Unsupervised Fine-grained Hate Speech Target Community Detection and Characterisation on Social Media

Anaïs Ollagnier, Elena Cabrio, Serena Villata

Unsupervised Fine-grained Hate Speech Target Community Detection and Characterisation on Social Media

Anaïs Ollagnier[1*], Elena Cabrio[1] and Serena Villata[1]

[1*]Université Côte d'Azur, Inria, CNRS, I3S, Polytech'Nice-Sophia, 930 route des Colles, Sophia Antipolis, 06903, France.

*Corresponding author(s). E-mail(s): ollagnier@i3s.unice.fr;
Contributing authors: elena.cabrio@inria.fr;
serena.villata@inria.fr;

Abstract

Recent studies have highlighted the importance to reach a fine-grained online hate speech characterisation to better understand how hate is conveyed, especially on social media. A key element in this scenario is the identification and characterisation of the hate speech target community, e.g., national, ethnic, religious minorities. In this paper, we propose a full pipeline relying on unsupervised methods to distinguish specific hate speech manifestations, i.e., targeted (group of) victim(s) and the protected characteristics (target-types) discriminated. Our contribution is threefold: (1) we leverage multiple data views to contrast different abusive behaviours; (2) we explore the use of clustering techniques to perform fine-grained hate speech target community detection, and (3) we address an in-depth content analysis of the generated hate speech target communities. Relying on multiple data views derived from multilingual pre-trained language models (i.e., multilingual BERT and multilingual Universal Sentence Encoder) and the Multi-view Spectral Clustering (MvSC) algorithm, the 69 experiments performed on the Multilingual Hate Speech dataset (MLMA) of tweets show that most of the configurations of the proposed pipeline significantly outperforms state-of-the-art clustering algorithms on French and English. Our experiments confirm the ability of the proposed approach to capture complex hate speech phenomena (i.e., intersections between victim-groups, target-types or both).

Keywords: Fine-grained hate speech target detection, Community detection, Multi-view clustering, Sentence embedding, Social media

1 Introduction

By empowering the freedom of expression and individual voices, social media platforms such as Twitter and Facebook, and community forums have witnessed an exponential growth, becoming an integral part of our daily lives. Whilst these online spaces have facilitated the communication and exchange of ideas and points of view, they have also opened the door to the proliferation of content that can be degrading, abusive, or otherwise harmful to people. An important and elusive form of such language is hateful speech, i.e., content that mocks or discriminates against a person or group based on specific characteristics such as colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [1, 2]. Over the last decade, the preoccupation for the use of electronic means of communication as a tool to convey hate, racist and xenophobic contents tremendously increased [3], and a large number of computational methods involving Natural Language Processing (NLP) and Machine Learning (ML) have been proposed for automated online hate speech detection, e.g., [4–6]. Most of the prior works have mainly considered the task as a binary classification problem seeking to distinguish hate and non-hate speech. However, recent works have highlighted the importance to consider the different hate speech facets (e.g., nature of the target, hate directness, hostility type) in order to better understand how hate is conveyed online [7–9]. On this purpose, a high number of resources and benchmark corpora aiming at evaluating the ability of automated methods to capture fine-grained hate speech phenomena (i.e., multi-level annotation accounting for different hate speech facets) were developed. The following examples extracted from the multilingual hate speech dataset (MLMA) [10] illustrate a multi-level annotation scheme. Here, we only detail the (group of) victim(s) targeted and the protected characteristics discriminated (e.g., references to racial or sexist stereotypes).

1. '*@user @user go back shithole country grab pussy*' – targets: immigrants, protected characteristic: ethnic origin
2. '*@user @user another disgusting lying feminazi.*' – targets: women, protected characteristic: gender
3. '*@url steve guy fucking retard.*' – targets: special needs, protected characteristic: disability

Besides the efforts made to develop methods enabling to distinguish fine-grained hate speech phenomena from each other, often only a specific facet is dealt with [11]. However, hate speech is a complex and multi-faceted notion relying on interconnected phenomena at stake. Despite existing works addressing the task from an intersectional perspective [12], the issue of understanding the multiple complex facets of hate speech phenomena is still an open challenge.

In addition, an important challenge when dealing with social media is to provide robust solutions enabling to cope with the fast and often unpredictable evolution of social media data. This evolution involves both the language level (i.e., neologisms and slang words) or content (i.e., expression of opinions in reaction to societal issues). However, most of existing approaches rely on supervised algorithms which suffer from well-known limitations including the lack of annotated data and the need to regularly update them in order to account for such continuous evolution of this kind of content. Such findings raise questioning about the ability of a such strategy to efficiently address alone the task of hate speech detection.

To undertake these issues, the **research objectives** of this paper are the following: (1) to encode harmful content to unveil key features aiming at contrasting different kinds of abusive behaviours; (2) to address the task of fine-grained hate speech target community detection adopting an unsupervised technique; and (3) to assess the ability of the proposed pipeline to establish partitioning reflecting complex hate speech phenomena.

To achieve these goals, we investigate the use of multiple data views (i.e., data structures) holding different properties to obtain meaningful hate speech representations. To do that, we exploit two well-known multilingual pre-trained language models: *mBERT* (multilingual BERT) [13] and *mUSE* (multilingual Universal Sentence Encoder) [14], from which syntactic, semantic and relationship information is derived. Then, the task of hate speech detection is transposed into a clustering problem allowing to benefit from unsupervised approaches, which do not need annotated corpora to be trained. Leveraging the last advances in clustering, we investigate the use of Multi-view Clustering (MvC), especially the Multi-view Spectral Clustering (MvSC) algorithm, in order to exploit complementary and consensual information across the multiple data views of a different nature (i.e., feature and graph spaces). To investigate the ability of the proposed pipeline to address challenges in the field (i.e., detecting fine-grained hate speech phenomena), we conduct experiments on a curated version of the French and the English of the MLMA dataset which only includes hostile tweets. We assessed the performances focusing on the target and group labels (hereinafter referred respectively to as target-type / victim-groups). Partitioning is evaluated regarding whether aggregated hate content correspond to existing hate speech target communities in the MLMA dataset, i.e., clusters containing content with the same target-type / victim-group labels. Conducted experiments show that the simultaneous clustering of multiple data views improves the clustering performance when compared to state-of-the-art clustering methods (k-means, k-medoids and spectral clustering) based on a single feature set on both languages. Furthermore, we also provide a study of the most frequent n-grams extracted from the generated clusters. The goal is to observe the ability of the proposed pipeline to generate semantic spaces reflecting hate speech manifestations used to offend a specific victim-group given a protected characteristic. In other words, we seek to identify whether the generated clusters may unveil an underlying structure of the

data similar to those provided by the MLMA multi-aspect hate speech analysis or - on the contrary - it is enabled to uncover different properties. From this study, we analyse the properties resulting from the automatic identification of fine-grained hate speech target communities with the purpose of providing key informational insights aiming at improving the design of machine learning tools dedicated to capture online hateful content.

