Stance Classification through Proximity-based Community Detection

Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, Mohand Boughanem

Stance Classification through Proximity-based Community Detection

Ophélie Fraisier
CEA-Tech Occitanie
IRIT, Université de Toulouse, CNRS
ophelie.fraisier@cea.fr

Guillaume Cabanac
IRIT, Université de Toulouse, CNRS
guillaume.cabanac@irit.fr

Yoann Pitarch
IRIT, Université de Toulouse, CNRS
yoann.pitarch@irit.fr

Romaric Besançon
CEA LIST, Nano-INNOV
romaric.besancon@cea.fr

Mohand Boughanem
IRIT, Université de Toulouse, CNRS
mohand.boughanem@irit.fr

ABSTRACT

Numerous domains have interests in studying the viewpoints expressed online, be it for marketing, cybersecurity, or research purposes with the rise of computational social sciences. Current stance detection models are usually grounded on the specificities of some social platforms. This rigidity is unfortunate since it does not allow the integration of the multitude of signals informing effective stance detection. We propose the SCSD model, or Sequential Community-based Stance Detection model, a semi-supervised ensemble algorithm which considers these signals by modeling them as a multi-layer graph representing proximities between profiles. We use a handful of seed profiles, for whom we know the stance, to classify the rest of the profiles by exploiting like-minded communities. These communities represent profiles close enough to assume they share a similar stance on a given subject. Using datasets from two different social platforms, containing two to five stances, we show that by combining several types of proximity we can achieve excellent results. Moreover, we compare the proximities to find those which convey useful information in term of stance detection.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Information systems** → *Web mining*;

KEYWORDS

Stance detection, Social media, Computational Social Science, Political discourse

1 INTRODUCTION

Since the launching of SixDegrees in 1997, social media sites became deeply embedded in users' lives [10]. This includes well-known networking sites, such as Facebook and Twitter, but also other platforms enabling users to interact with their content. Several domains evolved to take advantage of this overabundance of online activities, such as marketing [7, 31, 34] or cybersecurity [12, 26]. These domains have an increasing interest in stance detection due to its wide range of applications. It can be used to detect persons of interest [33], like Democrats and Republicans during a political campaign for example. Furthermore, automatic stance detection is also a precious tool for social scientists. Indeed, it enables researchers to have a better grasp on some research objects by extending observations to large corpora of data which used to be unusable. This task has already been tackled, but the proposed approaches often rely on large quantity of annotated text or specific social interactions [24, 38]. The written content is indubitably a rich source of information, and so is the pattern of interactions, but there are other types of information which could be helping us characterize profiles.

Profiles on social media are linked by a variety of information: elements of language, various social interactions, location, etc. If we try to infer their stance on a topic, we intuitively understand that using several bodies of evidence is beneficial: while one profile could heavily share the publications of a political candidate, another one might be more discreet in sharing but could use the same rethorics in her message, or could be passive but live in an area known to be in favor of a specific stance. Our hypothesis is that several evidences hinting at a strong similarity between two profiles are likely to mean that said profiles share a common stance. This is reinforced by the body of literature showing that social media are highly polarized on political topics [14, 25]. We model these *proximities* with a multi-layer graph, each layer representing a specific proximity, with profiles as vertices, and weighted edges as similarities. On this multi-layer proximity graph, we can extract communities, which can then be used for stance detection. We assume that when using adequate proximities, the extracted communities will be extremely homogeneous in terms of stance, allowing us to infer the stance of many profiles with only a handful of seed profiles. The contributions of this paper are:

(1) A detailed analysis strengthening our initial intuition that communities based on proximities tend to be very homogeneous in terms of stance.
(2) A semi-supervised ensemble model for stance detection employing this combination of proximities, and requiring a minimal amount of human investment. Our approach is flexible and generic in terms of platforms, proximities, and number of stances.
(3) Higher effectiveness than state-of-the-art methods, with F1-score as high as 95% while using only 1% of annotated data.

Section 2 presents the background of our work with related work on polarization on social media and stance detection. The data model used throughout this paper is detailed in Section 3. In Section 4, we expose the preliminary analysis questions we addressed before formalising our model, as well as the datasets and proximities used in our experiments. The stance detection model is presented and evaluated in Section 5 and Section 6.

This project is co-financed by the European Union – Europe is committed to Midi-Pyrénées with the European fund for regional development.

HT'18, July 2018, Baltimore, Maryland, USA

https://doi.org/http://dx.doi.org/10.1145/XXXXXXX.XXXXXXX

Table 1: Positioning in the existing literature

	[22]	[36]	[9]	[27]	[39]	[4]	[33]	[37]	[24]	[38]	[16]	SCSD
Textual features	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Social interactions	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓
Other information	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓
Few annotations	✓	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓	✓
More than 1 platform	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓
More than 2 stances	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓

2 RELATED WORK

2.1 Polarization on social media

Several studies showed that online social media were highly polarized in some contexts, particularly for political topics [23, 25]. They revealed the presence of *"echo chambers"* on several platforms [35]. This term describes a phenomenon characterised by users prefering to interact with like-minded people. Conservative and liberal blogs tend to mainly reference blogs from their own ideological camp, as shown by their linking patterns and discussion topics [2]. Similarly, Twitter's retweet networks concerning the 2010 and 2014 US midterms elections and the 2014 Scottish independance referendum were highly polarized between left- and right-leaning profiles, while people interacted more freely in the mention networks [14, 20]. Even on Wikipedia, controversies occur mainly in neighbourhoods of related topics [17]. This suggests that some topics tend to be particularly attractive for users promoting diverse mindsets. It is important to note however that on non-political topics, polarization is usually more nuanced [5].

2.2 Stance detection

Text is often the main piece of information used to determine stance. Several researchers studied debate sites and argumentative essays [22]. [36] used a topic model to discover viewpoints, topics, and opinions to classify texts on the Israeli-Palestinian conflict according to their ideological leaning. Other studies focused on less structured platforms: Twitter has, for instance, been largely used, due to its large popularity and the facility to collect data. [9] used a statistical model to determine the political stance of politicians from the Belgian Parliament on Twitter. [27] trained an SVM model including sentiments as features to detect if profiles were "for" or "against" given targets. Forums are other exploitable information silos: [39] used neural networks on a breast cancer forum to identify the profiles' stances on complementary and alternative medicine.

Alternatively, some works rely on social interactions between profiles. [4] built a bayesian model inferring ideology of profiles according to which political actors they are following. [33] identified pairs of profiles with differing opinions thanks to a retweet-based label propagation algorithm tied to a supervised classifier. [37] propose an unsupervised topic model taking into account the text and social interactions to identify viewpoints. [24] used an SVM on textual content, retweets, and mentions to predict the future attitude of profiles in the aftermath of a major event. Their results show that social features are of prime importance for this task. [38] also used a combination of text and retweets to quantify the political leaning of media outlets and prominent profiles. [16] consider users' discussions and interactions to predict stances using few annotations on any social media. These works are the most relevant ones for our task but they are focused on Twitter datasets, or limited by the fact that they require a large number of annotations or consider two main stances at most (see Table 1). In contrast, we promote a generic approach which needs significantly less annotated data to perform well, as exposed in the following sections.

