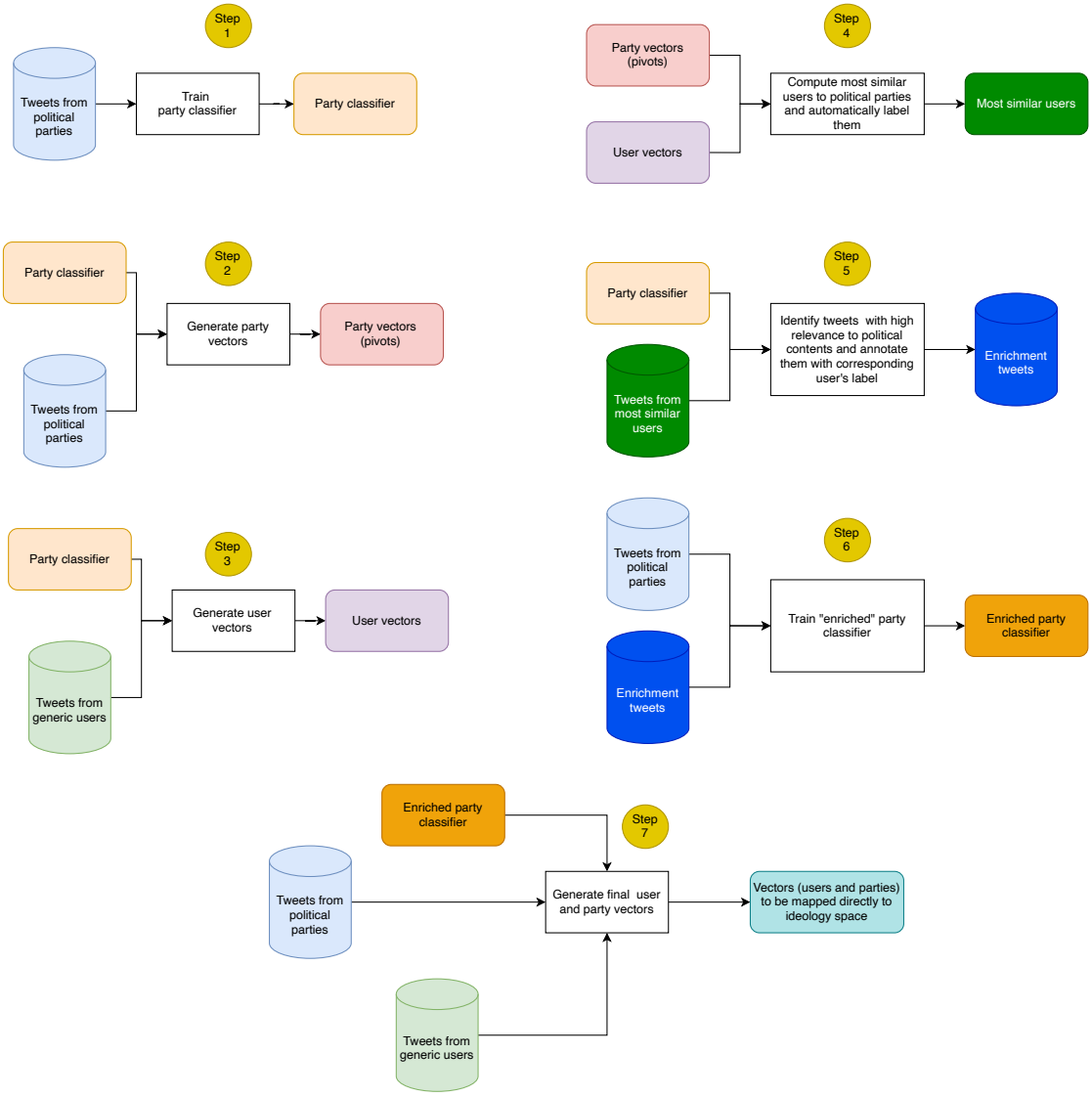


Figure 3: High-level overview of the proposed unsupervised strategy to map users to a latent political ideology space.

Given a tweet, it is possible to quickly determine if the text is politically relevant by leveraging an automatic multiclass classifier which has learned, from examples of political tweets, to predict if a tweet has been produced by a specific political party. Indeed, such classifier should be able to assign not only proper labels (i.e., the most probable party that could have produced that tweet) but also to estimate a confidence in its decision which can be seen as a "relevance level" of the tweet with respect to political topics[9]. Such type of solution normally requires a manual annotated dataset. Here, we obtained the same result in

9. The higher is the confidence score of the classifier, the higher is the probability that the tweet content is expressing something which is politically relevant.

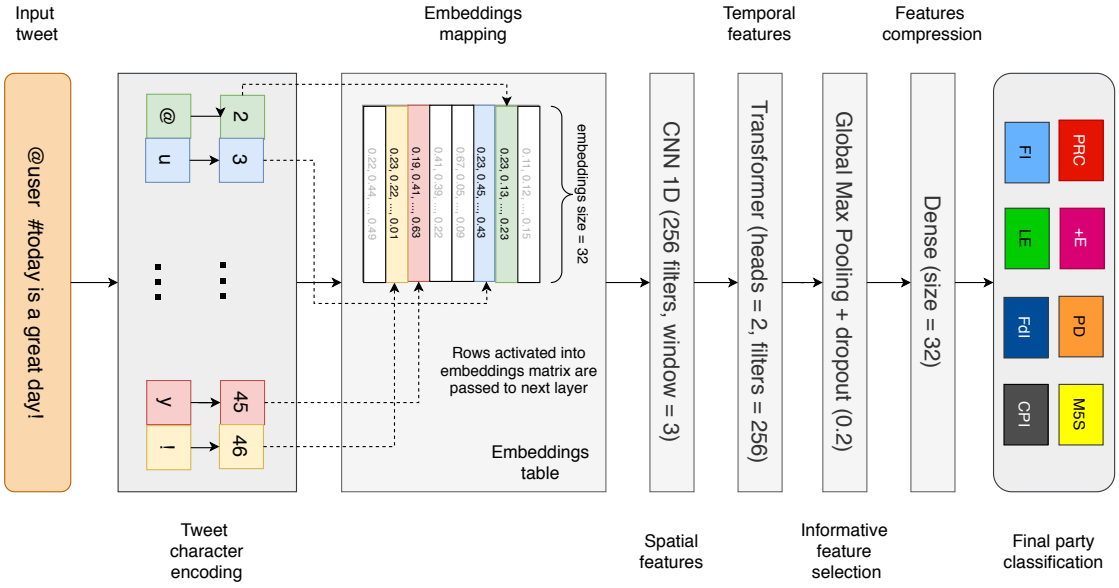


Figure 4: Neural network architecture of our party classifier.

an unsupervised way by exploiting the implicit relationship between tweets and the Twitter accounts that have produced them. In particular, we focus on the official accounts of the 8 considered parties and we use their timelines to automatically build a labeled dataset. In this way, each tweet posted by party account P is labeled as generated by political party P and the problem we solve is the prediction of the party that produced a tweet only based on the textual content of the tweet itself. Our approach thus resembles labeling schemes by *distant supervision* (Marchetti-Bowick & Chambers, 2012). Building our dataset by focusing on party accounts also has two important practical implications that simplify solving our task. The first implication is that we are certain that the labels assigned to tweets are correct. This allows us to train a classifier on a real gold-standard thus avoiding sub-optimal solutions caused by biases and labeling errors introduced during error-prone manual labeling operations (Misra et al., 2016; Pandey et al., 2019). The second and most important implication is that each considered timeline is "clean" (i.e., not noisy) and contains only politically-relevant tweets[10]. Such politically-relevant tweets are typically in favour of the party, of its leader, or of some action proposed, and only seldom against another political competitor.

By following the strategy described above, we built a dataset composed by all the tweets (in the form of original contents or retweets) posted in the timeline of the considered 8 political parties. For each party we selected the most recent 3,000 tweets. At the end of this process we obtained an almost balanced dataset composed of 23,791 labeled tweets. By leveraging these data, we built the political party classifier using the neural network architecture shown in Figure 4. We used a character-based encoding to obtain an initial vector for each tweet. Our method uses an embeddings character table which is learned during the

10. It is extremely rare that an official account of a political party posts something which is not related to politics.

training phase. Each tweet vector is thus mapped into a new vector using the embeddings table and next passed to a CNN layer (LeCun et al., 1998) with the aim of extracting spatial features – i.e., those that are invariant to the locations where they occur. This set of features is then processed by a transformer layer (Vaswani et al., 2017) that extracts the most informative temporal recurrent patterns from the data. The gathered information are thus filtered and compressed before being used to produce the final classification of the tweet.

4.2 Extracting a Politically Relevant User Vector

The party classifier built in the previous section can be employed to extract high level features that express the political attitude of a user with respect to all parties. In particular, by processing the entire timeline of a user with such classifier we can identify which tweets are politically relevant (i.e., tweets classified with medium/high confidence scores) and which political parties the user's opinion aligns with. More formally, let us define $T_U = \left\{t_i^U \mid i = 1, ..., min\left(200, |T_U|\right)\right\}$ as the timeline of user U where t_i is its i-th most recent tweet and $|T_U|$ is the number of tweets available in the whole timeline of U. Let us also declare $P = \{\mathsf{PRC}, \mathsf{+E}, \mathsf{PD}, \mathsf{M5S}, \mathsf{FI}, \mathsf{LE}, \mathsf{FdI}, \mathsf{CPI}\}$ as the set of the 8 considered parties such that $P_1 = \mathsf{PRC}$, $P_2 = \mathsf{+E}$, and so on. We define the party classifier as the function C mapping a tweet t to a score vector as in the following:

$$C : t \in \mathbb{R}^{280} \rightarrow [s_{P_1}, s_{P_2}, \ldots, s_{P_8}] \in \mathbb{R}^8 \tag{1}$$

where $s_{P_i} \in [0, 1]$ is the score assigned by classifier C to the tweet t for party P_i. Given the timeline of user U, we can compute $S_{U,i,k}$ as the set of the best k scores obtained for a specific party P_i, as in the following:

$$S_{U,i,k} = \{max_k \{C\left(t\right)_i\} \mid t \in T_U\} \in \mathbb{R}^k \tag{2}$$

Given the previous definitions, we can finally define how to extract a politically relevant user vector:

$$V_U^k = [S_{U,1,k}, S_{U,2,k}, \ldots, S_{U,8,k}] \in \mathbb{R}^{8k} \tag{3}$$

The vector V_U^k is a concatenation of the best tweet scores measured on the relevance to each party, which is indicative of the interests and the leaning shown by the user toward a specific political ideology. In this work, we fixed $k = 5$ in Equation (3) based on early experimentation demonstrating this value to yield reliable measures of the degree of interest shown by a user for a specific political party. Indeed on the one hand, a larger k would require the user to post a lot of political content in order not to penalize excessively the weight of a specific political stance. On the other hand, a smaller k would require to have an extremely accurate party classifier.