In summary, the **contributions** of our paper are summarised as follows:

- Leveraging syntactic/semantic and relationship information to enrich hate content representations;
- Exploring the use of clustering techniques as a mean to address the task of fine-grained hate speech detection;
- Assessing the ability of the proposed pipeline to address the task through an in-depth content analysis of the generated hate speech target communities.

The paper is organised as follows: Section 2 discusses the related literature on hate speech detection, on community detection in social media and on the multi-view data clustering. Section 3 describes the multilingual hate speech dataset used in this study. Section 4 presents the proposed clustering method and discusses the methodological choices. The experimental procedure is described in Section 5. Section 6 details the experimental setting and report on the obtained results. Finally, Section 7 presents the study about the characterisation of the different types of hate speech used to target communities. Conclusions end the paper, drawing directions for future work.

NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

2 Related Work

The following sections provide a panorama of existing works aiming at automatically identifying abusive behaviours. In particular, we focus on a review of datasets and approaches developed to address this task.

2.1 Hate Speech Datasets.

From 2016 onward, a high number of resources and benchmark corpora for many different languages were developed. As hate speech is a complex and multi-faceted notion, the scientific community tackles this issue by developing various semantic frameworks aiming at identifying different topical focus such as specific targets (groups targeted), nuances of hate speech (abusive, toxic, dangerous, offensive or aggressive language) or rhetoric devices (slurs, obscenity, offences or sarcasm). Several surveys describe the current state of the field providing a structured overview of existing datasets [7, 11, 15]. Most of available datasets come from Twitter and rely on a binary scheme: two mutually-exclusive values to mark the presence or absence of hate speech such as those introduced in [16], [17] and, [18]. Some datasets relies on multi-level

annotation, with finer-grained schemes accounting for different phenomena. Recently, [19] adopt a three-layer binary annotation for hate speech, aggressiveness and nature of the target (individual or group), while [10] provides a fine-grained annotation of tweets about both victim-groups and target-types, hate directness (whether the text is direct or indirect), hostility type and annotator's sentiment. With the purpose to facilitate access to information, the platform *hatespeechdata.com*[1] catalogues datasets annotated for hate speech, online abuse, and offensive language aiming at training natural language processing systems.

Several corpora have been developed with the purpose of organising open shared tasks at NLP-related conferences including TRAC[2] (Workshop on Trolling, Aggression and Cyberbullying at LREC 2018 & 2020), EVALITA 2020 Tasks[3] (automatic misogyny identification & hate speech detection) and SemEval 2019 Tasks 5 (HatEval [19], multilingual detection of hate speech against immigrants and women in Twitter) & 6 (OffensEval [20], identifying and categorising offensive language in social media) at NAACL HLT 2019, as well as Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC 2019 & 2020 [4]), among others.

2.2 Hate Speech Detection Methods.

Most of the prior works have mainly considered the task of hate speech detection as a supervised document classification problem. Following the taxonomy introduced in [2], existing methods can be divided into two main categories: classical ML methods and deep learning methods.
Classical ML methods require as input feature vectors derived from text representation techniques. Text representation consists of converting textual content into machine-readable format using a collection of meaningful features. While this task has been widely investigated by the NLP community, it is still an ongoing challenge as it involves addressing complex semantic phenomena such as figurative language and idiosyncratic style. As a part of hate speech detection models, this task is mainly addressed using two main text mining encoding techniques: surface features and linguistic-based features. Word and character n-grams are currently the prominent shallow lexical features, and also the most successful ones [21, 22]. In [23], character 4-grams features outperforms other surface feature representations in distinguish between profanity and hate comments. Concerning linguistic-based features, Part of Speech (PoS) tagging and dependency parsing obtain the best performances by providing a deeper understanding of hateful content [21, 22, 24]. Features derived from sentiment analysis, usually used as auxiliary features, are also considered as powerful linguistic cues. The use of sentiment polarity or

[1] https://hatespeechdata.com/ Date of access: 2nd November 2021.
[2] https://sites.google.com/view/trac2/home Date of access: 2nd November 2021.
[3] http://www.evalita.it/ Date of access: 2nd November 2021.
[4] https://hasocfire.github.io/hasoc/2020/index.html Date of access: 24th February 2023.

emotion tone prove effective in tasks on aggression identification and threat detection [25]. Recently, the use of clustering techniques, especially Brown clustering, show promising results in representing positive and negative sentiment data to enrich offensive comment representations [26]. Once feature vectors are generated given a text representation strategy, they are consumed by supervised algorithms. Several surveys report the use of different algorithms including Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Random Forest [2, 21, 24]. Among these, SVM is still one of the most popular algorithms to address shared tasks on hate speech detection [27–29].

Deep learning based methods consists of applying neuronal networks to automatically learn multi-layers of abstract features from raw data. Here, inputs can be simply the raw text data, or take various forms of feature encoding, including any of those used in the classic methods. Prior works based on neuronal networks exploit the network structure either to design classification models or to build language models. In the context of hate speech classification, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and, Long Short-Term Memory network (LSTM) have shown to perform better than classic ML methods [2]. Hybrid methods combining neuronal networks have also been proposed and shown promising performances outperforming state-of-the-art methods on a large collection of public hate speech datasets [4]. In parallel, pre-trained word representations significantly advanced the state of the art in various NLP tasks including shared tasks on hate speech detection [30]. Besides traditional neuronal network models, transformer-based language models like Bidirectional Encoder Representations from Transformers (BERT) [13] achieve also state-of-the-art performance. In the SemEval 2019 shared task: Offens-Eval [31], most of the top-ranked models relies on BERT models. Recent works address the topic bias issue and introduced fine-tuned *Transformer* neuronal network architectures achieving state-of-the-art performance in the task of abusive language detection in English [32, 33].