Table 2: Dataset sizes and stances

Dataset	$\lvert P\rvert$	Σ	$\lvert P_{T_\sigma}\rvert$	$\lvert P_T\rvert$
SR	604,399	Yes	564	1,101
		No	537	
ME	1,718,131	Democrat	761	1,571
		Republican	810	
PE	22,843	France Insoumise	5,113	18,649
		Parti Socialiste	1,832	
		En Marche !	3,962	
		Les Républicains	4,366	
		Front National	3,376	
GC	1,420	Prefers strict gun control	312	801
		Opposes strict gun control	489	

3 DATA MODEL

To take advantage from the numerous available proximities, we represent them as proximity graphs, and exploit their structure to predict unknown stances. Formally, let $P = \{p_1, p_2, \ldots\}$ a set of profiles, Σ the set of expressed stances, and $\sigma(p_i)$ the stance of p_i. Each profile has thus one stance, which does not evolve in time. Let us consider a setting where $S \subset P$ the set a profiles with a known stance, with $|S| \ll |P|$. $A = \{A_1, \ldots, A_k\}$ is a set of attributes, with A_i^j the value of attribute A_i for the profile p_j. We define a multiplex $\mathcal{G}$, i.e., a multi-layer graph with each layer sharing the same vertex set P. Each layer is defined as a graph $G_i = (P, \mathrm{Sim}_i)$ such that Sim_i expresses the proximity between profiles in P according to attribute A_i. We consider that proximities are symetrical measures, hence G_i is undirected. The proximities used in this paper are described in Section 4.3.

Problem formulation. Given $\mathcal{G} = \{G_i, i \in [1, k]\}$ and S, is it possible to effectively determine $\sigma(p_i)$ with $p_i \in P \setminus S$?

4 PRELIMINARY STUDY

4.1 Objectives

Our model is based on the double intuition that (1) communities extracted using proximities and similar stances are correlated and (2) one single proximity is not sufficient to effectively predict stances. In this section, we provide results of an extensive analysis of four real-world datasets to show the validity of our hypotheses. Specifically, we answer the following three questions:

AQ1. Do proximity graphs have a distinct community structure?
AQ2. Are the main communities homogeneous in terms of stance?
AQ3. Are the detected communities different according to the considered proximities?

4.2 Datasets

Real-world datasets usually represent a large number of profiles that cannot be fully manually annotated. We thus focus our analysis on a subset $P_T \subset P$ of profiles. We used three Twitter datasets, and a dataset from CreateDebate.com. Table 2 presents some basic statistics on these datasets.

4.2.1 2014 Scottish Independence Referendum (SR) [11]. Profiles in P_T were part of the Scottish Independence Referendum Electoral Commission, or unambiguously indicated their stance in their Twitter biographies, giving us Σ_{SR} = ("Yes", "No").

4.2.2 2014 US Midterm Elections (ME) [11]. Stances were determined thanks to several sources listing official Twitter accounts of campaigners: Σ_{ME} = ("Democrat", "Republican").

4.2.3 2017 French Presidential Elections (PE). This Twitter dataset was collected from November 25th 2016 to May 12th 2017, by monitoring several keywords referencing the campaign. These keywords were selected by researchers familiar with the French political landscape on Twitter. Profiles in P referenced one of the five main political parties in their names or biographies, and had published at least 10 posts (including tweets or retweets) during the seven months of collection. They were manually annotated by experts. We did not consider in P_T the profiles having undefined or multiple party affiliations, hence: Σ_{PE} = ("France Insoumise", "Parti Socialiste", "En Marche !", "Les Républicains", "Front National").

4.2.4 Gun control in the US (GC) [1]. This dataset contains discussions from CreateDebate.com, a debate site. Each discussion has two possible stances, decided by its creator, which were mapped by [1] to 2 global stances: Σ_{GC} = ("prefers strict gun control", "opposes strict gun control"). For example, in a discussion titled "The Right to Bear Arms, necessary?", the stance "Yes, to defend ourselves" is mapped to "opposes strict gun control" and "No, it only creates criminals" to "prefers strict gun control". When a profile adds a post to a discussion, it has to indicate its stance, and if the post supports, clarifies, or disputes another post. Profiles also contain biographical information, and other profiles indicated as allies. The 88 profiles whose global stance changed from debate to debate were not included here.

4.3 Proximities

Many proximities belonging to various categories, e.g., textual, social, can be exploited. This flexible definition enables a generic model that can be used on any social platform. In this work, several proximities exploiting different aspects of social media were used (Table 3). Since each proximity is used separately, it is not necessary to normalize the edges' weights.

4.3.1 Content-based proximities. This type of proximities links profiles using similar textual tokens in their posts.

- Use of *keywords* (*kw*): $\text{Sim}_{kw}(p_i, p_j) = \left| A^i_{kw} \cap A^j_{kw} \right|$

with A^i_{kw} being the hashtags (for Twitter datasets), or nouns (for CreateDebate), used in p_i's publications.

- *Reference* to a piece of information (*ref*): $\text{Sim}_{ref} = \left| A^i_{ref} \cap A^j_{ref} \right|$

with A^i_{ref} the websites mentionned by p_i.[1] Urls are truncated to domain names for CreateDebate due to the fact that, of all the full urls present in the publications, only three were shared by several profiles.

4.3.2 Social-based proximities. Proximities based on social context rely on the number of social interactions. We considered the following proximities:

- *Citation* (*cite*): share of another profile's post, via retweets for Twitter and quotes for CreateDebate.
- *Call*: interpellation of another profile, i.e. mentions for Twitter, and supporting and clarifying posts for CreateDebate (see Section 4.2.4).
- *Association* (*asso*): profiles explicitly monitored, friends on Twitter and allies on CreateDebate. Only associations of the profiles in P_T were collected.

We define $A^i_{soc} = \{s^i_k\}$ as the set of interactions $s^i_k = (p_k, t_k)$, with *soc* being the considered social interaction, p_k a profile which interacted with p_i, and t_k the date of the interaction. As explained by [13], symmetric and non-symmetric relationships rationale vary greatly. To take this into account, we considered three versions of these proximities. The first one ($_{all}$) considers all interactions, while the second one ($_{rec}$) focuses only on reciprocal interactions, and the third one ($_{\overline{rec}}$) on non-reciprocal interactions. Their formal definitions are given by equations 1, 2, and 3 respectively. By construction, $A_{soc_{all}} = A_{soc_{rec}} \cup A_{soc_{\overline{rec}}}$.

$$\text{Sim}_{soc_{all}}(p_i, p_j) = \left|\{s^i_k \mid p_k = p_j\}\right| + \left|\{s^j_l \mid p_l = p_i\}\right| \quad (1)$$

$$\text{Sim}_{soc_{rec}}(p_i, p_j) = \min\left(\left|\{s^i_k \mid p_k = p_j\}\right|, \left|\{s^j_l \mid p_l = p_i\}\right|\right) \quad (2)$$

$$\text{Sim}_{soc_{\overline{rec}}}(p_i, p_j) = \left|\left|\{s^i_k \mid p_k = p_j\}\right| - \left|\{s^j_l \mid p_l = p_i\}\right|\right| \quad (3)$$

CreateDebate profiles contained more information which was used as another way to measure social similarity between profiles:

- *Socio*-demographic criteria: $\text{Sim}_{socio}(p_i, p_j) = \left| A^i_{sex} \cap A^j_{sex} \right| + \left| A^i_{school\ level} \cap A^j_{school\ level} \right| + \left| A^i_{decade} \cap A^j_{decade} \right|$
- Religious and political *beliefs*: $\text{Sim}_{belief}(p_i, p_j) = \left| A^i_{religion} \cap A^j_{religion} \right| + \left| A^i_{party} \cap A^j_{party} \right|$

4.3.3 Geographic context. This type of proximities relies on the locations signalled in profiles. The construction is similar for *city*, *region*, and *country*: $\text{Sim}_{city}(p_i, p_j) = \left| A^i_{city} \cap A^j_{city} \right|$
Since variations in format made it necessary to manually clean the entries, for Twitter, only locations of the profiles in P_T were used.