4.3 Unsupervised Data Enrichment to Improve Tweet Party Classification

As for all supervised classifiers, the party classifier built in Section 4.1 works well when analyzing tweets whose writing style is similar to that of tweets used in the training dataset. In particular, using party accounts as positive seeds in the dataset construction phase, poses some limitations to the learned classifier for correctly handling the true tweet distribution.

Indeed, official party tweets are typically written in a clean, formal and institutional language. In addition and as previously anticipated, they also typically provide facts in support of the work of the party or of its political leader. On the contrary, political tweets from average users have different linguistic characteristics. Their writing style is informal and tweets contain abbreviations, slang, and jargon expressions. Regarding the opinions conveyed in a typical user tweet, sometimes users support a political party or leader. However, oftentimes users also express strong disagreement toward an opposing political opinion or politician. In particular, a considerable set of users tend to provide more destructive opinions (e.g., harsh comments against someone or something) than constructive ones (Nizzoli et al., 2021). In a few edge cases, the political opinions expressed in a user's timeline are exclusively against something or someone. Because of this, it is important to transfer such nuances to the party tweet classifier during its training phase, in order to be able to infer accurate user political ideologies.

Given these motivations, here we propose an unsupervised strategy to enrich original training data with labeled tweet examples coming from all types of users. This enrichment process is aimed at providing also negative tweet examples to the tweet party classifier, in addition to the positive tweets from the official parties, and can be summarized in the following steps:

1. Obtain the vector representations for the pivot (i.e., party) accounts. This can be achieved with Equation (3).

2. Select those training-set users that are most similar to each party account.

   2.1. For each training-set user, we obtain its vector representation with Equation (3) and we compute its cosine similarity with respect to the vector of each party.

   2.2. For each party, we sort users based on their similarity and we select users laying above the 99-th percentile of the similarity distribution (i.e., the most similar ones). We automatically assign the label of the party to each user matching this condition.

3. Select tweets to be used as an enrichment for training the tweet party classifier. To reach this goal, we analyze the timelines of the users selected at the previous step. For each selected user, we use the party classifier C to predict the political relevance of all the tweets in the user's timeline. Then, we retain only those tweets for which C yielded a score $s_{P_i} \geq Th$ for at least one party P_i. For large values of the threshold Th, this results in selecting only those tweets for which our classifier provided strong predictions. Such tweets are used as enrichment tweets in a second training phase of the classifier. The ground-truth label assigned to those tweets is that of its author, assigned at step 2.1 of this procedure. Notably, this label that we inferred in an unsupervised fashion is likely to be correct since we are considering users that are very similar to a given party in the political ideology space. Overall, this process allows to expand the training set by ingesting tweets from average users in addition to those of official party accounts, while still retaining a high confidence of the new tweets' labels.

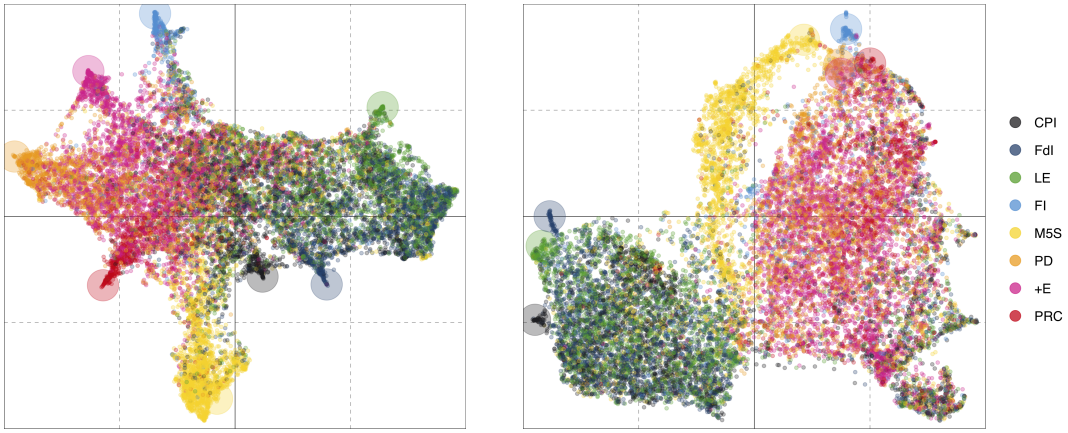
4. Build a new party classifier C' using both the original training dataset and the enrichment data, using the same architecture shown in Figure 4. As a consequence of the

	original tweet	**translated tweet**	L_c	L_o
√	@Mov5Stelle Invece di tagliare la rappresentanza, bastava dimezzare gli stipendi. Una legge ordinaria, sicuramente più veloce come iter di una legge costituzionale.	@Mov5Stelle Instead of cutting the representation, it was enough to halve the salaries. An ordinary law, a procedure certainly faster than a constitutional law.	PRC	M5S
√	Salvini chiede I pieni poteri (!!!), sappiatelo...	Salvini asks for full powers (!!!), you should be aware of this...	+E	LE
√	RT @gennaromigliore: A Genova la polizia rompe le ossa al cronista #StefanoOrigone e protegge le canaglie di #CasaPound: bisogna sanzionare...	RT @gennaromigliore: In Genoa the police breaks the bones of the reporter #StefanoOrigone and protects the villains of #CasaPound: you have to punish...	PD	CPI
√	#CasaPound Si arriverà davvero allo sgombero dei "fascisti del Terzo Millennio" dal palazzo di 6 piani che hanno occupato abusivamente da 15 anni nel centro di Roma?	#CasaPound Will there really be the evacuation of the "fascists of the Third Millennium" from the 6-storey building they have squatted in the center of Rome for 15 years?	M5S	CPI
√	@dariofrance @nzingaretti Sarebbe un governo peggio di questo. Peggio del #pd non ci sta niente in circolazione!	@dariofrance @nzingaretti It would be a worse government than this. There is nothing around that is worse than #pd!	FI	PD
√	Un Altro PD Idiota contro Salvini ciabattoni	Another PD-idiot against Salvini moron	LE	LE
√	Ennesimo strafalcione geografico per il #M5S: questa volta il vento del cambiamento sposta addirittura le regioni! Speriamo i fondi arrivino davvero in #Molise, senza nulla togliere alle #Marche..	Yet another geographical blunder for the #M5S: this time the wind of change even moves the regions! We hope the funds really arrive in #Molise, without detracting anything from the #Marche..	FdI	M5S
√	@virginiaraggi La compatisco. Fra un paio d'anni, allo scadere del suo mandato, lei finirà nell'oblio come merita. CasaPound sarà sempre al suo posto.	@virginiaraggi I sympathize with you. In a couple of years, at the end of your term, you will be forgotten as you deserve. CasaPound will always be in its place.	CPI	CPI
⊘	Le chiacchiere fanno i pidocchi, i maccheroni riempiono la pancia	The chatter makes the lice, the macaroni fill the belly	FI	CPI
⊘	In bocca al lupo alle ragazze e ai ragazzi della mia commissione. Domattina si parte #notteprimadegliesami	Good luck to the girls and boys of my commission. Tomorrow morning we begin #nightbeforeexams	PD	CPI

Table 2: Examples of tweets included in the enriched training-set of the party tweet classifier. Politically-relevant tweets are marked with √, irrelevant ones are marked with ⊘. The label initially assigned by C to each example is reported in column L_o, while the label corrected through the usage of the minimal distance from a party account is reported in the L_c column.

enriched training-set, the new classifier C' is more accurate than C, especially with regards to informal and negative political tweets.

In our preliminary experiments, we found that $Th = 0.5$ is a reasonable value to get both several thousands of new labeled examples and a large variety of new tweets featuring substantial differences in their writing style, with respect to the typical tweets of an official party account. In detail, by applying the aforementioned process: (i) in step 2 we selected 1,462 users equally distributed among all parties, and (ii) at the end of step 3 we obtained an enrichment dataset composed of 8,753 new labeled tweets. As shown in Table 2, this method is obviously not perfect but it ensures to obtain a plentiful variety of different examples that



(a) Our technique. (b) Word2vec.

Figure 5: UMAP projections of the latent political ideology spaces learned by our proposed technique and via word2vec. Colors encode ground-truth party labels. Larger circles highlight the position of the official accounts for each considered party (i.e., our pivots).

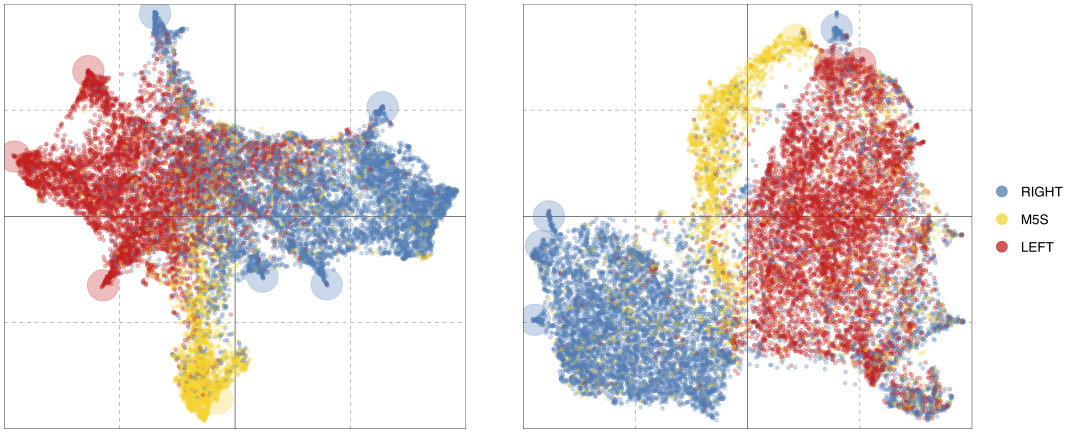
help to improve the precision and the generalizability of the enriched party classifier C', in an unsupervised fashion.