From the literature review, we highlight that the task of hate speech detection is mainly considered as a supervised classification problem. However, dealing with social media content means to cope with dynamic data. Moreover, recent studies report the proliferation of code words for communities aiming at countering moderation tools [34], while [35] highlight the difficulties in tracking all racial and minority insults due to the constant evolution of social phenomena and language. Applying supervised methods raise concerns also about the long-term robustness of such systems dealing with evolving social media data. Unsupervised approaches are mainly used in the literature to obtain richer text representations including deep learning to generate embeddings [36] or LDA and Brown Clustering to provide auxiliary features [26, 37]. Concerning this point, clustering techniques, especially community detection algorithms, constitutes a key tool for the analysis of complex networks by enabling the study of mesoscopic structures that are often associated with organisational and functional characteristics of the underlying networks. In the context of

social media, the analysis of such networks presents valuable resources from which it is possible to gain insights into the social phenomena and processes [38]. For instance, outcomes from analysing the community structure of networks have led to a wide range of intelligent services and applications in the fields of opinion mining [39], marketing activities [40] and data-driven decision making [41], among others. Often addressed to reveal communities from user-to-user interaction (e.g., mentions, follows), it has also been performed considering only the textual content allowing to uncover cohesive groups or clusters based on semantic knowledge [42]. Using clustering techniques allow to overcome aforementioned limitations by providing a scalable, adaptive and robust solution to deal with social data. As hate speech is a complex and multi-faceted notion, we investigate the use of multi-view clustering to handle data views holding different properties to leverage richer data information. Introduced in [43], the core idea behind *Multi-view Clustering* is to leverage effectively the diversity and the complementary of multi-view data to improve the clustering performance. Typically, each of these *data views* provides a different perspective of a given set of entities. The term "multi-view clustering" refers to algorithms that can utilise multiple feature spaces to describe distinct points of view of a phenomenon [43, 44]. Recently, several important related surveys about MvC have been published to summarise the theories, methodologies, taxonomies and applications of the existing MvC approaches [45–47]. As a part of this work, we rely on the MVSC-CEV (multi-view spectral clustering by common eigenvectors) algorithm introduced in [48] which allows the use of an arbitrary number of input views, possibly of a different nature (feature or graph space) and with different dimensions. To the best of our knowledge, this is the first approach based on multi-view clustering applied to the problem of hate speech detection.

3 The MLMA dataset and its extension

The multilingual hate speech dataset (MLMA) [10] provides a fine-grained annotation of 5.647 English tweets, 4.014 French tweets, and 3.353 Arabic tweets. Five facets characterising hate speech are labelled, including the directness of the speech (*directness*), the hostility type (*sentiment*), the protected characteristic discriminated (*target-type*), the group of victims (*victim-group*), and the annotator's sentiment (*annotator_sentiment*).

In this paper, we focus on the French and the English portions of the dataset and on the facets related to the *target-type* and *victim-group* labels. More precisely, the *target-type* denotes whether the tweet insults or discriminates against people based on their *origin, religious affiliation, gender, sexual orientation, special needs* or *other*. In total, 16 common *victim-group* have been identified denoting whether the tweet is aimed at *women, people of African descent, Hispanics, gay people, Asians, Arabs, immigrants in general, refugees*; people of different religious affiliations such as *Hindu*, *Christian*, *Jewish people*, and *Muslims*; or different political ideologies as *socialists*, and others. The

individual covers hate directed towards one individual, which cannot be generalised. Both target-type and victim-group referring to *other* correspond to utterances which do not fit with MLMA facets' annotation guidelines. In the dataset, 49 hate speech target communities are identified in French and 70 in English (i.e., combining *target-type_victim-group* labels).

Table 1: Distribution of hate speech target communities in the French and English corpora.

	FR		**EN**	
Target	Group	no. of Tweets	Group	no. of Tweets
origin	other	619	other	814
	indian/hindu	324	specNeeds	378
	africanDesc	298	individual	301
	arabs	289	women	162
	individual	276	refugees	128
	leftWing	103	immigrants	123
	asians	87	leftWing	101
	immigrants	84	hispanics	96
	specNeeds	61	africanDesc	81
	-	-	muslims	75
	-	-	indian/hindu	67
	-	-	asians	61
other	individual	556	other	507
	other	442	women	145
	leftWing	280	specNeeds	120
	immigrants	20	individual	55
religion	muslims	71	-	-
	jews	56	-	-
sexOrient	-	-	other	133
	-	-	gay	111
	-	-	individual	98
	-	-	specNeeds	67
disability	specNeeds	83	specNeeds	944
	individual	80	other	58
	-	-	women	55
gender	women	22	women	474
	-	-	other	55
	-	-	specNeeds	50
Total	-	3.701	-	5.209

A serious skewed distribution of the hate speech target communities is observed in each corpus, with 69.3% of hate speech target communities in French below the average imbalance ratio (i.e., the average proportion of the number of instances in the majority community to the number of instances in the minority ones) and 72.8% in English. Most of the studies on the behaviour of machine learning applications have shown a significant loss of performance facing imbalanced datasets [49]. As one of the purposes of this study is to

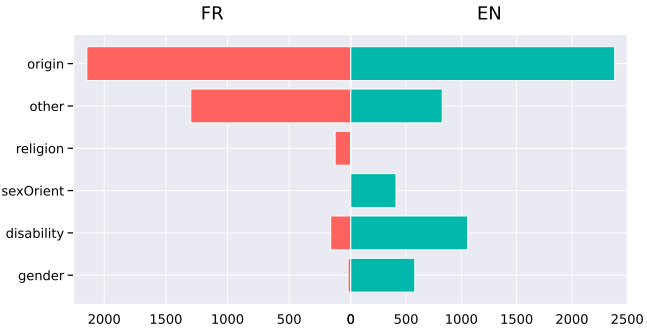


Fig. 1: Distribution of the *target-type* facet in both French and English corpora.

evaluate the ability of the proposed pipeline to identify automatically fine-grained hate speech phenomena relying on the MLMA labels, the corpora are filtered to alleviate the bias towards the majority communities. Only communities getting an imbalance ratio above 20 for the French corpus and 15 for the English one are considered in this study. In addition, as we focus on hostile tweets conveying abusive or threatening speech only tweets labelled as *abusive, hateful, offensive, disrespectful or fearful* are considered. The resulting datasets (hereinafter referred respectively to as the *FR hate speech target community* corpus and the *EN hate speech target community* corpus) comprise 3.701 and 5.209 tweets divided respectively into 16 and 26 hate speech target communities, as reported in Table 1.

From this Table, we can observe that English tweets tend to target people with *special needs* and *women* over the different target-type facets, while French tweets are more offensive towards the victim-group *individual*. Figure 1 illustrates the global distribution of target-types on both languages. From this figure, we can observe a flagrant difference between the protected characteristics discriminated according to each language. In detail, tweets discriminating or insulting against people based on their origin is the highest accounted target-type. Concerning the other target-type facets, disparities are more important according to the language. Whilst the target-type *sexual orientation* is among one of the main protected characteristics discriminated in English (7.85% in total considering the whole English tweets), in French it is pruned due to a low number of samples. The same phenomenon is observed on the target-type *religion* which is pruned in English while it represents 3.4% of the hateful content in the French corpus. In English, *disability* is the second most frequent target-type followed by *other* and *gender*. In French, the second most populate target-type facet is *other* followed by *disability* and *religion*. Table 2 presents a sample of hateful tweets with the corresponding annotations extracted from the MLMA dataset.