4.4 Metrics

To answer **AQ1**, we compute the *transitivity* of our proximity graphs. The transitivity measures the probability that two neighbors of a vertex are connected, and the higher it is, the stronger the community structure in the graph is [29]. It is the average weighted clustering coefficient [6]. To answer **AQ2**, we measure the homogeneity of the communities for each proximity graph. We used the average intra-community *purity* [21]. Communities are extracted using label propagation [32]. It has been shown to be efficient for this task, despite not being deterministic [20]. To measure how much information was shared between proximities and thus answer **AQ3**, we computed the *Normalized Mutual Information* [15].

4.5 Findings

AQ1. Table 4 presents the transitivity of our proximities. Most proximities have a non-negligible transitivity, suggesting an underlying community structure, more or less pronounced depending on the dataset. Despite being often successfully used in stance detection models, all versions of *cite* exhibit a low or medium transitivity.
AQ2. To confirm that some proximities give us homogeneous communities in terms of stance, we measured the average purity of the 10 biggest communities containing at least 5 profiles in P_T (see Table 5). While this is an incomplete picture – we have a median of 50% of annotated profiles – we can see that some proximities have extremely high purity. On Twitter datasets, $cite_{all}$, $cite_{rec}$, and $asso_{all}$ consistently obtain a mean purity close to or higher than 0.90. *Ref* and $call_{rec}$ are not as efficient but still give good results. On the other hand, scores for $call_{all}$, $asso_{rec}$, and $asso_{\overline{rec}}$ vary a lot depending on the considered dataset.
AQ3. Table 6 presents the normalized mutual information between proximities. The first observation we can make is that the relationships between proximities is heavily dataset-dependent. In most cases, each proximity brings unique information about profiles in regards to others. The proximities having similar communities structure are not surprising: *city* and *region* are often strongly related,

[1] Urls coming from shortening services, such as `bit.ly` or `goo.gl`, were expanded for an optimal matching when possible.

Table 3: Number of edges in $\mathbf{Sim}_i$. avg_w represents their average weight, and $\max_w$ their maximum weight.

Interaction	*SR*			*ME*			*PE*			*GC*		
	$\lvert\text{Sim}_i\rvert$	avg_w	$\max_w$	$\lvert\text{Sim}_i\rvert$	avg_w	$\max_w$	$\lvert\text{Sim}_i\rvert$	avg_w	$\max_w$	$\lvert\text{Sim}_i\rvert$	avg_w	$\max_w$
ref	315,326	16	3,257,193	16,865	47	169,368	17,367,840	10	99,704	74	3	32
kw	422,767	5,142	5,319,272	1,304,290	24	222,379	78,440,288	80	29,006	236,976	49	29,854
$cite_{all}$	1,426,966	1	324	713,052	2	672	1,262,484	3	7,983	128	6	68
$cite_{rec}$	11,911	1	65	7,986	2	73	123,853	7	7,983	23	8	53
$call_{all}$	335,191	3	1,296	66,179	3	377	1,787,127	5	10,473	226	2	14
$call_{rec}$	24,205	2	1,255	4,351	2	356	227,400	8	8,658	48	2	11
$asso_{all}$	1,914,298	1	1	3,330,134	1	1	3,135,200	1	1			
$asso_{rec}$	1,740,077	1	1	2,817,081	1	1	1,532,825	1	1	4,034	1	1
$asso_{\overline{rec}}$	295,315	1	1	473,271	1	1	1,602,375	1	1			
socio										312,795	1	3
beliefs										240,323	1	2
city	11,833,595	1	1	2,023	1	1	2,929,622	1	1			
region	12,196,762	1	1	20,447	1	1	7,780,336	1	1			
country										445,407	1	1

Table 4: Proximity graphs transitivity

	ref	*kw*	$cite_{all}$	$cite_{rec}$	$call_{all}$	$call_{rec}$	$asso_{all}$	$asso_{rec}$	$asso_{\overline{rec}}$	*socio*	*beliefs*	*city*	*region*	*country*
SR	0.61	0.95	0.02	0.09	0.15	0.11	0.16	0.16	0.03			1.00	1.00	
ME	0.44	0.79	0.06	0.12	0.11	0.04	0.12	0.10	0.03			0.73	1.00	
PE	0.69	0.87	0.38	0.24	0.48	0.24	0.41	0.31	0.26			0.95	0.99	
GC	0.59	0.91	0.03	0.07	0.05	0.01		0.09		0.80	0.82			0.98

Table 5: Average purity of the 10 biggest communities containing at least 5 profiles in P_T. Values above 0.80 are bolded for readability.

	ref	*kw*	$cite_{all}$	$cite_{rec}$	$call_{all}$	$call_{rec}$	$asso_{all}$	$asso_{rec}$	$asso_{\overline{rec}}$	*socio*	*beliefs*	*city*	*region*	*country*
SR	**0.82**	0.52	**0.94**	**1.00**	0.52	0.79	**0.98**	**0.99**	0.75			0.61	0.61	
ME	**0.84**	0.62	**0.99**	**0.98**	0.78	**0.96**	**0.94**	0.52	0.52			0.69	0.58	
PE	0.68	0.25	**0.93**	**0.95**	**0.93**	**0.85**	**0.87**	0.79	**0.84**			0.33	0.30	
GC	0.76	0.63	0.67	0.60	0.66	0.66		0.66		0.64	0.64			0.62

as well as reciprocal versions of proximities with their complete versions. There is a lot more of redundant information between proximities on *ME*, with $call_{all}$, $cite_{all}$, and *ref* being closely related. This is interesting since we do not observe this phenomenon on the other Twitter datasets.

4.6 Implications for Stance Detection

The results of these experiments demonstrate that communities detected on social media elements can yield extremely high homogeneity in terms of stance, and therefore be an effective way to propagate stance from some known profiles. Moreover, as indicated by moderate NMI values, the communities extracted from the different considered proximities look different. This suggests that each one brings a specific piece of information about the profiles entourage, allowing for a better characterization. Unsurprisingly, reciprocal versions of the proximities are semantically close to their complete versions (we see high NMI scores between the pairs) but their higher purities may be of interest for our task.

Even when extracted from the same platform, each dataset has its own particularities. Indeed we can see that some proximities can be useful or hurtful depending on the dataset, and that the similarity between proximities varies across datasets. On Twitter, $cite_{all}$ and $asso_{all}$ seem particularly encouraging for our task: they have very homogeneous communities and bring unique information compared to other proximities (apart from their reciprocal version). On CreateDebate, *ref* and $asso_{rec}$ seem interesting for the same reasons. These measures could help us determine which proximity to discard in order to optimize our process, but for the time being we will consider all the defined proximities.