5. Predicting Political Leaning

By using the encoding scheme presented in Section 4.2, we are able to analyze the proposed method from a qualitative point of view and to map each user into a position within a shared latent political ideology space. This mapping is built directly onto user vectors with the aim of projecting users over a bidimensional geometric space in such a way to (i) minimize the distance between similar users having close political ideas and (ii) maximize the distance between users having different political opinions. To perform feature reduction and to map the latent user vectors in $\mathbb{R}^{40}$ to an equivalent space in $\mathbb{R}^2$, we leveraged UMAP with default parameters (McInnes et al., 2018). The mapping obtained from training data[11] is shown in Figures 5a and 6a, where users are respectively colored according to their party and pole labels.

Regarding party projections, the first observation is that many users supporting a specific party are concentrated in the neighborhood of the party itself (indicated by large circle points). This is a quite strong indication that user feature representations provided by Equation (3) properly describe the political stance of the parties. In addition, when users have enough political content in their timeline, the same method also allows to position them near to their preferred party. This first result is verified for all parties, and particularly so for the left-leaning ones and for the M5S. Regarding right-leaning parties (right-hand side of the figures), although this trend is confirmed, the situation is more fluid with users of FI clearly separated from the users of the other 3 right-leaning parties (CPI, FdI and LE).

11. Here, to better highlight data distribution in the political ideological space, we used training data because the amount of users is far bigger than those contained in test data and the distribution of points is practically the same (i.e., there is no drift between training and test data).



(a) Our technique. (b) Word2vec.

Figure 6: UMAP projections of the latent political ideology spaces learned by our proposed technique and via word2vec. Colors encode ground-truth pole labels (right-leaning, left-leaning, and M5S). Larger circles highlight the position of the official accounts for each considered party (i.e., our pivots).

Indeed, supporters of the latter parties, in addition to forming clear clusters positioned around official party accounts, are spread over wide areas of the political ideology space, also creating regions where users of different parties are mixed together. This feature of our learned political ideology space is in agreement with the Italian political landscape, where these 3 far-right parties hold similar stances with respect to many political issues (Pasquino, 2019), and with the opinions expressed by their electors. By analyzing Figure 5a, it is also worth noting that even the central area of the ideology space contains a mixture of users belonging to different parties. Also this situation is expected and understandable, since it represents undecided users and users that hardly share any political content at all.

When considering pole projections shown in Figure 6a, we can see that there is a clear separation between the three poles, with only the central region of the ideology space characterized by a physiological group of users whose political stance is not uniform, for the same motivations given before. Quite naturally, these qualitative results suggest that the fine-grained task (i.e., party prediction) represents a much more challenging problem than the coarse-grained one (i.e., pole prediction). This naturally results from the minimal differences between some of the considered parties.

For comparative purposes, in Figures 5b and 6b we show the latent political ideology space obtained with a user encoding based on word2vec (Mikolov et al., 2013) instead of the one learned with our method. Word2vec is a very popular embeddings method that already demonstrated an excellent encoding power on many NLP tasks. In this case we used word2vec algorithm provided by gensim library[12] to build from scratch a custom model optimized for this specific context by learning the latent space directly from the used Twitter dataset. In order to build the new word2vec model, we applied a minimal step of preprocessing to raw textual data. In particular, we transformed texts into lowercase and removed

12. `https://radimrehurek.com/gensim/`

all stopwords. With the resulting data, we built a word2vec model keeping only the most frequent 50,000 words. Each tweet is thus vectorized by computing the mean of the sum of vector embeddings of each word occurring in the text. The user vector is finally obtained by computing the mean of the tweet vectors extracted from the user's timeline. Differently from our method, the word2vec encoding seems unable to clearly separate the different political parties, as demonstrated by several regions of the ideology space featuring a mixture of users from different parties. Another major drawback of this approach is represented by the vicinity between the accounts of several different parties. While in Figure 5a each pivot held a specific position in the ideology space, clearly separated from that of other parties, in Figure 5b several pivots end up laying in the same area of the ideology space, which inevitably hinders party separability and the prediction of users' political leaning. Regarding pole predictions, the situation improves. However, it does not reach the level of data separation obtained with our proposed technique. In summary, these findings suggest that word2vec encoding, in this particular context, is sub-optimal and not able to properly model the semantics of political ideologies of the different parties.

Based on the favorable properties of our learned latent political ideology space, the unsupervised prediction of user political leaning can be achieved by applying a clustering algorithm directly to the bidimensional projected user vectors. Without loss of generality, in this work we assume that we know the number of clusters we want to obtain at the end of clustering process (i.e., 8 clusters for the party prediction task and 3 clusters for pole prediction task). The steps needed by the clustering process are summarized the following:

1. Projection of the users into a new bigger latent space based on the similarity of users. Each distinct user is seen as a separated feature in this new space and the feature vector of each user is generated by computing its pairwise distance to all the other users. This step was originally proposed by Darwish et al. (2020) and demonstrated to improve the subsequent clustering step.

2. Feature reduction using UMAP to prevent the curse-of-dimensionality due to data sparseness (Domingos, 2012).

3. Feature standardization by subtracting the mean and scaling the features to unit variance.

4. Data clustering using the KMeans, GaussianMixture, or MeanShift algorithms. This clustering step is based on the implementations provided by the *sklearn* Python software package[13].

The first 3 steps of the above list are optional and can be used only in specific cases where they improve clustering accuracy. For the experiments reported in the next section, we followed the approach used by both Darwish et al. (2020) and Di Giovanni et al. (2018), and we evaluated different configurations on the validation partition of the dataset. Then, we used the best configuration obtained on the validation set to label users of the test set. The details about the specific configurations that we used are given in Section 6.1.2.

13. `https://scikit-learn.org/stable/`

6. Experiments and Results

Given the previously described method for predicting the political leaning of social media users, in this section we evaluate its predictions for test-set users of our dataset, with respect to the ground truth labels and to the predictions yielded by a number of baselines and other state-of-the-art techniques. Furthermore, we also validate our method by applying it to predict the political leaning of the Italian members of the European parliament (MEPs). Finally, we carry out a set of additional experiments to assess the sensitivity of our method to a number of relevant factors (e.g., distance from the pivots in the latent ideology space, number of tweets, number of retweets) and we provide a thorough analysis and discussion of our classification errors.

6.1 Experimental Settings

This section provides details on the evaluation settings and on the techniques used in our performance comparisons.

6.1.1 Evaluation

All techniques considered in our experiments are evaluated on two tasks. The aim of the first task is the prediction of fine-grained political leaning, which concerns with associating each user to its preferred political party. The second – simpler – task is the prediction of coarse-grained political leaning, where each user is assigned to a political pole (e.g., left-leaning, right-leaning, or leaning towards M5S). Similarly to previous work, we evaluate each task as a multiclass classification task. However, our experiments are considerably more challenging than those typically performed in previous works, due the larger number of involved classes. In fact, previous methods for predicting political leaning were typically evaluated in a binary classification setting (e.g., predicting left- *vs* right-leaning users). Here instead, our fine-grained task encompasses 8 possible classes, while our coarse-grained task admits 3 classes. Given the moderate class imbalance for both the fine- and coarse-grained tasks, visible in Table 1, for each evaluated method we report both the *micro* and *macro* versions of *precision*, *recall* and *F1-measure*.

6.1.2 Comparisons

In the upcoming sections, we report experimental results for several techniques, including different configurations of our present proposal, strong and weak baselines, and other state-of-the-art techniques. In the following, we briefly introduce each technique that we implemented and evaluated, starting from 3 interesting configurations of our proposal. Wherever meaningful, each technique is labeled by separately specifying the approach used for learning ideologies and that used for making predictions (i.e., ideologies + predictions).

Parties + clustering: This method is based on the latent ideologies learned with our proposed unsupervised approach. For learning ideologies, we apply only the steps described in Sections 4.1 and 4.2, without the enrichment step introduced in Section 4.3. Predictions are performed via clustering, as detailed in Section 5, by applying step 1, step 2 with a feature reduction to 64 features, and step 4 using GaussianMixture with default parameters. These clustering settings are used for both prediction tasks.

Parties enriched + distance: For this method we use unsupervised ideologies learned with all 3 steps described in Section 4, thus also including the enrichment step. Predictions are obtained by assigning each user to the party of the pivot the user is nearest to. This method represents a strong unsupervised baseline that leverages our learned ideologies, combined with a simple prediction strategy.

Parties enriched + clustering: This is our most complete method. It is fully unsupervised, it leverages all steps for learning latent ideologies as well as clustering for obtaining predictions. The clustering process for the party prediction task is performed by applying only step 3 and step 4 using KMeans as the clustering algorithm. For the pole prediction task, we used instead step 1, step 2 with a feature reduction to 64 features, and step 4 using the KMeans algorithm.

In addition to our 3 unsupervised contributions reported above, we also experimented with a number of baselines and other approaches, which we briefly describe in the following.

Random classifier: Simple unsupervised baseline that outputs random predictions.

Majority classifier: Simple supervised baseline that always outputs the majority class.

Word2vec + clustering: Unsupervised method where latent ideologies are learned by leveraging word2vec embeddings, while predictions on both tasks are obtained via clustering applying only step 4 with the KMeans algorithm.