Table 2: Annotation examples in the MLMA dataset.

smh women r retarder @url	quel mongol ! jsuis sur c un rebeu
Target type: disability **Victim group**: women	**Target type**: origin **Victim group**: specNeeds **Translation**: what a retard ! I'm sure it's a raghead
(a) English.	(b) French.

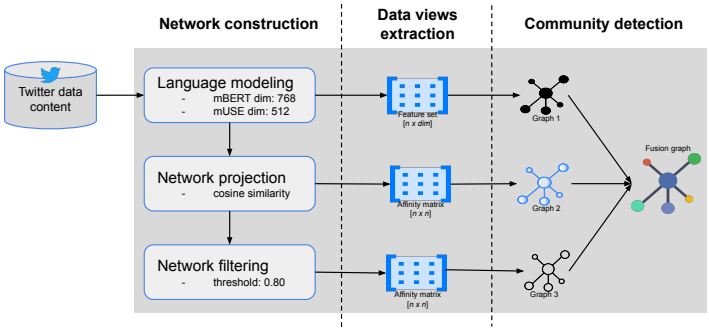


Fig. 2: Multi-view clustering workflow from Twitter data.

4 A Framework for Fine-grained Hate Speech Target Community Detection

This section describes the proposed pipeline for the task of fine-grained hate speech target community detection. As visualised in Figure 2, it consists of three main steps including the network construction, the data views extraction and the target community detection.

4.1 Network Construction

Twitter data are noisy and unstructured, and include linguistic errors and idiosyncratic style. Handling such content implies using adapted cleaning processes to fully leverage data content and extract relevant information. Several works have proposed specific preprocessing pipelines showing significant improvements in model performances [50, 51]. As a part of the proposed work, we have implemented the benchmark preprocessing framework presented in [6] aiming at dealing with heterogeneous hate speech datasets extracted from various social media. In short, the applied cleaning process consists of performing data normalisation including hashtags (i.e., split into single words), emojis (i.e., replaced by their textual description), user mentions and URLs (i.e., replaced by canonical forms: username and url), and next, tokenises and lemmatises the textual content.

Table 3: Description of the Pre-trained language models.

Name	Architecture	Training data
mBERT [13]	12-layer, 768-hidden, 12-heads	Trained on uncased texts in the top 102 languages with the largest Wikipedias.
mUSE [14]	6-layer, 512-hidden, 8-heads	Originally trained on mined question-answer pairs, SNLI data, translated SNLI data and parallel corpora over 16 languages. Here, we use the V2 extended to 50+ languages.

After data cleaning and normalisation, the step consists of encoding the *hate speech target community* corpora to obtain node-like structures. As pre-trained language models have become a popular method achieving state-of-the-art results in a wide range of NLP tasks, especially as a part of feature construction methods [52, 53], we decide to exploit two well-known multilingual pre-trained language models to generate sentence vector-based features. From empirical studies we conducted on the *hate speech target community* corpora, *mBERT* (multilingual BERT) and *mUSE* (multilingual Universal Sentence Encoder) achieved better performances. These two pre-trained language models - evaluated against XLM-RoBERTa and Quora distilbert multilingual on the community detection task using k-means and spectral clustering with a single feature set - are used to generate sentence embeddings for each tweet. Table 3 provides a brief overview of both pre-trained models describing the encoder architecture and the training dataset. In detail, *mBERT* relies on the transformer model *BERT* [13] (Bidirectional Encoder Representations from Transformers) and is pre-trained on a large corpus of multilingual data. More precisely, it is pre-trained with two objectives: Masked Language Modelling (MLM, 15% of tokens are masked and *BERT* is trained to predict them from context) and Next Sentence Prediction (NSP, it is trained to predict if a chosen next sentence was probable or not given the first sentence). Concerning *mUSE*, it relies on the *USE* [54] (Universal Sentence Encoder) architecture and is originally pre-trained on Wikipedia, web news, web question-answer pages and discussion forums and augmented with the Stanford Natural Language Inference (SNLI) corpus translated to 15 languages. From a multi-task dual-encoder model, this transformer constructs sentence embeddings based on cross-lingual representation learning that combines methods for multi-task learning of monolingual sentence representations. Here, the pre-trained *mBERT* model used is publicly available on HuggingFace[5]. In order to generate sentence embeddings, we perform a mean pooling of the model outputs. Conversely, the *mUSE* model used in this study, also released by HuggingFace[6], allows to generate directly sentence embeddings.

[5] https://huggingface.co/mBERT-base-multilingual-uncased Date of access: 5th October 2021.
[6] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2 Date of access: 5th October 2021.

Once the feature sets are built, the next step is the *network construction*, which consists of reshaping the data into node-like structures. To identify tweet pairs sharing key textual features, we use the cosine similarity measure. Widely used to determine similarities between vectors [55], this similarity measure has the advantage to allow comparisons between inputs with different lengths.

Figure 3 shows samples of sentence similarity scores obtained from the corresponding affinity matrix of each pre-trained language model. As illustrated, the inner representation of the languages learnt by each pre-trained model results in affinity matrices conveying different sources of rich semantic information. From the affinity matrices, a network (graph) is generated in which individual tweets are the nodes, and similar tweets are grouped together. As a result we obtain a weighted and undirected graph G. Edges have positive weights, and the graph topology is described by the affinity matrix W, where the generic element $W_{uv} = W_{vu} > 0$ if there is a weighted edge between nodes u and v, while $W_{uv} = W_{vu} = 0$ otherwise. However, keeping the full information about the network may be problematic. A network with a high edge density may be intractable by traditional tools of network analysis. It may especially pose a serious obstacle for graph clustering techniques [56]. To overcome this issue and considering the weighted character of the edges in the proposed approach, we perform information reduction using a simple weight thresholding method. Weight thresholding removes all edges with weight lower than a threshold value. This means that the resulting graph $\tilde{G}$ has a thresholded weight matrix $\tilde{W}$, whose generic element $\tilde{W}_{uv} = \tilde{W}_{vu} = W_{uv}$ if $W_{uv} \geq \theta$ and $\tilde{W}_{uv} = \tilde{W}_{vu} = 0$ otherwise. The thresholded graph $\tilde{G}$ is therefore a subgraph of G with the same number of nodes. Concerning the connectivity of this subgraph, as there are so many ways to express hatred some users' utterances obtain similarity scores below the defined threshold. However, these tweets constitute relevant informational segments which can allow to better capture how hate is spread online. In addition, the lack of connectivity in a graph can impact its analysis leading to decrease the quality of the partitioning. To bridge isolated nodes towards the main connected components, we connect them to their closest neighbours (i.e., nodes getting the highest similarity scores). From empirical studies we conducted on the *hate speech target community* corpora, the weight threshold used throughout the paper is set to 0.8 when referring to the thresholded graphs.