5 SEQUENTIAL COMMUNITY-BASED STANCE DETECTION MODEL

Based on the previous results, we propose the Sequential Community-based Stance Detection model, or **SCSD**. Algorithm 1 exposes the main mecanism of our model, i.e., the assignation of one stance per community iteratively: first, different sets of communities are built according to different profile proximities, then a specific function is used to order them, and a set of seed profiles is selected from these communities. Finally, an iterative stance detection step is performed based on these seed profiles to find the stance of the remaining profiles in P. This algorithm features two crucial elements:

(1) the order in which the proximities are considered,
(2) and the profiles selected as seeds.

The following sections present our answers to these questions.

Algorithm 1 : SCSD framework – $X = (x_1, \ldots, x_n)$ is the sequence of proximities to be used. The functions designed to order the proximities and to select the seed profiles are presented in Section 5.1 and Section 5.2 respectively. The algorithm of getMajorityStance is detailed in Algorithm 2.

```
for x_i in X do                                          ▷ Initialisation
    C_{x_i} ← detectCommunities(x_i)
X^ord ← orderProximities(X, ω)
S ← selectSeedProfiles(x_1^ord, s, s_com, s_min, φ)
for p_i in S do
    P(p_i) ← getGroundTruth(p_i)
for x_i^ord in X^ord do                                  ▷ Stance detection process
    for c in C_{x_i^ord} do
        for p_i in c do
            if σ(p_i) is undefined then
                σ(p_i) ← getMajorityStance(c)
```

5.1 Proximities ordering

In SCSD, a profile's stance cannot change once it has been assigned. The order in which the proximities are considered in the stance detection loop is thus crucial. The ordering function ω computes the ordered sequence of proximities to be used during the rest

Table 6: Normalized Mutual Information between proximity communities for each dataset. Values above 0.80 are bolded for readability.

(a) *SR*

	region	*city*	$asso_{\overline{rec}}$	$asso_{rec}$	$asso_{all}$	$call_{rec}$	$call_{all}$	$cite_{rec}$	$cite_{all}$	*ref*
kw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ref	0.25	0.29	0.29	0.26	0.27	0.37	0.25	0.38	0.27	
$cite_{all}$	0.21	0.22	0.36	0.60	0.57	0.37	0.49	0.52		
$cite_{rec}$	0.49	0.52	0.38	0.54	0.53	0.69	0.40			
$call_{all}$	0.22	0.21	0.43	0.51	0.47	0.32				
$call_{rec}$	0.51	0.50	0.40	0.38	0.36					
$asso_{all}$	0.22	0.24	0.48	**0.84**						
$asso_{rec}$	0.21	0.23	0.47							
$asso_{\overline{rec}}$	0.28	0.29								
city	**0.97**									

(b) *ME*

	region	*city*	$asso_{\overline{rec}}$	$asso_{rec}$	$asso_{all}$	$call_{rec}$	$call_{all}$	$cite_{rec}$	$cite_{all}$	*ref*
kw	0.35	0.40	0.29	0.18	0.21	0.45	0.39	0.48	0.38	0.42
ref	0.59	0.71	0.41	0.21	0.37	**0.87**	**0.81**	**0.88**	0.73	
$cite_{all}$	0.56	0.65	0.37	0.22	0.40	0.79	0.73	**0.81**		
$cite_{rec}$	0.71	**0.81**	0.44	0.23	0.41	**0.96**	**0.93**			
$call_{all}$	0.66	0.77	0.38	0.20	0.34	**0.94**				
$call_{rec}$	0.71	**0.82**	0.41	0.22	0.39					
$asso_{all}$	0.24	0.29	0.49	0.59						
$asso_{rec}$	0.19	0.19	0.43							
$asso_{\overline{rec}}$	0.32	0.33								
city	**0.90**									

(c) *PE*

	region	*city*	$asso_{\overline{rec}}$	$asso_{rec}$	$asso_{all}$	$call_{rec}$	$call_{all}$	$cite_{rec}$	$cite_{all}$	*ref*
kw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ref	0.02	0.05	0.51	0.51	0.48	0.24	0.51	0.30	0.54	
$cite_{all}$	0.03	0.08	0.62	0.67	0.63	0.38	0.76	0.46		
$cite_{rec}$	0.31	0.49	0.37	0.43	0.40	0.78	0.44			
$call_{all}$	0.03	0.08	0.60	0.66	0.62	0.38				
$call_{rec}$	0.28	0.45	0.30	0.37	0.33					
$asso_{all}$	0.02	0.07	0.62	0.73						
$asso_{rec}$	0.04	0.09	0.63							
$asso_{\overline{rec}}$	0.02	0.06								
city	0.70									

(d) *GC*

	country	*beliefs*	*socio*	$asso_{rec}$	$call_{rec}$	$call_{all}$	$cite_{rec}$	$cite_{all}$	*ref*
kw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ref	0.25	0.26	0.00	0.10	0.67	0.51	0.50	0.54	
$cite_{all}$	0.33	0.27	0.00	0.34	0.74	0.72	**0.94**		
$cite_{rec}$	0.27	0.27	0.00	0.26	0.73	0.70			
$call_{all}$	0.26	0.20	0.02	0.31	**0.98**				
$call_{rec}$	0.34	0.18	0.05	0.43					
$asso_{rec}$	0.09	0.05	0.00						
socio	0.03	0.01							
beliefs	0.15								

of the model, X^{ord}. The ideal strategy is to order them in order of decreasing like-mindedness of the communities. Since, in most cases, it is impossible to know this element beforehand, let us discuss two options to design this ordering function ω:

Manual ordering function. This mode is useful if the user already has expertise on the dataset, and when he wants to choose the order in which the proximities are to be used.

Automatic ordering function. The model can also automatically arrange proximities according to a user-defined function. Some examples are:

Algorithm 2 : getMajorityStance – c is the community whose majority stance must be determined.

```
stances ← ∅                                   ▷ Table mapping stances with frequencies
for p_i in c do
    if σ(p_i) is defined then
        stances[σ(p_i)].increment()
sorted ← sortByDecreasingFreq(stances)
if stances ≠ ∅ and sorted[0].freq ≠ sorted[1].freq then
    return sorted[0].stance
else
    return undefined
```

Mod Ordering by decreasing modularity.
NbC$_{Asc}$ Ordering by ascending number of communities.
NbC$_{Desc}$ Ordering by descending number of communities.
NbP$_{Asc}$ Ordering by ascending number of profiles.
NbP$_{Desc}$ Ordering by descending number of profiles.

Section 6.3 presents a comparison of these functions.

5.2 Seed selection

Algorithm 3 : Seed selection – s is the size of the seed, s_{com} the minimum number of seed communities to consider, s_{min} the minimum number of seed profiles in each seed community, and φ the importance function determining how the seed profiles are selected.

```
seedP ← ∅                                     ▷ Selected seed profiles
seedC ← ∅                                     ▷ Selected seed communities
nbPossibleP ← 0                               ▷ Number of possible seed profiles
                                              ▷ in selected seed communities
for c in sortByDecreasingSize(C_{x_1^ord}) do
    if |c| ≥ s_min then
        seedC ← seedC ∪ c
        nbPossibleP ← nbPossibleP + |c|
    if |seedC| ≥ s_com and nbPossibleP ≥ s then
        break
for c in seedC do
    nbProfiles ← min(|c|, max(s_min, |c| × s / nbPossibleP))
    newSeedProfiles ← selectSeedProfiles(c, nbProfiles, φ)
    seedP ← seed P ∪ newSeedProfiles
return seedP
```

We now detail our strategy for seed selection. Indeed, manually annotating data is time-consuming and expensive. SCSD is designed to be efficient with a very small seed (i.e., less than 5% of the number of profiles to classify), and seed selection is dependent on the global cost of annotation s the final user wishes to put into it: s is the size of our seed S, i.e., the total number of seed profiles to manually annotate. *Seed profiles* are profiles contained in the seed, and *seed communities* the communities containing seed profiles.