Retweets + clustering: This method implements the state-of-the-art unsupervised technique proposed by Darwish et al. (2020). It first learns user representations based on user retweets, then it obtains predictions on both tasks via clustering by applying step 1 and step 4 implemented with the MeanShift clustering algorithm.

Supervised enriched + clustering: This method is similar to the parties enriched + clustering one, with the exception of how the enrichment step is carried out. To this end, this method feeds back into the tweet classifier those tweets for which the classifier outputted wrong predictions despite having a high confidence. Given that this method exploits ground-truth labels for the enrichment step and clustering for obtaining predictions, it is classified as semi-supervised. The clustering process is performed by applying step 1, step 2 and step 4. On the party prediction task, we reduced features to 64 dimensions and we used KMeans as the clustering algorithm. Differently, on the pole prediction task, we reduced features to 2 dimensions and we applied the GaussianMixture algorithm with default parameters.

Parties enriched + SVC: This overall semi-supervised method leverages all our proposed steps for learning unsupervised latent ideologies, including enrichment. Then, predictions are obtained by training a supervised SVC classifier.

Supervised enriched + SVC: Here we exploit our supervised enriched ideologies in conjunction with an SVC classifier. Both steps used in this method are thus supervised.

Word2vec + SVC: Ideologies used in this semi-supervised method are obtained via word2vec embeddings, while predictions are yielded by an SVC classifier.

6.2 Results

In the remainder of this section we present and discuss experimental results for the 2 considered tasks: coarse- and fine-grained prediction of political leaning. We first compare results of our method to those of the several others that we evaluated. While discussing

	method		macro			micro		
	ideologies	*predictions*	*precision*	*recall*	*F1*	*precision*	*recall*	*F1*
	–	random classifier	0.124	0.125	0.120	0.143	0.126	0.131
	word2vec	clustering	0.128	0.111	0.106	0.171	0.139	0.142
‡	retweets	clustering	0.370	0.293	0.301	0.420	0.346	0.354
	our contributions							
	parties	clustering	0.390	0.344	0.342	0.443	0.354	0.372
	parties enriched	distance	0.419	0.370	0.324	0.489	0.339	0.330
	parties enriched	clustering	**0.472**	**0.434**	**0.426**	**0.517**	**0.421**	**0.426**

‡: (Darwish et al., 2020)

Table 3: Performance comparison of *unsupervised* methods for fine-grained (party) prediction of political leaning. The best result for each evaluation metric is shown in bold font.

such comparisons, we particularly focus on the differences in performance between our best proposal and the technique introduced by Darwish et al. (2020), which is considered the state-of-the-art for unsupervised prediction of political leaning. Then, we measure and discuss the performance gap between unsupervised approaches with respect to semi-supervised and supervised ones. Finally, we investigate the impact of retweets and distance from pivots, for predicting the political leaning of social media users.

6.2.1 Prediction of Political Leaning: Unsupervised Approaches

We begin by evaluating the performance of unsupervised techniques on the fine-grained prediction task. This is the core contribution of our work and results of this evaluation and comparison are shown in Table 3. Our 3 contributions are compared to a random classifier, to the technique proposed by Darwish et al. (2020) and to an approach based on word2vec + clustering.

All results reported in Table 3 are moderate, at best. This shows the difficulty of the fine-grained task. Among the evaluated techniques, our parties enriched + clustering method achieves the best results in each evaluation metric, with *micro* and *macro F1* = 0.426. This is our most complete proposal that takes full advantage of all the steps described in Sections 4 and 5. The second-best method is parties + clustering, with *micro F1* = 0.372 and *macro F1* = 0.342. The difference in performance between these 2 methods is solely due to the enrichment step, that we introduced in Section 4.3, and that allows learning better user representations as demonstrated by these results. The state-of-the-art technique by Darwish et al. (2020) achieves the third-best results, confirming its overall value. Interestingly, the parties enriched + distance strong baseline achieves performances that are only slightly worse than those of the previous methods. This further demonstrates the informativeness of the latent user ideologies that we learned. Contrarily, both the word2vec + clustering and the simple random classifier baseline obtain unsatisfactory performances, with *micro F1* = 0.142 and 0.131, respectively. This result is particularly interesting for the word2vec + clustering approach. In fact, it suggests that the user representations learned by word2vec in this context, are not suitable for a prediction step via clustering.

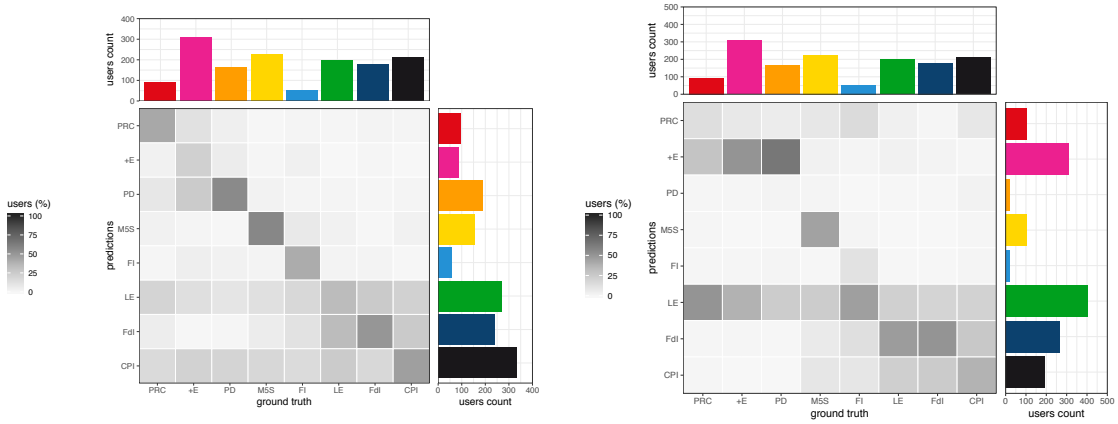
	method		macro			micro		
	ideologies	*predictions*	*precision*	*recall*	*F1*	*precision*	*recall*	*F1*
	–	random classifier	0.323	0.321	0.310	0.370	0.324	0.339
	word2vec	clustering	0.429	0.426	0.415	0.491	0.480	0.471
‡	retweets	clustering	**0.804**	0.657	0.688	0.758	0.719	0.710
	our contributions							
	parties	clustering	0.599	0.587	0.586	0.665	0.633	0.640
	parties enriched	distance	0.758	0.575	0.569	0.728	0.698	0.659
	parties enriched	clustering	0.751	**0.752**	**0.750**	**0.776**	**0.772**	**0.772**

‡: (Darwish et al., 2020)

Table 4: Performance comparison of *unsupervised* methods for coarse-grained (pole) prediction of political leaning. The best result for each evaluation metric is shown in bold font.

Table 4 shows the evaluation results of the same methods for the coarse-grained task. Difficulty-wise, this task is similar to those tackled in previous works (Barberá, 2015; Kulshrestha et al., 2017; Di Giovanni et al., 2018; Darwish et al., 2020). As such, these results are comparable to those reported in existing literature. In particular, all methods greatly improve and the best ones obtain rather good performances. As for the fine-grained task, the parties enriched + clustering method achieves the best results, with a balanced and promising *micro F1* = 0.772 and *macro F1* = 0.750. Also the method by Darwish et al. (2020) greatly improves, reaching the second-best overall result with *micro F1* = 0.710 and *macro F1* = 0.688, and the best *macro precision*. The 2 other techniques based on our approach obtain comparable results, with *micro F1* $\approx$ 0.65 and *macro F1* $\approx$ 0.57. Finally, word2vec + clustering and the random classifier again obtain markedly worse results, thus confirming the underwhelming performance already measured for the fine-grained task.

The large improvement measured by both our parties enriched + clustering method and the technique by Darwish et al. (2020) between Tables 3 and 4, demonstrate that many of the mistakes made by these techniques in the fine-grained task consisted in misclassifying a user of a given party to a different party of the same pole, rather than to a party to the opposite of the political spectrum. This is expected and is due to the difficulty of the fine-grained task. Furthermore, it explains why the same techniques obtain strikingly better results when evaluated for the prediction of poles instead of parties. Figure 7 helps to clarify this point. It shows the fine-grained confusion matrices of the 2 techniques, together with the marginal distributions of both ground-truth and predicted labels. First of all, the figure clearly highlights that more data points lay on the matrix diagonal in Figure 7a compared to Figure 7b. This visually explains the better overall performance of our technique with regards to that by Darwish et al. (2020). In addition, it shows the existence of two darker-colored 3×3 squares laying in the bottom-right and top-left corner of Figure 7a. To a lower extent, the same also occurs in Figure 7b. These regions of the confusion matrices allow to visualize the mistakes that we mentioned earlier – that is, wrong party classifications that become correct predictions when techniques are evaluated pole-wise. The fact that these regions are more visible in Figure 7a than in Figure 7b explains the better results



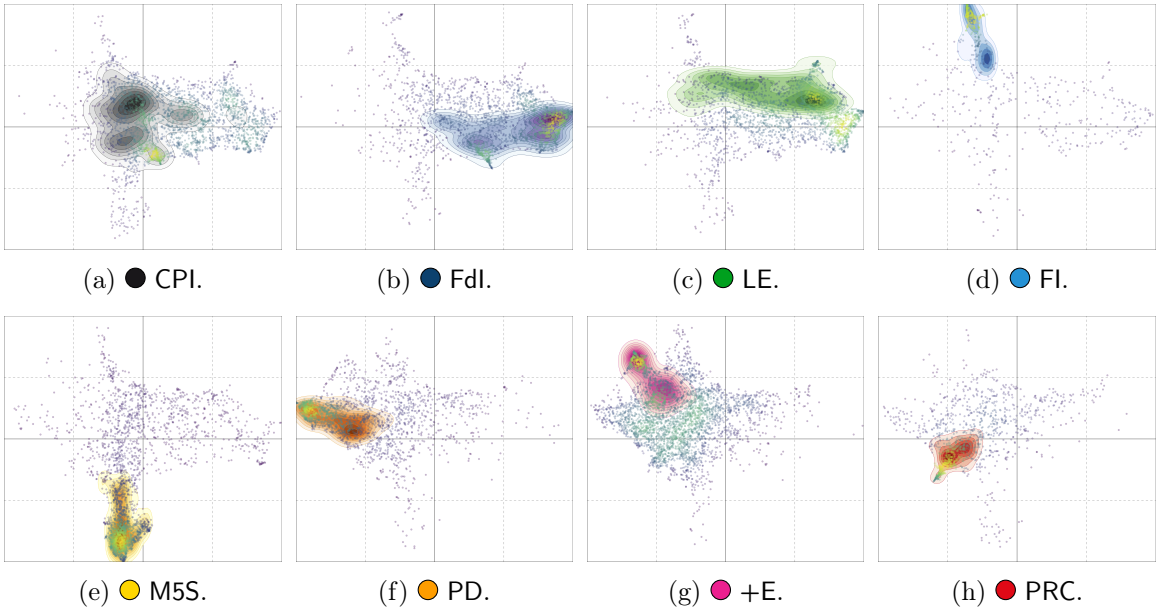
(a) Our technique.