4.2 Data views Construction

The MvSC algorithm maximises clustering quality considering the diversity and complementary of different data views. In this regard, one of the main contributions of this work is to derive different views holding specific properties to unveil a more robust network partitioning. Here, data views of different nature are considered including feature sets (syntactic/semantic information) and affinity matrices (relationship information) resulting from the *network construction* step. Data views used in this work are detailed below:

- Feature sets (syntactic/semantic information)

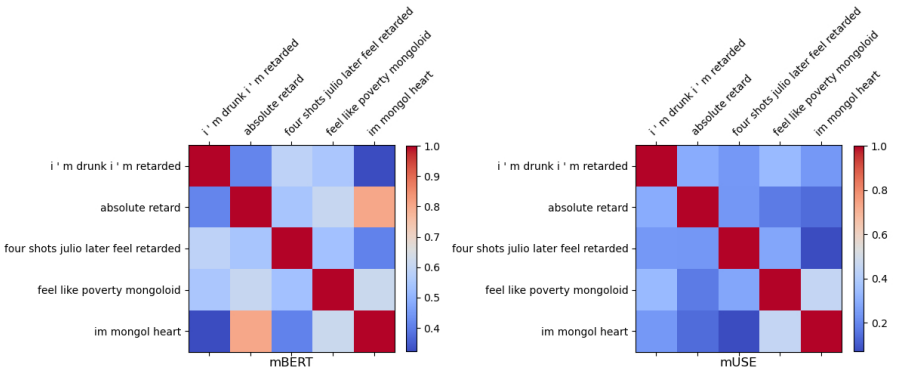


Fig. 3: Sentence similarity scores using embeddings from *mBERT* (left) and *mUSE* (right).

– *mBERT view (`mBERT`).* This view refers to the sentence embedding representations encoded using the pre-trained *mBERT* model. As a result, we get feature sets of 5.209 data points in English and 3.701 in French with an embedding size of 768.
– *mUSE view (`mUSE`).* This view consists of the sentence vector-based features generated from the pre-trained *mUSE* model. Resulting feature sets are composed of 5.209 data points in English and 3.701 in French with an embedding size of 512.

- Affinity matrices (relationship information)

– *Projection view (`mBERT-PRO` & `mUSE-PRO`).* These views consist of affinity matrices resulting from the cosine similarity computed respectively on the `mBERT` view for `mBERT-PRO` and on the `mUSE` view for `mUSE-PRO`.
– *Network view (`mBERT-NET` & `mUSE-NET`).* These views refer to affinity matrices relying on the thresholded subgraphs (Section 4.1). The `mBERT-NET` and `mUSE-NET` views are respectively built from the corresponding original view, namely the `mBERT` view and the `mUSE` view.

While the feature sets capture relations, similarities and semantic relationships between sentences, the graph spaces contain information about nodes (i.e., tweets) connectedness (i.e., whether pairs of nodes are adjacent or not in a graph and the nature of their connection). The deriving views obtained from different language models – relying on different encoding strategies and exhibiting different properties – is beneficial to accurately describe the textual data we analyse. In the following section, we describe how these views are consumed by the MvSC algorithm.

4.3 Community Detection

Relying on a MvSC algorithm, the final module of the proposed pipeline aims at identifying communities. In the proposed approach, target communities refer

to the different hate speech targets introduced in Section 3. In order to unveil mesoscopic structures among tweets related to the expected communities, this step relies on the MVSC-CEV [48]. The main structural difference between prevailing multi-view clustering methods and MVSC-CEV lies in the step where the information from the multiple views is collapsed into a single view to produce the final clustering assignment. Roughly speaking, multi-view clustering methods consider as input data views and obtain an affinity matrix for each view. Then, a projection of each affinity matrix is computed into a space suitable for clustering, to produce a consensus partition. As a part of the MVSC-CEV algorithm, it clusters all views separately. Then, it takes the obtained clustering assignments and loops back to the data projection step in order to improve the projections with the clustering information previously obtained.

Formally, the algorithm considers as input a set of D data views $\overline{V} = \{V_1, V_2, \ldots, V_D\}$ of the data with n samples each. For each data view $V_D \in \overline{V}$, a similarity matrix is computed using the Gaussian similarity function resulting in a set of similarity matrices $\overline{S} = \{S_1, S_2, \ldots, S_D\}$. In turn, for each $S_D \in \overline{S}$ its corresponding Laplacian matrix L_D is generated, where $L_D \in \overline{L}$ and $\overline{L}$ corresponds to the set of computed Laplacian matrices. The set of Laplacian matrices $\overline{L}$ is passed to the S-CPC algorithm [57] along with k desired number of clusters in order to compute their common eigenvectors. As a result, a matrix X is obtained, where the k largest eigenvectors of the Laplacian matrices $L_D \in \overline{L}$ are the first eigenvectors $x_1^{(i)}$ of each submatrix $L_D^{(i)}$. Finally, k-means is applied to the matrix Y producing the partitioning of the input data samples common to the D input views, where Y is the result of the normalisation of the matrix X. Therefore, it is a co-training approach, since it uses the results of one iteration to further improve the results of the final clustering. Figure 4 presents a visualisation of the best partitioning (cf. Section 6) resulting from the proposed pipeline applied to the *FR hate speech target community* corpus. In this example, the consensus partition is obtained computing eigenvectors common to the `mUSE` view and the following affinity matrices: `mBERT-PRO`, `mUSE-PRO` and `mUSE-NET`. As a result, we obtain a graph composed of 3.701 nodes and 16.945 edges. Each colour refers to one of the 16 ground-truth (labels already used in the MLMA dataset) fine-grained hate speech target communities of the French corpus.

5 Experimental setting

In this section, we firstly describe the reference methods evaluated in this study, and secondly, we introduce the evaluation metrics used to assess the quality of the clustering produced by each method. Finally, we detail the procedure of hypothesis testing used to perform significance tests.

Fig. 4: Hate speech target community network relying on a 2-views configuration (the `mBERT` view and the `mUSE` view). The network is obtained from the *FR hate speech target community* corpus and it is composed of 16 node communities. Each node colouring corresponds to an hate speech target community. Individual nodes represent the tweets and are linked to each other considering their semantic similarity.

5.1 Reference methods

To evaluate the advantages of mixing multiple data views, we compare the MvSC clustering algorithm against three well-known clustering techniques: k-means [58], k-medoids [59] and spectral clustering [60]. All these state-of-the art clustering techniques are based on the Scikit-learn implementation. As k-means only allows feature sets as input, here k-medoids is used to evaluate the performances of the proposed single-view configurations relying on similarity matrices. Conversely, the spectral clustering algorithm allows inputs of a different nature (feature or graph space). Both k-means and k-medoids use default parameters. Concerning spectral clustering, the affinity parameter for the feature spaces is set to 'nearest_neighbors' (construct the affinity matrix by computing a graph of nearest neighbours) and for the similarity matrices it is set to 'precomputed_nearest_neighbors' (interpret precomputed distances and construct a binary affinity matrix from the n_neighbors nearest neighbours of each instance). As all the clustering methods analysed and compared require to define a number of desired clusters, the number defined for the French and

the English datasets is respectively 16 and 26 clusters, numbers corresponding to their ground-truth hate speech target communities.