In order to accurately detect the $|\Sigma|$ stances present in the considered dataset, we sample at least s_{com} seed communities, with $s_{com} \geq |\Sigma|$. The number of seed profiles in each seed community is proportional to its size, and each seed community contains at least s_{min} seed profiles. The importance function φ indicates the method of selection for the seed profiles. It can be deterministic or not – random for example. Algorithm 3 details this process.

The importance functions translate greater importance by larger values, and can be divided into 2 families. The first family is based on raw information from the profiles, for example:

- For Twitter datasets: *number of followers* (#Fo), *number of friends* (#Fr), *number of posts* (#P), *number of retweets of the profile's posts* (#Rt), *seniority* (Se).
- For the CreateDebate dataset: *efficiency* (E), *number of debates* (#D), *number of posts*, *number of relations* (#R), *seniority*.

The second one is based on the proximity graphs: *degree* (D), *strength* (St), *pagerank* (Pr) [30], *closeness centrality* (CC) [8], *eigenvector centrality* (EC) [28]. Section 6.1 presents a comparison of these functions.

6 EXPERIMENTS

In this section we examine the effectiveness of the SCSD model. We use standard metrics in a classification settings, namely the macro-averaged *precision*, *recall*, and *F1-score*, computed on the subset P_T defined in Section 4.2. We consider several questions in order to measure the impact of the different components of the SCSD model:

Q1. What is the influence of the importance function φ, used in the seed selection step, on the performance?
Q2. What is the contribution of each proximity?
Q3. How well do the automatic ordering functions perform compared to a manual ordering of proximities?
Q4. How robust is SCSD with regards to seed size variations?
Q5. How does SCSD perform compared to baselines?

6.1 Comparison of importance functions φ

6.1.1 Study of compared importance functions. We suppose that some importance functions are heavily correlated to others. In order to consider relevant functions only, we study their pairwise similarities. Table 7 presents the Spearman correlation [18] between profiles ranks returned by the importance functions indicated in Section 5.2. As expected, a lot of functions seem to be redundant. *Degree* (D), *pagerank* (Pr), and *strength* (St) returns similar profiles, as shown by their high correlations. It is not surprising given their definitions and the existing literature [19]. The *closeness* (CC) and *eigenvector centralities* (EC) are close to each other, suggesting their evaluation of the profiles' importance is also similar. The *number of posts* (#P) is often correlated to the *number of retweets* (#Rt) and of *followers* (#Fo) for Twitter, and to the *number of debates* (#D) and *relations* (#R) for CreateDebate. For the Twitter datasets, the numbers of *followers* and *friends* (#Fr) returns a majority of similar profiles. This is probably due to the atypical profiles (i.e. public figures) having a large number of friends and of followers. Based on these observations, we can omit some functions from our analysis. We will only consider *degree*, *closeness centrality*, *number of posts*, and *seniority* for the remaining analyses.

6.1.2 Influence of importance function on seed selection. In order to compare the importance functions in an optimal setting, we chose to manually order the proximities. We built an optimal sequence per dataset: the proximities are manually ordered by decreasing purity (see Table 5). The first observation we can make is that, with the correct proximities ordering, we successfully assign stance to a large majority of profiles. And, surprisingly, the scores do not vary while using different functions (see Table 8). This is probably a consequence of seed communities being so homogeneous in terms of stance that the importance metric used does not make a difference.

This result has a very interesting practical application: once the seed communities, and how many profiles to pick in each one, has been determined, it is possible to select profiles from which the stance is known beforehand. This is particularly useful when used by an expert user, since there is a high probability that he / she already knows the stance of a fraction of the profiles in the studied dataset.

Table 7: Spearman correlation between profiles ranks returned by different importance functions φ – values above 0.50 are bolded for readability.

(a) *SR*

	EC	CC	Pr	St	D	Se	#Fo	#Fr	#Rt
#P	0.22	0.29	**0.52**	0.48	**0.51**	0.47	**0.58**	**0.57**	**0.71**
#Rt	**0.53**	**0.59**	**0.61**	**0.73**	**0.75**	0.20	0.42	0.47	
#Fr	0.23	0.23	0.27	0.27	0.30	0.34	**0.60**		
#Fo	0.16	0.21	0.39	0.33	0.34	0.42			
Se	-0.13	-0.09	0.12	0.02	0.03				
D	**0.66**	**0.75**	**0.88**	**0.97**					
St	**0.65**	**0.70**	**0.89**						
Pr	0.37	0.45							
CC	**0.92**								

(b) *ME*

	EC	CC	Pr	St	D	Se	#Fo	#Fr	#Rt
#P	0.10	0.14	0.24	0.24	0.24	**0.62**	**0.80**	**0.64**	**0.50**
#Rt	0.36	0.43	**0.74**	**0.79**	**0.76**	0.21	0.32	0.38	
#Fr	0.01	0.04	0.13	0.11	0.14	0.48	**0.58**		
#Fo	0.01	0.02	0.03	0.02	0.02	**0.63**			
Se	0.03	0.06	-0.01	0.02	0.01				
D	0.45	**0.59**	**0.90**	**0.94**					
St	0.45	**0.57**	**0.89**						
Pr	0.25	0.34							
CC	**0.85**								

(c) *PE*

	EC	CC	Pr	St	D	Se	#Fo	#Fr	#Rt
#P	0.23	0.40	0.23	0.36	0.39	0.47	**0.69**	**0.53**	0.25
#Rt	**0.52**	**0.59**	0.37	**0.76**	**0.75**	0.02	0.40	0.31	
#Fr	0.10	0.25	0.18	0.29	0.29	0.39	**0.78**		
#Fo	0.09	0.25	0.26	0.36	0.36	0.45			
Se	0.06	0.05	0.01	0.07	0.06				
D	**0.68**	**0.80**	0.35	**0.88**					
St	**0.56**	**0.61**	**0.52**						
Pr	-0.10	-0.03							
CC	**0.80**								

(d) *GC*

	EC	CC	Pr	St	D	Se	E	#R	#D
#P	0.21	0.20	0.28	0.37	0.23	0.08	-0.21	**0.78**	**0.79**
#D	0.12	0.05	0.22	0.46	0.09	0.09	-0.19	**0.71**	
#R	0.13	0.27	0.14	0.22	0.09	0.16	-0.03		
E	0.02	-0.24	-0.07	-0.08	-0.31	-0.40			
Se	**-0.51**	0.01	0.31	0.06	-0.06				
D	**0.51**	**0.52**	0.48	**0.55**					
St	0.46	-0.06	**0.68**						
Pr	0.00	-0.17							
CC	0.41								

Table 8: SCSD model scores using $(s, s_{com}, s_{min}) = (3\%\times|P_T|, 3\times|\Sigma|, 3)$ and ω = *Manual*. These scores are identical when using the following importance functions φ: *degree*, *closeness centrality*, *number of posts*, *seniority*, and *random* (average scores on 10 runs).