(b) Darwish *et al.*

Figure 7: Comparison of the confusion matrices, with marginal distributions, for fine-grained (party) predictions between our proposed technique and the unsupervised method by Darwish et al. (2020). Correct predictions lay on the matrix diagonal.

in Table 4 of our technique with respect to Darwish et al. (2020), especially regarding the *macro F1* metric where we achieve 0.750 *vs* 0.688. Figure 7 also shows that our technique is particularly good at predicting center-leaning parties, such as PD, M5S and FI, for which we obtain almost flawless predictions. Contrarily, we have more difficulties in predicting far-right and far-left parties. Moreover, both techniques exhibit a bias towards overestimating right-leaning parties. On top of that, Darwish et al. also overestimate +E and almost completely neglect PD and FI.

Overall, results presented in Tables 3 and 4 and in Figure 7 demonstrate that it is very challenging to distinguish between the different parties that lay on the same side of the political spectrum. In order to provide an even more detailed breakdown of our party predictions, in Figure 8 we show ground-truth and prediction densities, for each party. Specifically in each subfigure, the scatter plot distribution shows where ground-truth users of a given party are positioned within the shared ideology space. Overlaid, the contour lines show the distribution of the test-set users predicted by our technique for that party. In other words, the maps of our ideology space shown in Figure 8 somewhat resemble the saliency maps used in computer vision systems for diagnostic purposes (Adebayo et al., 2018). Following from our previous explanation, the best results are achieved when the highest contour density perfectly overlaps the bright-colored dots. Examples of this kind are Figures 8d, 8e, 8f, 8h. Contrarily, many mistakes are made in those cases where regions of bright-colored dots are not contained within any contour line, as in Figures 8c and 8g. In addition to highlighting parties for which we are able to yield good predictions and those for which we are not, Figure 8 also allows to understand some of the reasons for our mistakes. For example it is evident that for each party, the majority of users is tightly clustered in a restricted area of the ideology space. At the same time however, a minority of users appear to be spread out throughout all the ideology space, possibly also invading a region populated by members of another party. This represents a limitation of our method for learning



(a) CPI. (b) FdI. (c) LE. (d) FI.

(e) M5S. (f) PD. (g) +E. (h) PRC.

Figure 8: Comparison between the distribution of ground-truth users and of our predictions, within the latent ideology space. For each party, the distribution of ground-truth users is shown as a density-colored scatter plot. The distribution of our predictions is shown with contour lines.

ideologies or an intrinsic limitation of working with noisy textual data, which inevitably results in wrong predictions at clustering time. For some parties, this drawback is more prominent than for others. For instance, a large portion of LE users completely overlaps the highest-density region of FdI. Such users will be erroneously predicted as supporters of FdI by our technique. Similarly, despite having a high-density cluster that we correctly detected, users of +E also appear to be spread-out across the top-left quadrant of the ideology space, which makes it difficult to cluster them all together. For the future, it will be interesting to evaluate and diagnose novel techniques for learning latent political ideologies and for predicting political leaning, by means of this visualization technique.

6.2.2 Comparison with Supervised and Semi-Supervised Approaches

Results presented in the previous section highlighted the advantages of the proposed parties enriched + clustering technique with respect to all other unsupervised techniques and baselines. However, previous works showed that the additional information exploited by supervised and semi-supervised techniques (e.g., ground-truth labels of the training-set) typically allow to yield better prediction performance compared to unsupervised approaches. Such performance is however hardly generalizable, since it strongly depends on the training set used for learning models. As a consequence, performances reported for supervised and semi-supervised techniques often represent overestimations of the capability to predict political leaning in the wild (Cohen & Ruths, 2013; Yan et al., 2019).

Following this previous line of research, here we are interested in evaluating the performance gap between the best unsupervised technique (parties enriched + clustering) and

method			macro			micro		
ideologies	*predictions*	*type*	*precision*	*recall*	*F1*	*precision*	*recall*	*F1*
parties enriched	clustering	○	0.472	0.434	0.426	0.517	0.421	0.426
supervised enriched	clustering	◐	0.433	0.394	0.392	0.485	0.384	0.411
parties enriched	SVC	◐	0.555	0.453	0.474	0.532	0.513	0.500
word2vec	SVC	◐	0.601	**0.468**	**0.485**	**0.574**	**0.554**	**0.536**
–	majority classifier	●	0.027	0.125	0.044	0.046	0.215	0.076
supervised enriched	SVC	●	**0.606**	0.453	0.481	0.551	0.517	0.504

○ : unsupervised ◐ : semi-supervised ● : supervised

Table 5: Performance comparison of the *best unsupervised* method against *semi-supervised* and *supervised* ones, for fine-grained (party) prediction of political leaning. The best result for each evaluation metric is shown in bold font.

method			macro			micro		
ideologies	*predictions*	*type*	*precision*	*recall*	*F1*	*precision*	*recall*	*F1*
parties enriched	clustering	○	0.751	0.752	0.750	0.776	0.772	0.772
supervised enriched	clustering	◐	0.748	0.745	0.745	0.787	0.785	0.785
parties enriched	SVC	◐	0.828	0.769	0.789	0.821	0.819	0.816
word2vec	SVC	◐	**0.877**	**0.822**	**0.841**	**0.875**	**0.875**	**0.871**
–	majority classifier	●	0.131	0.333	0.188	0.156	0.395	0.223
supervised enriched	SVC	●	0.822	0.761	0.780	0.822	0.823	0.816

○ : unsupervised ◐ : semi-supervised ● : supervised

Table 6: Performance comparison of the *best unsupervised* method against *semi-supervised* and *supervised* ones, for coarse-grained (pole) prediction of political leaning. The best result for each evaluation metric is shown in bold font.

semi-supervised and supervised ones. Table 5 shows results of this comparison for the fine-grained prediction task, while Table 6 presents results for the coarse-grained task. Results presented in both tables confirm previous findings and show that the best unsupervised technique is outperformed by the best supervised and semi-supervised ones. The best overall results are achieved by the semi-supervised word2vec + SVC technique, with *micro F1* = 0.536 on the fine-grained task and *micro F1* = 0.871 on the coarse-grained one. Macro results are only slightly worse in both tasks. Thus, the performance gap between the best unsupervised technique and the best overall technique is in the region of 0.11 *micro F1* on the fine-grained task and 0.10 *micro F1* on the coarse-grained one. Taking into account the differences previously reported in Tables 3 and 4 between unsupervised techniques, these last results represent non-negligible yet modest differences in performance. Notably, the parties enriched + clustering unsupervised technique is also capable of beating the supervised enriched + clustering technique in the challenging fine-grained task, in addition to largely beating the simple majority classifier supervised baseline in both tasks.

An interesting result that clearly emerges from Tables 5 and 6 is the superiority of all the approaches based on SVC classifiers for the prediction step. Independently on the methodology used for obtaining political ideologies and on the overall approach to the task (e.g.,

semi-supervised or supervised), the 3 methods leveraging an SVC consistently obtained the 3 best overall results in both the fine- and coarse-grained tasks. An important contribution to these positive results is given by the data distribution of the different splits of our dataset. In fact, as anticipated in Section 3.2, our data splitting strategy implied that no drift is present between the training and test partitions of our dataset. Under this favorable laboratory condition, supervised classifiers are able to maximize their learning phase on data instances in the training-set, and to effectively carry over what they learnt to the test-set. However, it is known that real-world conditions are characterized by issues such as concept drift that limit the generalizability of supervised approaches (Lu et al., 2018). In presence of concept drift, or of any other factor that shifts the test distribution away from the one used in training, supervised approaches end up being unreliable. Instead, unsupervised approaches, such as the one proposed in our work, are able to better adapt to possible drifts. For instance, with reference to the ideology space shown in Figure 8, while a supervised classifier learns fixed decision boundaries for the different parties based on the data distribution of the training-set, our unsupervised clustering approach is capable of highlighting regions of the ideology space featuring high density, independently on their position.

6.2.3 Validation: Concept Drift

The results presented so far are computed on the test-set split of our dataset, obtained via a stratified random sampling of the users as explained in Section 3.2. However, as anticipated in the previous section, a more rigorous evaluation can be conducted by assessing the performance of our technique on a time-dependent test-set, by assigning users to either the training-, validation-, or test-set according to the time when they tweeted. The advantage of this evaluation strategy is that, in general, time-wise splits are more representative of the conditions in which machine learning models are used, as they allow to test a model's ability to withstand issues that emerge through time, such as concept-drift (Lu et al., 2018). In turn, a model capable of withstanding such issues would open up the possibility to carry out longitudinal analyses and even to nowcast political leanings (Lampos & Cristianini, 2012; Avvenuti et al., 2017; Tsakalidis et al., 2018). For these reasons, we performed an additional experiment by evaluating our method in this, more stringent, condition.