5.2 Performance assessment

Following the methodology described in [61], the quality of the clustering methods is evaluated using clustering *purity*, clustering *Adjusted Rand Index* (ARI) and the *Normalised Mutual Information* (NMI) metric. Briefly, purity is the ratio of the summation of how many maximum points of each algorithmic cluster $c \in k$ match with a considered gold set cluster $g \in t$ and the total number of points in data, such as:

$$purity(c_k, g_t) = \frac{\sum_{i=1}^{k} max_{j=1}^{t}(c_i \cap g_j)}{N} \tag{1}$$

the higher the purity the better the clustering outcome is. The maximum purity value is 1.0. The *Rand Index* calculates a similarity between two clusterings (i.e, sets of clusters), by looking at each peer of individuals and counting those that are or are not in the same cluster, depending on whether you are in actual or predicted clustering:

$$RI = (a + b)/({}_nC_2) \tag{2}$$

where a is the number of times a pair of elements belongs to the same cluster across two clustering methods, b is for those that belong to difference clusters across two clustering methods, and ${}_nC_2$ is the number of unordered pairs in a set of n elements. As such, the RI does not guarantee that random assignment will produce a value close to 0. This is why this raw index is 'adjusted to account for chance', which gives the ARI score:

$$ARI = \frac{RI - Expected\_RI}{max(RI) - Expected\_RI} \tag{3}$$

The ARI, which is symmetrical, measures the similarity and the consensus of two assignments, ignoring permutations and normalising against what would have happened by chance. The NMI, built on the Shannon entropy of information theory, tries to quantify the amount of shared information between two clusterings C and T. Formally:

$$MI(C, T) = \sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \log \left(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}} \right) \tag{4}$$

It measures the dependence between the observed joint probability p_{ij} of C and T, and the expected joint probability $p_{C_i} \cdot p_{T_j}$ under the independence assumption. When C and T are independent then $p_{ij} = p_{C_i} \cdot p_{T_j}$, and thus $MI(C, T) = 0$. In other words, $MI(C, T)$ can be thought of as the informational

'overlap' between C and T, or how much we learn about C from knowing T (and about T from knowing C). In order to normalise the MI value, the NMI of C and T is defined as follows:

$$NMI(C,T) = \sqrt{\frac{MI(C,T)}{H(C)} \cdot \frac{MI(C,T)}{H(T)}} = \frac{MI(C,T)}{\sqrt{H(C) \cdot H(T)}} \tag{5}$$

where $H(.)$ corresponds to the computation of the Shannon entropy. All these indices lie in the range $[0, 1]$, and values tending towards unity indicate a perfect correlation between the partitions. In our experiments we use the Scikit-learn implementation of the ARI and NMI while the *purity* is implemented following the instructions reported in [61].

5.3 Statistical tests

For all the analysed and compared stochastic clustering methods in this study, a total of 31 independent executions for each dataset were performed. For each clustering method and evaluation metric, we carried out statistical tests from all the executions to determine the significance of the reported performances. To define the appropriate procedure of hypothesis testing, we checked the normality assumption using the *Shapiro-Wilk test* [62] and the homogeneity of variance through the *Levene's test* [63]. As a result, and for both assumptions, we rejected the null hypotheses indicating that the residuals are not normally distributed and that their variability is statistically insignificant. Therefore, we use the *Wilcoxon signed ranks test* [64] to compare every pair of approaches on each dataset and across all datasets. Furthermore, we use the *Friedman test* [65] to compare multiple clustering methods on each dataset or across all datasets. In the case where the latter test reveals significant differences between the results, a *post hoc Nemenyi test* [66] is performed to establish a hierarchy between the approaches. In all tests, we assume the significance level $\alpha = 0.05$. From these tests, we expect to assess evidences concerning the plausibility of the following hypotheses:

1. Question: Considering each evaluation metric individually, are the performances observed on each model equal?
   - Null Hypothesis (H0) — Models achieve similar performances.
   - Alternative Hypothesis (HA) — Some models perform better considering the evaluation metrics.
2. Question: Considering all evaluation metrics, are there models they do globally perform better?
   - Null Hypothesis (H0) — Proposed models achieve similar performances considering all the evaluation metrics.
   - Alternative Hypothesis (HA) — Some models provide global better performances.

Both hypotheses are claimed on each dataset and also across all the datasets. Outcomes of these hypotheses allow to observe whether the combination of certain methods and view configurations perform better depending on the language and the tested evaluation metrics, or not.

6 Results

This section investigates the ability of the MVSC-CEV algorithm to generate high-quality clustering solutions on both the *FR hate speech target community* corpus and the *EN hate speech target community* corpus by varying the combination of the data views introduced in Section 4.2. Experiments were conducted on an Intel i7-4600U CPU @ 2.10 GHz with 32GB of RAM, using single-threaded processes. Table 4 summarises the results achieved by the two baseline clustering techniques and the top-3 models relying on MVSC-CEV. In total 69 experiments were conducted, all the results are reported in Appendix A Table A1.

English/French corpus performance – Concerning the English corpus, most of the configurations relying on the MVSC-CEV algorithm outperforms the baseline clustering techniques, except the {`mBERT/mUSE/mUSE-PRO`} configuration w.r.t. the ARI score. Concerning each evaluation metric, the *post hoc Nemenyi test*, allowing to evaluate the consistency of the ranking of each model throughout all the iterations, ranks the {`mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET`} and {`mBERT-PRO/mUSE-PRO/mBERT-NET`} configurations respectively at the first and the second position. Considering all the metrics, the *post hoc Nemenyi test* establishes that {`mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET`} outperforms other models followed by {`mBERT-PRO/mUSE-PRO/mBERT-NET`} and {`mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET`}. In French, the majority of the data view configurations combined to the MVSC-CEV algorithm achieves good performances as well as the baseline clustering techniques relying on the feature set and the projection affinity matrix derived from *mUSE*. The *post hoc Nemenyi test* allows to establish that {`mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET`} outperforms the other models w.r.t. the *purity* score, {`mBERT-PRO/mUSE-PRO`} regarding the ARI score and {`SC view mUSE`} w.r.t. the NMI score. Considering all the metrics, the *post hoc Nemenyi test* establishes that {`mUSE/mBERT-PRO/mUSE-PRO`} outperforms the other models followed by {`mBERT-PRO/mUSE-PRO`} and {`mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET`}.