Dataset	*s*	p	r	F1
SR	*33*	0.95	0.95	0.95
ME	*47*	0.95	0.95	0.95
PE	*560*	0.90	0.86	0.87
GC	*24*	0.58	0.52	0.45

6.2 Contribution of each proximity

In order to measure the contribution of each considered proximity, we computed scores for SCSD-Basic, a simplified version of SCSD with $X = (x)$ for each proximity x, with $(s, s_{com}, s_{min}, \varphi) = (3\%\times|P_T|, 3\times|\Sigma|, 3, \textit{Degree})$. We note that despite the small seed

Table 9: `SCSD-Basic` model scores with $(s, s_{com}, s_{min}, \varphi) = (3\%\times |P_T|, 3\times|\Sigma|, 3, Degree)$.

	SR			ME			PE			GC		
	p	r	F1	p	r	F1	p	r	F1	p	r	F1
ref	0.51	0.37	0.33	0.87	0.15	0.25	0.48	0.51	0.45	0.38	0.02	0.03
kw	0.25	0.46	0.32	0.57	0.34	0.38	0.05	0.13	0.07	0.31	**0.50**	**0.39**
$cite_{all}$	0.91	0.71	0.78	0.97	0.29	0.37	0.91	0.84	0.87	0.58	0.04	0.08
$cite_{rec}$	**1.00**	0.34	0.50	**1.00**	0.03	0.06	**0.96**	0.22	0.35	**0.83**	0.02	0.04
$call_{all}$	0.25	0.42	0.31	0.72	0.02	0.04	0.90	**0.88**	**0.89**	0.60	0.05	0.09
$call_{rec}$	0.78	0.23	0.35	0.94	0.03	0.06	0.84	0.27	0.39	0.72	0.03	0.05
$asso_{all}$	0.97	0.84	0.90	0.97	**0.81**	**0.88**	0.89	0.74	0.79			
$asso_{rec}$	0.99	**0.85**	**0.91**	0.25	0.43	0.31	0.89	0.70	0.78	0.59	0.17	0.24
$asso_{\overline{rec}}$	0.55	0.39	0.31	0.25	0.38	0.30	0.67	0.63	0.63			
socio										0.39	0.21	0.25
beliefs										0.56	0.21	0.27
city	0.60	0.08	0.14	0.64	0.05	0.09	0.41	0.06	0.08			
region	0.56	0.10	0.16	0.56	0.14	0.22	0.37	0.10	0.12			
country										0.64	0.23	0.25

Table 10: Comparison of ordering functions on SCSD model F1-scores, using $(s, s_{com}, s_{min}, \varphi) = (3\%\times|P_T|, 3\times|\Sigma|, 3, Degree)$.
Scores presented for *Random* are the average scores on 10 runs.

	Random	*Manual*	*Mod*	NbC_{Asc}	NbC_{Desc}	NbP_{Asc}	NbP_{Desc}
SR	0.67	**0.95**	0.49	0.38	0.79	0.50	0.90
ME	0.53	**0.95**	0.50	0.43	0.40	0.55	0.91
PE	0.63	0.87	0.83	0.11	0.74	0.21	**0.89**
GC	0.43	0.45	0.47	0.43	**0.48**	0.41	0.34

size, some proximities obtain extremely high precision scores, but they usually retrieve few profiles, leading to a poor recall. This phenomenon is less present in *PE*, where we see that we can obtain excellent results in detecting the five different stances. The weak scores obtained by $asso_{rec}$ on *ME* are surprising. After further investigation, this is due to the fact that on this dataset, profiles are highly connected using this proximity. When detecting communities, the majority of profiles in P_T are assigned to the same cluster, leading to one stance being largely ignored during the classification process. Despite gun control being often presented as a "right-wing versus left-wing" debate, or like a generational conflict, the *socio* and *beliefs* proximities did not perform well. This is probably due to the small size of the dataset, and to the profiles in *GC* not being representative of the US society.

6.3 Comparison of ordering functions ω

The order of the proximities is a key element of the model. The interactions providing the most homogeneous communities should be used first in order to obtain the best results. We decided to compare the performances of our model using different proximity sequences, given by the ordering functions ω presented in Section 5.1, with $(s, s_{com}, s_{min}, \varphi) = (3\%\times|P_T|, 3\times|\Sigma|, 3, Degree)$.[2] We also consider a random ordering of the proximities (whose scores are averaged on 10 runs). Figure 1 presents the evolution of precision and recall during the process, and Table 10 the final F1-scores. For Twitter, we see that $NbProfiles_{Desc}$ is a strong contender to the manual ordering. For *SR* and *PE*, $nbCom_{Desc}$ could be an acceptable alternative, probably because a higher number of communities means smaller, more homogeneous ones on these datasets. For *GC*, *modularity* and $nbCom_{Desc}$ give us slightly better results than the manual ordering, and only $nbProfiles_{Desc}$ gives us significantly weakened scores. Contrary to our initial intuition that modularity could be a good indicator for proximities ordering, its results shows it vastly depends on the studied dataset. Good results characterise *PE* and *GC*, but we see a great loss of performance on *SR* and *ME*.

[2]Given the results presented in Section 6.1, we use only one importance function for the remaining analyses.

Table 11: Influence of s and s_{com} on SCSD model F1-scores, using $(s_{min}, \varphi) = (3, Degree)$ and $\omega = Manual$.
The missing values represent cases where the seed selection is impossible because $s < s_{com} \times s_{min}$.

s	$1\%\times\|P_T\|$		$3\%\times\|P_T\|$		$5\%\times\|P_T\|$		$7\%\times\|P_T\|$	
s_{com}	$1\times\|\Sigma\|$	$2\times\|\Sigma\|$	$1\times\|\Sigma\|$	$2\times\|\Sigma\|$	$1\times\|\Sigma\|$	$2\times\|\Sigma\|$	$1\times\|\Sigma\|$	$2\times\|\Sigma\|$
SR	0.95		0.95	0.95	0.95	0.95	0.95	0.95
ME	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
PE	0.47	0.87	0.47	0.87	0.47	0.88	0.48	0.88
GC	0.29		0.45	0.45	0.45	0.45	0.45	0.45

Table 12: Baselines scores.