Specifically, we first obtained a new time-wise test-set that contains 5,524 users that only tweeted after August 15, 2019. All other users from our dataset are assigned to either the training- or validation-set, which contain users that tweeted before the threshold date. Next, we used our best method (i.e., parties enriched + clustering) to repeat both the party (fine-grained) and the pole (coarse-grained) prediction tasks. Finally, we compared the results obtained by our method on the time-dependent test-set with those obtained on the original (random) one. Regarding party predictions, our method obtains *macro F1* = 0.388 and *micro F1* = 0.389 on the time-dependent test-set, whereas it obtained both *macro* and *micro F1* = 0.426 on the random one. For pole predictions, our method obtains *macro F1* = 0.721 and *micro F1* = 0.771 on the time-dependent test-set, whereas it obtained *macro F1* = 0.750 and *micro F1* = 0.772 on the random one. Summarizing, the more stringent evaluation resulted in a maximum of 9% performance decrease on the party prediction task, and in a maximum of 4% performance decrease on the pole prediction one. The overall positive results of our unsupervised method are confirmed.

6.2.4 Validation: Members of the European Parliament

The ground-truth labels for users of our dataset are implicitly derived from user likes to party tweets, as detailed in Section 3.3. On the one hand this labeling strategy removed the need for a manual labeling of our dataset and avoided possible human errors and biases entering our ground-truth (Pandey et al., 2019). On the other hand however, an automatic labeling strategy does not necessarily exclude the risk of inconsistencies or wrong labels. In order to further validate the correctness of our predictions, we also applied our technique to a small set of users whose preferred party and pole are publicly known. Specifically, we focused on the Italian members of the European parliament (MEPs) as of 2019. These represent active politicians whose party and pole affiliations can be used as reliable ground-truth labels in our prediction tasks. Notably, the majority of MEPs also has an official Twitter account linked to its public page on the European parliament website[14].

Similarly to the users in our Twitter dataset, we thus collected 200 tweets for each of the 34 Italian MEPs with an official Twitter account. Then, we applied our best method (i.e., parties enriched + clustering) to predict the party leaning of the MEPs. Finally, we evaluated our method by comparing its predictions with the party affiliations of the MEPs. We repeated these steps also for the method by Darwish et al. (2020), thus enabling a comparison of our results with those of this state-of-the-art technique. On this party prediction task, our method achieved *macro F1* = 0.651 and *micro F1* = 0.823. Instead, the method by Darwish et al. (2020) achieved *macro F1* = 0.377 and *micro F1* = 0.662. Overall, our results are very positive, confirming the good performance of our technique and its superiority with respect to the best existing unsupervised competitor. We also note that the results reported in this section are better than those reported in Tables 3 and 5. This is expected by considering that MEPs are very politically-active users with clear political inclinations. As such predicting their political leaning represents a simpler task with respect to predicting the political leaning of generic Twitter users.

6.2.5 Sensitivity Analysis: Distance

In this section and in the subsequent ones, we report results of a sensitivity analysis that we carried out on the best unsupervised technique that we proposed: parties enriched + clustering.

The rationale for the analysis discussed in this section stems from results of Tables 3 and 4. In particular, the evaluation of the unsupervised approaches highlighted the promising performance of the parties enriched + distance baseline. Given an ideology space, this technique assigns a label to each user based on its distance to the pivots. In other words, the simple distance-based prediction strategy employed in this baseline was capable of yielding positive results. This suggests that the distance from the pivots in the ideology space is a relevant parameter that has an impact on our predictions. With the goal of evaluating this facet, in Figure 9 we show results of an analysis where we evaluated the performance of our parties enriched + clustering technique, as a function of user distance from the pivots. Results shown in figure confirm our previous intuition. When only evaluating predictions for users that lay near to one of the pivots, our results are extremely accurate. For instance, when considering only users whose min-max normalized distance ≤ 0.2, our technique obtains

14. `https://www.europarl.europa.eu/meps/en/home`

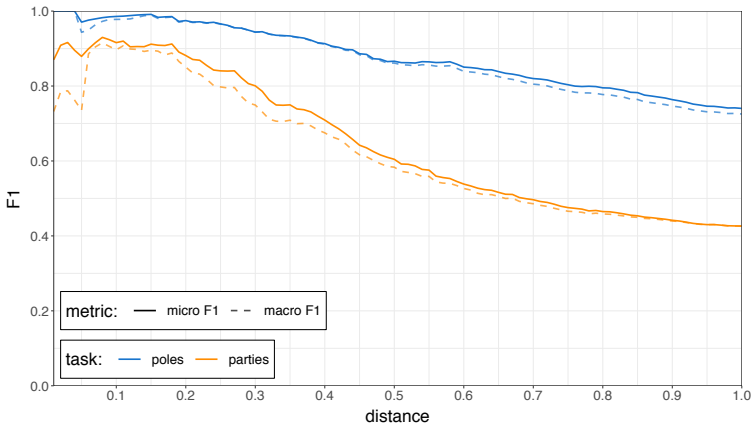


Figure 9: Performance evaluation of our parties enriched + clustering unsupervised technique as a function of user distance from the pivots. For users that lay near to one of the pivots, we are able to provide fine- (party) and coarse-grained (pole) predictions with exceptional accuracy. Instead, most of our mistakes occur for users that are positioned far away from all pivots.

micro F1 > 0.90 and > 0.98 on the fine- and coarse-grained task, respectively. Exceptional results indeed, considering the difficulty of the tasks at hand. As we include in our evaluation also users who lay further away from any pivot, our results worsen. At the end of our evaluation, when we consider all users independently on their distance, we end up with the same results already reported in Tables 3 and 4.

The decreasing trend shown in Figure 9 demonstrates that the accuracy of our predictions strongly depends on a user's distance to the pivots. In turn, this facet can be exploited to complement our predictions with a confidence score that states how likely a given prediction is to be correct. For users that lay near to one of our pivots in the ideology space, we are able to provide predictions with large confidence scores (e.g., $> 90\%$). Conversely, for users that lay far away from all pivots, we are still able to provide a prediction, but with a much lower confidence.

These results also suggest one possible strategy for quickly improving the performance of our technique – that is, increasing the number of pivots. This simple operation would reduce the average distance of users from the pivots, thus allowing to obtain overall better performances. However, having a large number of pivots requires additional manual effort and it would also move the approach towards a semi-supervised one, with all the implications previously discussed. We remark that for all the experiments in this work, we used the minimum possible number of pivots: one for each considered political party.

6.2.6 Sensitivity Analysis: Tweets

Our proposed technique is based on the analysis of the textual content of the tweets shared by social media users. One important aspect to consider when evaluating our technique is thus its sensitivity to the number of available tweets per user. Intuitively, users for which only a small number of tweets are available represent more challenging predictions than users who shared many tweets. With the goal of evaluating this aspect, Figure 10 shows the

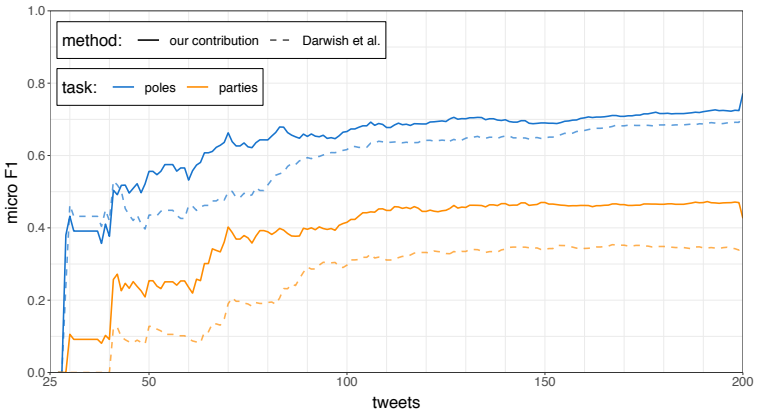


Figure 10: Performance evaluation of Darwish et al. (2020) and of our parties enriched + clustering unsupervised technique, as a function of the number of user tweets. Our technique consistently beats the competitor, especially in the challenging party prediction (i.e., fine-grained) task. Both techniques show overall stable performance for users having ≥ 100 tweets. For users having < 100 tweets the performance of both techniques starts decreasing. For < 40 tweets performances rapidly plummet.

performance of our technique and of Darwish et al. (2020), as a function of the number of user tweets, for both tasks. We recall that we collected maximum 200 tweets per user and that we discarded users for which we could collect < 25 tweets.

Results in Figure 10 show that our technique consistently beats the competitor in both tasks and for all users, independently on the number of their tweets. The only exception is represented by users having < 40 tweets, on the fine-grained party prediction task, for which Darwish et al. obtain slightly better results than us. Apart from this, our method always achieves superior results, especially in the challenging party prediction task. Interestingly, both techniques show overall stable performance for users having ≥ 100 tweets. Instead, as expected, for users having < 100 tweets the performance of both techniques starts decreasing. The decreasing trend is particularly steep for users having < 40 tweets, for which the performances of both techniques rapidly plummet.

In addition to serving as further evaluation of our technique, these results also provide guidance for applying tweet-based predictors of political leaning in-the-wild. In fact, our experiments suggest that near-optimal results can be expected for users with ≥ 100 tweets, and reduced performance otherwise. In particular, predictions obtained for users having < 40 or 50 tweets should undergo additional scrutiny and validation, since misclassifications are frequent under these operating conditions.