Overall evaluation – Table 5 shows the top-5 models obtained from the *post hoc Nemenyi test* performed on each evaluation metric considering both corpora. The latter test allows to reject the null hypothesis formulated in Question 1 (cf. Section 5.3). Indeed, the resulting hierarchy confirms that some models achieve better performances regarding the tested evaluation metrics. In detail, {`mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET`} outperforms the other methods on both the *purity* score and the NMI score. This finding highlights

Table 4: Detailed results w.r.t. *purity*, ARI and NMI. The mean and standard deviation from 31 independent runs are reported for each language. According to the *post hoc Nemenyi test* the best values reported for each metric are in bold while the highlighted rows refer to the best models considering all the metrics for each language.

Single-view configs.	EN			FR		
	K-Means & Spectral clustering					
	Purity	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**
{KM view mBERT}	0.286±0.006	0.053±0.005	0.128±0.008	0.258±0.003	0.039±0.003	0.121±0.003
{KM view mBERT-PRO}	0.299±0.007	0.045±0.007	0.139±0.008	0.254±0.010	0.040±0.007	0.113±0.009
{KM view mBERT-NET}	0.191±0.004	0.002±0.004	0.045±0.011	0.177±0.005	0.000±0.000	0.026±0.007
{KM view mUSE}	0.372±0.007	0.096±0.006	0.259±0.004	0.348±0.004	0.093±0.007	0.229±0.003
{KM view mUSE-PRO}	0.336±0.012	0.088±0.010	0.218±0.011	0.324±0.014	0.091±0.011	0.197±0.013
{KM view mUSE-NET}	0.187±0.003	0.001±0.002	0.033±0.009	0.183±0.007	0.000±0.001	0.032±0.009
{SC view mBERT}	0.352±0.003	0.106±0.009	0.234±0.004	0.306±0.000	0.071±0.000	0.178±0.000
{SC view mBERT-PRO}	0.347±0.000	0.094±0.000	0.238±0.000	0.300±0.001	0.065±0.001	0.182±0.001
{SC view mBERT-NET}	0.251±0.000	0.013±0.000	0.151±0.000	0.237±0.000	0.011±0.000	0.118±0.000
{SC view mUSE}	0.344±0.000	0.090±0.001	0.242±0.000	0.339±0.000	0.101±0.000	**0.237±0.000**
{SC view mUSE-PRO}	0.345±0.000	0.101±0.000	0.234±0.000	0.342±0.002	0.103±0.001	0.231±0.003
{SC view mUSE-NET}	0.232±0.000	0.004±0.000	0.131±0.000	0.291±0.000	0.031±0.000	0.166±0.000
Multi- view configs.	MVSC					
	Purity	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**
{mUSE/mBERT-PRO}	0.382±0.003	0.098±0.003	0.263±0.003	0.360±0.002	0.107±0.004	0.227±0.002
{mBERT-PRO/mUSE-PRO}	0.385±0.002	0.101±0.003	0.267±0.002	0.362±0.002	**0.110±0.001**	0.227±0.002
{mBERT/mUSE/mUSE-PRO}	0.360±0.002	0.085±0.004	0.247±0.002	0.362±0.001	0.104±0.004	0.222±0.002
{mUSE/mBERT-PRO/mUSE-PRO}	0.384±0.003	0.101±0.004	0.268±0.003	0.364±0.003	0.109±0.003	0.227±0.002
{mBERT-PRO/mUSE-PRO/mBERT-NET}	0.391±0.004	**0.108±0.005**	0.273±0.003	0.343±0.005	0.088±0.001	0.216±0.003
{mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}	**0.390±0.002**	0.108±0.005	**0.276±0.001**	**0.367±0.001**	0.099±0.005	0.231±0.001
{mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.383±0.004	0.106±0.006	0.270±0.004	0.343±0.003	0.088±0.001	0.208±0.002
{mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.380±0.003	0.105±0.005	0.271±0.002	0.350±0.003	0.093±0.003	0.221±0.002

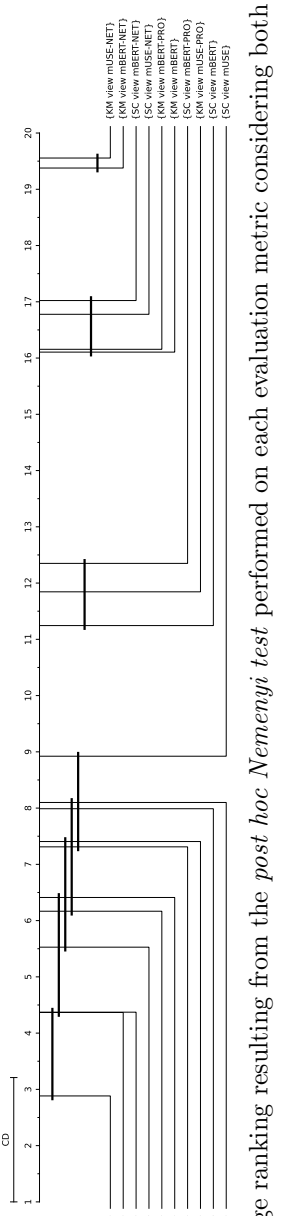


Fig. 5: Models' average ranking resulting from the *post hoc Nemenyi test* performed on each evaluation metric considering both corpora.

Table 5: Top-5 models w.r.t. the *post hoc Nemenyi test* performed on each evaluation metric considering both corpora. Figure A1 in Appendix A details the whole hierarchies obtained for each metric.

Rank	Evaluation Metrics		
	Purity	ARI	NMI
1	{mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}	{mBERT-PRO/mUSE-PRO}	{mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}
2	{mBERT-PRO/mUSE-PRO}	{mUSE/mBERT-PRO/mUSE-PRO}	{mUSE/mBERT-PRO/mUSE-PRO}
3	{mUSE/mBERT-PRO/mUSE-PRO}	{mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}	{SC view mUSE}
4	{mUSE/mBERT-PRO}	{SC view mUSE-PRO}	{mBERT-PRO/mUSE-PRO}
5	{mBERT-PRO/mUSE-PRO/mBERT-NET}	{mUSE/mBERT-PRO}	{KM view mUSE}