		`SVM`	$\texttt{RF}_{cite}$	$\texttt{RF}_{asso}$	$\texttt{LP}_{cite}$	$\texttt{LP}_{asso}$	`SCSD`
		Annotations: 80%			*Annotations: 3%*		
SR	p	0.92	0.94	**0.95**	0.89	**0.95**	**0.95**
	r	0.92	0.91	0.94	0.76	0.93	**0.95**
	F1	0.92	0.90	0.94	0.78	0.94	**0.95**
ME	p	0.88	0.93	**0.96**	0.20	0.11	0.95
	r	0.87	0.92	0.93	0.33	0.08	**0.95**
	F1	0.87	0.92	0.93	0.24	0.07	**0.95**
PE	p	0.75	**0.91**	0.87	0.21	0.11	0.90
	r	0.76	**0.89**	0.82	0.19	0.09	0.88
	F1	0.75	**0.89**	0.83	0.20	0.10	**0.89**
GC	p	0.43	**0.62**	0.55	0.45	0.40	0.57
	r	0.37	**0.54**	0.51	0.04	0.17	0.53
	F1	0.37	**0.51**	0.50	0.08	0.19	0.48

6.4 Influence of seed size

We then looked at the influence of the seed size and the number of initial communities on scores. Table 11 presents the F1-scores of the SCSD model when s and s_{com} vary, using $(s_{min}, \varphi) = (3, Degree)$ and $\omega = Manual$. With our compared configurations, *SR* and *ME* do not vary, however *PE* and *GC* show the influence of s and s_{com}. *GC* shows degraded performances when considering a minimal seed, i.e. $(s, s_{com}) = (1\%\times|P_T|, 1\times|\Sigma|)$. For *PE*, s_{com} seems to hold more importance than s, since with $s_{com} = 1\times|\Sigma|$ we see a drastic drop in F1-score which is not compensated by the increase in s, while we obtain good results with $s_{com} = 2\times|\Sigma|$ even when taking into account only 1% of annotated profiles. Finally, these results confirm that SCSD model is built to be effective with a very small seed, the benefit given by a seed larger than 3% of the number of profiles being non-existent.

6.5 Comparison with baselines

We compare the performance of the SCSD model to several baselines traditionally used for stance detection. For the social aspect, we use the two main proximities used for this task: $cite_{all}$ and $asso_{rec}$.[3] These baselines being supervised models, we use a 5-fold cross-validation.

`SVM` We use an SVM model based on the concatenation of each profile's posts, with a vocabulary consisting of the 10,000 most distinctive tokens according to χ^2 stats.

$\texttt{RF}_{cite}$ For the social aspect, we build a Random Forest classifier based on the $cite_{all}$ proximity.

$\texttt{RF}_{asso}$ Similar to $\texttt{RF}_{cite}$ but using the $asso_{rec}$ proximity.

[3]We use $asso_{rec}$ because $asso_{all}$ is undefined for *GC*, the ally interaction being inevitably reciprocal.

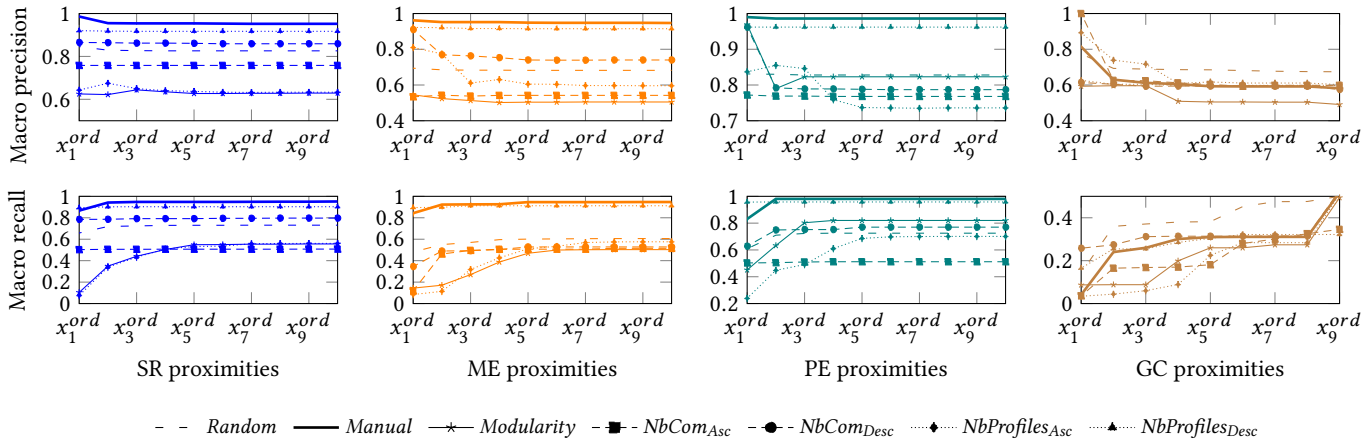

Figure 1: Comparison of ordering functions on SCSD model precision and recall evolution, using $(s, s_{com}, s_{min}, \varphi) = (3\%\times|P_T|, 3\times|\Sigma|, 3, Degree)$**.** Scores presented for *Random* are the mean scores on 10 runs.

These baselines being supervised, they need a lot more of annotated data than the SCSD model. In order to compare to models using the same amount of annotated data, we define the following baselines:

LP_{cite} A semi-supervised label propagation process using the $cite_{all}$ proximity. The seed selection being random, we present the average scores on 10 runs.

LP_{asso} Similar to LP_{cite} but using the $asso_{rec}$ proximity.

Table 12 presents the results of SCSD using the optimal configurations, deduced from previous experiments, compared to our baselines. We can see that SCSD obtains, with at least 30 times less annotated data, scores higher than or close to our supervised baselines. When comparing to the semi-supervised baselines using the same amount of annotated data, it has an average gain in F1-score of 49 points (ranging between 1 and 88 percentage points). For *ME*, the low scores obtained by LP_{asso} are probably linked to the overabundance of links observed in Section 6.2. This is a problem posed by relying on a unique proximity: if it is not adapted to the studied dataset, there is no way to rectify the situation. Since we observed high purities for $cite_{all}$ and $asso_{rec}$ on both *SR* and *PE*, the differences in scores are probably due to the multiple stances of *PE* which are a lot harder for label propagation to deal with than the simple bipartite problematic represented by *SR*. We see that for all models, *GC* profiles were significantly harder to classify, probably because of the small size of the dataset.

6.6 Discussion

The analysis of the different proximities supports previous findings in the literature. The precision scores of $cite_{all}$ and *ref* suggest people tend to construct their discourse on social media by sharing arguments they agree on rather than refuting opposing ones. Note that while people can call out their opponents, they tend to engage a lot more to their allies. This is demonstrated by the better precision usually obtained by $call_{rec}$ compared to $call_{all}$. The $asso_{all}$ precision tends to demonstrate that they also select the profiles they follow [3, 14]. Interestingly, and contrary to the observation made above, another behaviour appears. Some profiles, mainly profiles of public figures, decided to follow their opponents as well as their allies. This is probably a way of monitoring their actions and discourse, suggesting that on Twitter, retweets and following are not always endorsements. This phenomenon is visible in the 2014 US midterms elections, when profiles are so closely linked that they are assigned to the same cluster.

The results seem to indicate that the majority of keywords are more related to the topic than to a specific stance. Moreover, the importance of the geographical proximity appears to vary depending on the topic. Unfortunately it was not possible to investigate all possible granularities for the geographical proximity. In addition, finer granularities were not always provided and inferring which granularity would be the most effective for a given topic is not an easy task. Since the focus of the study was stance detection, we used a naive version of geographical proximity, but this issue requires further attention.

7 CONCLUSION

Our aim was to detect profiles' stance using their likeliness to other profiles with a very small seed, to reduce annotation costs. We proposed SCSD, a generic semi-supervised model which can easily be customized to suit any requirement. The results of this study on 4 corpora suggest that it is possible to accurately predict stance using this method, even when using as little as 1% of annotated profiles and considering more than two stances. Moreover, they showed that since all proximities do not carry as much information in terms of stance, using several proximities allows to strengthen stance assignment. We do think this model could be a great help for computational social scientists wishing to exploit large datasets without having the resources to manually annotate them. The basic principles do not require advanced technical skills to grasp. Moreover, social scientists usually have the expert knowledge needed to order the proximities in an optimal manner according to the platform they wish to study.