6.2.7 Sensitivity Analysis: Retweets

In many recent works, retweets have been used as a strong signal in several prediction tasks on social media, including the estimation of stance and political leaning (Aldayel & Magdy, 2019), degree of automation (Mazza et al., 2019), extent of coordination among users (Nizzoli et al., 2021) and percentage of fake messages in an online discussion (Tardelli et al., 2022), to name but a few notable examples. In particular, one of the techniques that we evaluated in

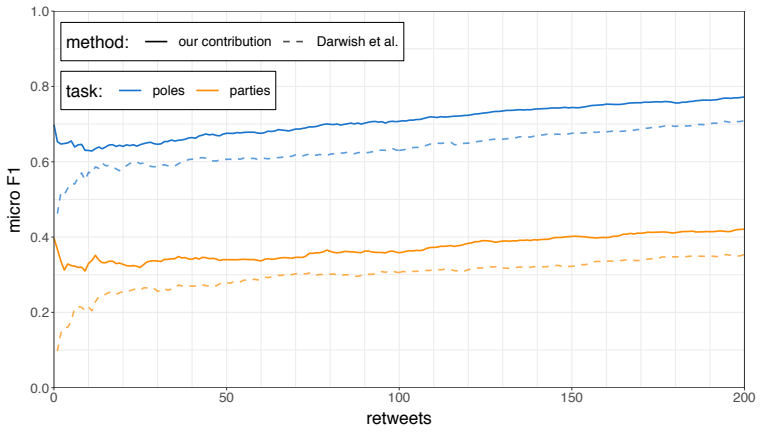


Figure 11: Performance evaluation of Darwish et al. (2020) and of our parties enriched + clustering unsupervised technique, as a function of the number of user retweets. Both techniques show degraded performance when classifying users with few retweets, showing the importance of this signal for the task. However, our technique consistently outperforms the competitor and does not suffer from a steep performance drop for users with ≤ 15 retweets.

this work solely depends on retweets for estimating political leaning (Darwish et al., 2020). In addition, also our proposed technique uses that information, although not as explicitly as by Darwish et al. (2020). In fact, as explained in Section 4, in our work retweets only partly contribute to document embeddings, which in turn contribute to our latent user representations.

Similarly to the previous experiment, here we were interested in evaluating the impact that retweets have on the predictions of political leaning generated by our technique and by that of Darwish et al. (2020). To carry out this experiment, we repeatedly evaluated both techniques on subsets of test-set users featuring different numbers of retweets, starting from users with no retweets at all, and concluding our experiment with users with 200 retweets (the maximum number of tweets that we collected per each user). Results are shown in Figure 11. As expected, both techniques achieve worse results for users with few retweets, confirming the informativeness of this feature. Our proposed technique consistently outperforms the one from Darwish et al. (2020) and the gap between the 2 shows only minor fluctuations along the x axis. However, a marked difference is shown for users that feature an extremely low number of retweets. Indeed for users with ≤ 15 retweets, the performance of Darwish et al. (2020) plummets in both prediction tasks. On the contrary, our technique exhibits a different behavior, as it does not appear to be impacted so negatively by an extremely low number of retweets. The difference between the behavior of the 2 techniques is explained by considering that retweets are the only information exploited by Darwish et al. (2020), while they are an important – yet minor – part of all the information that our technique leverages.

6.2.8 Limitations and Open Challenges

In this section we carry out a detailed analysis of the main types of errors made by our proposed method. To reach this goal, we manually selected a set of users that were projected by our technique to a region of the latent ideology space that is not associated with their ground-truth party label. For these users that were projected far from their party – and thus, that were subsequently wrongly labeled by the clustering step – we manually analyzed their Twitter timelines, so as to identify the root causes for our misclassifications. This analysis contributes to highlighting current limitations in content-based unsupervised approaches to the prediction of social media political leaning, also highlighting open challenges and valuable directions for future research.

We first evaluated *major* misclassifications – namely, cases where users were projected to a region of the ideology space related to parties of the opposite pole with respect to the ground-truth party of the users. For example, users favoring a left-leaning party (e.g., PD) that were erroneously projected to a region of the ideology space associated with extreme-right parties (e.g., LE, FdI, CPI). These cases yield errors both in the fine-grained party prediction task, as well as in the coarse-grained pole prediction task. Overall, the total number of these major misclassifications is small, but it is nonetheless interesting to assess the causes for these errors. The analysis of these major misclassifications revealed that some of our ground-truth labels contrast with the information contained in the tweets from the user timeline. Ground-truth labels were automatically obtained from user likes to party tweets. Instead, our classifications are derived from user tweets. Thus, the majority of cases of major misclassifications are related to users that liked many tweets by a given party, but that support a different party in their own tweets. This is a peculiar and interesting behavior that, to the best of our knowledge, is undocumented. The existence of a subset of users exhibiting this behavior mandates to carefully consider the source of ground-truth labels in future works, since considering user likes or following relationships might convey different and contrasting information with respect to that obtainable from user tweets. In the remaining cases of major misclassifications by our system, we were not able to correctly detect the political leaning of the user mainly due to: (i) wrong understanding of tweet semantics (more on this in the following); or (ii) an objective difficulty in understanding the political orientation of the user, due to ambiguous and contrasting political content. This latter case is not a limitation of our technique, since also human evaluators would struggle to reliably provide predictions for certain users, but rather an inherent challenge in the classification of users that express few or ambiguous political positions. Such challenge has already been noted in earlier works on this same task (Cohen & Ruths, 2013), as well as on other social media-related tasks (Cresci et al., 2018).

We also assessed causes of *minor* misclassifications – namely, cases where a user is labeled with a wrong party in the fine-grained party classification task, but it is correctly labeled in the coarse-grained pole prediction task. In such cases, several misclassifications are caused by a wrong interpretation of tweet semantics or by the weight (i.e., the importance) that our system assigned to certain political tweets. In fact, the political orientation of a user is not a binary concept, but it is rather a nuanced concept often involving ideas and opinions that align with the political line of more than one party. In particular, it is common for a user to support opinions from multiple politically-close parties. To this regard our system,

original tweet	translated tweet	party	score
Type 1: tweets in favour of a party/politician, that receive low scores for that party:			
@matteosalvinimi Matteo Salvini gli italiani sceglieranno il miglior Matteo (Salvini)	@matteosalvinimi Matteo Salvini Italians will choose the best Matteo (Salvini)	LE	0.126
@LegaSalvini Mitico Matteo Salvini sei il nostro capitano	@LegaSalvini Mythical Matteo Salvini you are our captain	LE	0.066
Type 2: tweets against a party/politician, that receive high scores for that party:			
Deve andare in pensione. Berlusconi ormai è fulminato.	He has to retire. Berlusconi is stoned.	FI	0.454
@DSantanche @FratellidItaIia tornare? devi cominciare a crescere Santanchè, hai novant'anni e "ragioni" come una lattante	@DSantanche @FratellidItaIia coming back? you have to grow Santanchè, you are ninety years old and you still "think" like a baby	FdI	0.470
Type 3: tweets with limited/no political information, that receive high scores for a party:			
RT @oss_romano: #27agosto #rassegnastampa Un mondo di fraternità e pace è possibile. Il #Papa incoraggia le iniziative per dare attuazione ...	RT @oss_romano: # 27agosto #rassegnastampa A world of fraternity and peace is possible. The #Pope encourages initiatives to implement ...	PRC	0.658
RT @visit_lazio: Tra le 100 esperienze al mondo da vivere, il settimanale @TIME include il @Castello_Severa nella lista world's greatest ...	RT @visit_lazio: Among the 100 experiences in the world to live, the @TIME magazine includes @Castello_Severa in the world's greatest ...	M5S	0.480
Type 4: tweets with local, subjective, or very specific information, that receive high scores for a party:			
RT @c_appendino: Asfalto nuovo per via Cigna. Una buona notizia per i tanti cittadini che transitano su questa importante arteria	RT @c_appendino: New asphalt on via Cigna. Good news for the many citizens who move through this important thoroughfare	M5S	0.773
RT @virginiaraggi: Partiti i lavori di restauro della Fontana delle Rane nel quartiere Coppedè. L'intervento è il primo di questa portata	RT @virginiaraggi: Restoration work on the Fontana delle Rane in the Coppedè district has begun. The intervention is the first of this magnitude	M5S	0.602

Table 7: Examples of problematic tweets that were incorrectly assessed by our tweet party classifier. For each of the 4 main types of problematic tweets, we report some examples specifying the reference political party to which the error is referred and the corresponding score computed by the tweet party classifier. Scores are in the $[0, 1]$ range.

as a human evaluator would do, weighs the available information as best as possible, but with an inevitable degree of uncertainty. Finally, we also analyzed errors for users projected to the central (i.e., most uncertain) area of the ideology space of Figure 5a, finding that the projection errors are mainly due to one of the following reasons: (i) users with insufficient political content in their timeline; (ii) automated accounts that produce extremely varied content (i.e., news bots); and (iii) users that repeatedly attack certain political parties and leaders, but that do not explicitly support any party[15]. Challenges related to the analysis of the first category of users are well-known in literature. For instance, Cohen and Ruths

15. For these users, we know who they *do not* support, but we do not know who they *do* support.

(2013) refer to them as *politically inactive* users. Instead, challenges related to the analysis of bots and political antagonists, are rather undocumented, despite the widespread presence of both these types of accounts in our online ecosystems (Lokot & Diakopoulos, 2016; Nizzoli et al., 2021).

Given that our projections are based on the scores assigned to user tweets by the tweet party classifier, wrong predictions by our technique are typically due to initial errors by the tweet party classifier. We now turn our attention to these errors, so as to provide practical examples of problematic tweets. Our analysis highlighted 4 main categories of problematic tweets, summarized in Table 7. A first set of errors is due to problematic tweets of type 1 (tweets in favour of a party/politician, that receive low scores for that party). Here, our classifier was unable to provide high scores for the correct party because of the limited number of these tweets used to train it. Increasing our dataset, or anyway feeding more tweets of this type to the classifier, would likely remove this type of error. Errors due to the second type of problematic tweets (tweets against a party/politician, that receive high scores for that party) are more challenging. First of all, those tweets do not express support for any party nor candidate. Thus, there is an intrinsic difficulty in assigning a high score for a party. Moreover, they negatively – yet explicitly – mention a party, which tricked our classifier into giving a high score for that party. This second issue implies that our deep learning tweet party classifier was unable to correctly "understand" the meaning of those tweets. This problem can be mitigated by implementing the classifier with more complex and powerful deep learning architectures, such as those based on modern pretrained language models (e.g., BERT, T5). These state-of-the-art natural language understanding systems are capable of grasping subtle nuances in the language used for or against a given political party. As such, they would contribute to reducing this type of errors. The third type of problematic tweets (tweets with limited/no political information, that receive high scores for a party) are due to the challenges of classifying items that do not convey any useful information for the machine learning task at hand. In this situation, classifiers usually yield unreliable predictions. One possible way of solving this issue is by carrying out an additional filtering step in our analysis pipeline. For instance, we could train a separate binary classifier to distinguish between politically-related and unrelated tweets. Then, only politically-related tweets would be given to the party tweet classifier for computing a party score. The last type of problematic tweets (tweets with local, subjective, or very specific information, that receive high scores for a party) represents another big challenge. In the examples from the bottom rows of Table 7, a user is expressing positive opinions about the local administration. Notably, the high scores given by our classifier do not necessarily represent errors, in a strict sense. In fact, appreciation for the work of a local administration surely conveys a certain extent of political information. However, for these users our classifier should have given more weight to other, more explicit, political tweets with respect to those supporting the local administration.

7. Conclusions

We proposed a novel unsupervised technique for estimating the political leaning of social media users. Our solution leverages a deep neural network in a representation learning task, for analyzing user tweets and for learning latent political ideologies. Then, users are

projected in a low-dimensional ideology space and are subsequently clustered. The political leaning of a user is automatically derived from the cluster to which the user is assigned. We evaluated our technique on two prediction tasks and we compared it to baselines and other state-of-the-art approaches. The fine-grained task aims to infer the preferred political party of each user, out of 8 possible parties. The – easier – coarse-grained task aims to infer the high-level political leaning of each user, in a 3-class classification task. Among all unsupervised techniques that we evaluated, our proposed one achieved the best results in both tasks, with *micro F1* = 0.426 and 0.772, respectively for the fine- and coarse-grained task. It also achieved comparable results to some of the semi-supervised and supervised techniques. However, the best unsupervised technique is outperformed by the best semi-supervised and supervised ones, given the additional information that the latter exploit. Moving forward, we demonstrated that we can exploit the topology of our learned ideology space to assign a confidence score to our predictions, thus allowing to retain only those predictions for which the confidence meets a desired threshold. Finally, we analyzed the relationship between our predictions and the number of tweets and retweets performed by users, showing that our technique is able to provide accurate predictions also for users who tweet or retweet sporadically, contrarily to other state-of-the-art methods.

Our results advance the state-of-the-art for unsupervised prediction of political leaning – an increasingly popular task. For the future we aim to provide additional contributions by devising better techniques for learning latent political ideologies, a step where there is still large room for improvement. To this regard, another interesting direction of research involves providing interpretations to the dimensions of the latent ideology space. As shown in our results, the different parties seem to position themselves in different regions of the space. Hence, being able to interpret the main dimensions of the ideology space could provide additional and valuable information. For the future we also aim at learning aspect-based stances on a number of politically-relevant issues (e.g., immigration, economy, rights, and more). Finally, we plan to leverage our technique for carrying out a longitudinal analysis aimed at investigating fluctuations in leaning due to important real-world events, such as during electoral campaigns. This latter experimental setup would also be valuable toward assessing the robustness of our system, as well as of others tackling the same task, to known issues such as concept drift and other temporal variations.

Acknowledgments

Correspondence should be addressed to Dr. Stefano Cresci (*stefano.cresci@iit.cnr.it*).This research is supported in part by the EU H2020 Program under the scheme "INFRAIA-01-2018-2019: Research and Innovation Action" grant agreement #871042 *SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics.*

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *The 32nd Annual Conference on Advances in Neural Information Processing Systems (NeurIPS'18)*, pp. 9505–9515.

Ahmed, S., Jaidka, K., & Skoric, M. M. (2016). Tweets and votes: A four-country comparison of volumetric and sentiment analysis approaches. In *The 10th International AAAI Conference on Web and Social Media (ICWSM'16)*. AAAI.

Aldayel, A., & Magdy, W. (2019). Your stance is exposed! Analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–20.

Avvenuti, M., Cresci, S., La Polla, M. N., Meletti, C., & Tesconi, M. (2017). Nowcasting of earthquake consequences using big social data. *IEEE Internet Computing*, *21*(6), 37–45.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior*, *116*, 106633.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology?. *Political Behavior*, *39*(3), 553–583.

Benkler, Y., Faris, R., Roberts, H., & Zuckerman, E. (2017). Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*, *3*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: Theory and experiment*, *2008*(10), P10008.

Bond, R., & Messing, S. (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, *109*(1), 62–78.

Busch, K. B. (2016). Estimating parties' left-right positions: Determinants of voters' perceptions' proximity to party ideology. *Electoral studies*, *41*, 159–178.

Cinelli, M., Cresci, S., Galeazzi, A., Quattrociocchi, W., & Tesconi, M. (2020). The limited reach of fake news on Twitter during 2019 European elections. *PLoS One*, *15*(6), e0234689.

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy!. In *The 7th International AAAI Conference on Web and Social Media (ICWSM'13)*. AAAI.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of Twitter users. In *The 3rd IEEE International Conference on Social Computing (SocialCom'11)*. IEEE.

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72–83.

Cresci, S., Cimino, A., Avvenuti, M., Tesconi, M., & Dell'Orletta, F. (2018). Real-world witness detection in social media via hybrid crowdsensing. In *The 12th International AAAI Conference on Web and Social Media (ICWSM'18)*, pp. 576–579. AAAI.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, *80*, 56–71.

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, *110*(15), 5791–5796.

Darwish, K. (2018). To Kavanaugh or not to Kavanaugh: That is the polarizing question. arXiv preprint arXiv:1810.06687.

Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised user stance detection on Twitter. In *The 14th International AAAI Conference on Web and Social Media (ICWSM'20)*, Vol. 14, pp. 141–152. AAAI.

Di Giovanni, M., Brambilla, M., Ceri, S., Daniel, F., & Ramponi, G. (2018). Content-based classification of political inclinations of Twitter users. In *The 2018 IEEE International Conference on Big Data (BigData'18)*. IEEE.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, *3*, 271–277.

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2016). Quantifying controversy in social media. In *The 9th International Conference on Web Search and Data Mining (WSDM'16)*. ACM.

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2017). Reducing controversy by connecting opposing views. In *The 10th International Conference on Web Search and Data Mining (WSDM'17)*. ACM.

Grčar, M., Cherepnalkoski, D., Mozetič, I., & Novak, P. K. (2017). Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, *4*(1), 6.

Hegelich, S., & Janetzko, D. (2016). Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet. In *The 10th International AAAI Conference on Web and Social Media (ICWSM'16)*. AAAI.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media. In *The 20th Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'17)*. ACM.

Lahoti, P., Garimella, K., & Gionis, A. (2018). Joint non-negative matrix factorization for learning ideological leaning on Twitter. In *The 11th International Conference on Web Search and Data Mining (WSDM'18)*. ACM.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social Web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 1–22.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lokot, T., & Diakopoulos, N. (2016). News bots: Automating news and information dissemination on Twitter. *Digital Journalism*, *4*(6), 682–699.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363.

Luceri, L., Deb, A., Badawy, A., & Ferrara, E. (2019). Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of the 28th Web Conference (WWW'19 Companion)*, pp. 1007–1012. IW3C2.

Marchetti-Bowick, M., & Chambers, N. (2012). Learning for microblogs with distant supervision: Political forecasting with Twitter. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pp. 603–612.

Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLoS One*, *8*(11), e74516.

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting temporal patterns for botnet detection on Twitter. In *The 11th International ACM Web Science Conference (WebSci'19)*, pp. 183–192. ACM.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Misra, I., Lawrence Zitnick, C., Mitchell, M., & Girshick, R. (2016). Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *The 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 2930–2939.

Nikolov, D., Flammini, A., & Menczer, F. (2021). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *The Harvard Kennedy School Misinformation Review*, *0*(0), 1–13.

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., & Tesconi, M. (2021). Coordinated behavior on social media in 2019 UK General Election. In *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*. AAAI.

Pandey, R., Castillo, C., & Purohit, H. (2019). Modeling human annotation errors to design bias-aware systems for social stream processing. In *The 11th ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM'19)*, pp. 374–377. IEEE/ACM.

Pasquino, G. (2019). The state of the Italian Republic. *Contemporary Italian Politics*, *11*(2), 195–204.

Pennacchiotti, M., & Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: User classification in Twitter. In *The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM.

Pla, F., & Hurtado, L.-F. (2014). Political tendency identification in Twitter using sentiment analysis techniques. In *The 25th International Conference on Computational Linguistics (COLING'14)*. ACL.

Poole, K. T., & Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, *29*(2), 357–384.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: political ideology prediction of Twitter users. In *The 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pp. 729–740.

Rizoiu, M.-A., Graham, T., Zhang, R., Zhang, Y., Ackland, R., & Xie, L. (2018). #DebateNight: The role and influence of socialbots on Twitter during the 1st 2016 US presidential debate. In *The 12th International AAAI Conference on Web and Social Media (ICWSM'18)*. AAAI.

Stefanov, P., Darwish, K., & Nakov, P. (2019). Predicting the topical stance of media and popular Twitter users. In *The 11th International Conference on Social Informatics (SocInfo'19)*.

Tardelli, S., Avvenuti, M., Tesconi, M., & Cresci, S. (2022). Detecting inorganic financial campaigns on Twitter. *Information Systems*, *103*, 101769.

Tsakalidis, A., Aletras, N., Cristea, A. I., & Liakata, M. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the Greek referendum. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 367–376. ACM.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature*. William and Flora Hewlett Foundation.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *The 4th International AAAI Conference on Web and Social Media (ICWSM'10)*. AAAI.

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among twitter users in Italy. *Journal of Computer-Mediated Communication*, *20*(2), 221–239.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *The 31st Annual Conference on Advances in Neural Information Processing Systems (NeurIPS'17)*, pp. 5998–6008.

Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*, *3*.

Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2158–2172.

Yan, H., Das, S., Lavoie, A., Li, S., & Sinclair, B. (2019). The congressional classification challenge: Domain specificity and partisan intensity. In *The 20th ACM Conference on Economics and Computation (EC'19)*, pp. 71–89. ACM.

Yan, H. Y., Yang, K.-C., Menczer, F., & Shanahan, J. (2021). Asymmetrical perceptions of partisan political bots. *New Media & Society*, *23*(10), 3016–3037.