The present study did not investigate the whole Twitter follow graphs due to their computational cost, and the excellent performance of more focused graphs. It would be interesting in the future to compare the performance of follow graphs to the friends graphs. A more complete and finer use of the textual content would surely be helpful as well, since keywords alone are not easily exploitable.

REFERENCES

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go with It. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), 23–28.

[2] Lada A. Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proceedings of the 3rd International Workshop on Link Discovery*. ACM Press, 36–43. https://doi.org/10.1145/1134271.1134277

[3] Jisun An, Daniele Quercia, and Jon Crowcroft. 2013. Fragmented Social Media: A Look into Selective Exposure to Political News. In *Companion Proceedings of the 22nd International Conference on World Wide Web*. ACM Press, 51–52. https://doi.org/10.1145/2487788.2487807

[4] Pablo Barberá. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* 23, 01 (2015), 76–91. https://doi.org/10.1093/pan/mpu011

[5] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science* 26, 10 (2015), 1531–1542. https://doi.org/10.1177/0956797615594620

[6] A. Barrat, M. Barthélemy, R Pastor-Satorras, and A Vespignani. 2004. The Architecture of Complex Weighted Networks. *Proceedings of the National Academy of Sciences* 101, 11 (2004), 3747–3752. https://doi.org/10.1073/pnas.0400087101

[7] Carl Julien Barrelet, Sebnem Sahin Kuzulugil, and Ayşe Başar Bener. 2016. The Twitter Bullishness Index: A Social Media Analytics Indicator for the Stock Market. In *Proceedings of the 20th International Database Engineering & Applications Symposium (IDEAS '16)*. 394–395. https://doi.org/10.1145/2938503.2938508

[8] Alex Bavelas. 1950. Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America* 22, 6 (1950), 725–730. https://doi.org/10.1121/1.1906679

[9] Michaël Boireau. 2014. Determining Political Stances from Twitter Timelines: The Belgian Parliament Case. In *Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia*. ACM Press, 145–151. https://doi.org/10.1145/2729104.2729114

[10] danah m. boyd and Nicole B. Ellison. 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13, 1 (2007), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

[11] Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2015. Analyzing Discourse Communities with Distributional Semantic Models. In *Proceedings of the ACM Web Science Conference*. ACM Press, 1–10. https://doi.org/10.1145/2786451.2786470

[12] Ming Cheung, Xiaopeng Li, and James She. 2017. An Efficient Computation Framework for Connection Discovery Using Shared Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 13, 4 (2017), 58:1–58:21. https://doi.org/10.1145/3115951

[13] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter. *Journal of Communication* 64, 2 (2014), 317–332. https://doi.org/10.1111/jcom.12084

[14] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. 2011. Political Polarization on Twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. 89–96.

[15] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing Community Structure Identification. *JSTAT* 09 (2005), P09008. https://doi.org/10.1088/1742-5468/2005/09/P09008

[16] Rui Dong, Yizhou Sun, Lu Wang, Yupeng Gu, and Yuan Zhong. 2017. Weakly-Guided User Stance Prediction via Joint Modeling of Content and Social Interaction. 1249–1258. https://doi.org/10.1145/3132847.3133020

[17] Shiri Dori-Hacohen, David Jensen, and James Allan. 2016. Controversy Detection in Wikipedia Using Collective Classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 797–800. https://doi.org/10.1145/2911451.2914745

[18] E. C. Fieller, H. O. Hartley, and E. S. Pearson. 1957. Tests for Rank Correlation Coefficients. I. *Biometrika* 44, 3-4 (1957), 470–481. https://doi.org/10.1093/biomet/44.3-4.470

[19] Santo Fortunato, Marián Boguñá, Alessandro Flammini, and Filippo Menczer. 2008. Approximating PageRank from In-Degree. In *Algorithms and Models for the Web-Graph*. Vol. 4936. Springer Berlin Heidelberg, 59–71.

[20] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besancon, and Mohand Boughanem. 2017. Uncovering Like-minded Political Communities on Twitter. In *International Conference on the Theory of Information Retrieval (ICTIR), Amsterdam, 01/10/17-04/10/17*. ACM.

[21] M. Girvan and M. E. J. Newman. 2002. Community Structure in Social and Biological Networks. *NAS* 99, 12 (2002), 7821–7826. https://doi.org/10.1073/pnas.122653799

[22] Kazi Saidul Hasan and Vincent Ng. 2013. Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 1348–1356.

[23] Shanto Iyengar and Sean J. Westwood. 2015. Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science* 59, 3 (2015), 690–707. https://doi.org/10.1111/ajps.12152

[24] Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. #ISISisNotIslam or #DeportAllMuslims?: Predicting Unspoken Views. In *Proceedings of the 8th ACM Conference on Web Science*. ACM Press, 95–106. https://doi.org/10.1145/2908131.2908150

[25] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

[26] Pasquale De Meo, Katarzyna Musial-Gabrys, Domenico Rosaci, Giuseppe M. L. Sarnè, and Lora Aroyo. 2017. Using Centrality Measures to Predict Helpfulness-Based Reputation in Trust Networks. *ACM Trans. Internet Technol.* 17, 1 (2017), 8:1–8:20. https://doi.org/10.1145/2981545

[27] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology* 17, 3 (2017), 1–23. https://doi.org/10.1145/3003433

[28] Mark Newman. 2010. *Mathematics of Networks*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

[29] Keziban Orman, Vincent Labatut, and Hocine Cherifi. 2013. *An Empirical Study of the Relation between Community Structure and Transitivity*. 99–110. https://doi.org/10.1007/978-3-642-30287-9_11

[30] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

[31] Christopher Phethean, Thanassis Tiropanis, and Lisa Harris. 2015. Assessing the Value of Social Media for Organisations: The Case for Charitable Use. In *Proceedings of the ACM Web Science Conference (WebSci '15)*. Article 32, 32:1–32:9 pages. https://doi.org/10.1145/2786451.2786457

[32] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. *Physical Review E* 76, 3 (2007), 036106. https://doi.org/10.1103/PhysRevE.76.036106

[33] Ashwin Rajadesingan and Huan Liu. 2014. Identifying Users with Opposing Opinions in Twitter Debates. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Vol. 8393. Springer International Publishing, 153–160. https://doi.org/10.1007/978-3-319-05579-4_19

[34] Dmitry Saprykin, Galina Kurcheeva, and Maxim Bakaev. 2016. Impact of Social Media Promotion in the Information Age. In *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia (EGOSE '16)*. 229–236. https://doi.org/10.1145/3014087.3014109

[35] Cass R Sunstein. 2009. *Republic. Com 2. 0*.

[36] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. 2016. VODUM: a Topic Model Unifying Viewpoint, Topic and Opinion Discovery (regular paper). In *European Conference on Information Retrieval (ECIR), Padua, Italy, 20/03/2016-23/03/2016 (LNCS)*, Vol. 9626. Springer, 533–545.

[37] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. 2017. Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks. 87–96. https://doi.org/10.1145/3132847.3132897

[38] Felix Ming Fai Wong, Chee-Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*.

[39] Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We Make Choices We Think Are Going to Save Us: Debate and Stance Identification for Online Breast Cancer CAM Discussions. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, 1073–1081. https://doi.org/10.1145/3041021.3055134