the ability of this configuration to generate homogeneous clusters (i.e. each cluster gathers tweets belonging for the most part to the same ground-truth community) and to maximise mutual information (i.e. uncertainties of dependencies between clusters and labels is decreased). Concerning the ARI score, {mBERT-PRO/mUSE-PRO} achieves the best performances providing the most similar partitioning in comparison to the ground-truth communities considering the frequency of occurrence of agreements over the total pairs. In short, most of the top-ranked models rely on the MVSC-CEV algorithm except the fourth-ranked model of the ARI score and both the third and the fifth-ranked models of the NMI. Figure 5 allows to visualise the hierarchy among the models according to *post hoc Nemenyi test* performed on each evaluation metric considering both corpora. From the latter test we can reject the null hypothesis formulated in Question 2 (cf. Section 5.3). Indeed, {mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET} outperforms all the other models reaching the best balance considering all the evaluation metrics on both languages. {mUSE/mBERT-PRO/mUSE-PRO} and {mBERT-PRO/mUSE-PRO} are respectively ranked at the second and the third position. From these findings, we can confirm that using data views of a different nature is beneficial in this context to improve the clustering performance considering all the tested aspects. Additionally, we can also observe that encapsulating the complementary information of different language representations unveil relevant data properties improving partitioning. Moreover, the consistency of the occurrence of affinity matrices in top-ranked models, especially the ones derived from the network projection, allows to establish that these affinity matrices exhibit relevant properties describing accurately data connections. Furthermore, the majority of the configurations relying on data views based on *mUSE* achieves better performances considering both the baseline clustering methods and the MVSC-CEV algorithm.

To sum up, the majority of the models relying on the MVSC-CEV algorithm achieves the best results on both languages and tested evaluation metrics. We can observe a consistency of the performances on each metric highlighting the portability of the proposed pipeline across other languages. In addition, the *purity* and the NMI score achieve the highest results witnessing the ability of this pipeline to also generate homogeneous clusters and to decrease the uncertainty among clusters. Considering results reported in Appendix A, we can notice that affinity matrices (relationship information) derived from the

network projection appear to be a rich source of information as they are a part of most of the top-ranked models. In addition, despite the fact that {mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET} provides the best balance considering both languages and metrics, performances vary throughout the configurations. For instance, to get the best ARI {mBERT-PRO/mUSE-PRO} has to be preferred in French and {mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET} in English.

7 Hate speech target community Characterisation

In this section, we explore the partitioning resulting from the proposed pipeline. Our goal here is not to perform an error analysis as in standard supervised approaches, but more to identify whether the properties emerging from the generated clusters correspond to the facets used in the MLMA multi-aspect hate speech analysis, i.e., our gold-standard corpus.

For each language, we explore the partitioning resulting from the best balance model identified in Section 6, namely {mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}. The methodology consists of observing the ground-truth communities' behaviours within the generated clusters in terms of scatteredness and equivocality. To disambiguate, in this study scatteredness refers to the level of dispersion of the communities and equivocality attempts to qualify the ability of each community to federate their own cluster(s). From these two aspects, we investigate the ability of victim-group and target-type to provide salient properties allowing to generate fully-fledged communities corresponding to the MLMA labels. To establish these two aspects, we consider the most frequent ground-truth hate speech target community label within each cluster as the cluster-head. Next, the content of each cluster is analysed using n-gram co-occurrence statistics.

Figure 6 describes the behaviours of the communities observed in the partitioning obtained from the *EN hate speech target community* corpus. Considering the equivocality, 7 ground-truth communities federate their own cluster(s). Among them 4 are the heads of one cluster including *other_other*, *sexual_orientation_other*, *origin_refugees* and *sexual_orientation_individual* while the others are the heads of multiple clusters. *origin_other* reaches the highest number of clusters by being the head of 10 different clusters representing 38.4% of the total number of clusters. The second and the third communities obtaining the highest degree of equivocality are *disability_special_needs* and *gender_women* representing respectively 30.7% and 15.3% of clusters' heads. Considering the scatteredness, we can observe that some communities being cluster-heads are also assimilated to other clusters. For instance, *other_other* and *origin_refugees* are above 70% of scatteredness. Entries composing both of these two clusters occur frequently in clusters having *origin_other* and *disability_special_needs* as head. From the mixing of these ground-truth communities, new communities emerge characterising how the hate is expressed towards a victim-group or the different ways of linguistically discriminate a target-type or

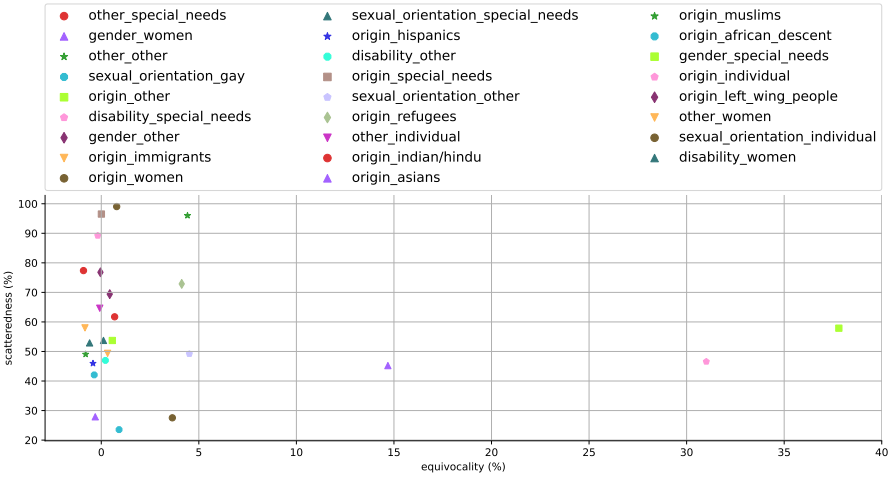


Fig. 6: Scatteredness and equivocality of the ground-truth communities in the *EN hate speech target community* corpus.

both. For instance, clusters mixing entries from the ground-truth communities *origin_refugees* and *origin_other* gather tweets using terms such as 'go back country' or 'shithole countries'. These segments are frequently associated with the victim-group *refugees* but their usage is also extended to discriminate and insult more globally people based on their origin. From the generated communities mixing *other_other* and *origin_other* salient patterns emerge combining vocabularies used to offend people based on their origin or political views such as 'radical leftist', 'leftist terrorist' or 'conspiracy terrorist'. In parallel, 19 ground-truth communities are fully assimilated into the generated clusters regarding various degrees of scatteredness. Among them, *origin_women* reaches 100% of scatteredness showing that tweets composing this ground-truth community are also scattered throughout all the clusters. Entries from this ground-truth community occur frequently within clusters being federated by the community *gender_women*. Studying these combinations highlighted salient properties (identified as a part of two different clusters) characterising hate speech expresses towards women such as 'fucking cunt' and 'little cunt' in the first cluster and 'absolute twat' and 'fucking twat' in the second one. *other_special_needs*, *origin_special_needs* and *origin_individual* are highly dispersed with a degree of scatteredness above 70%. Considering both of these aspects, we can observe that the cluster-head *sexual_orientation_individual* is the less equivocal and scattered ground-truth community. Indeed, this community assimilates only entries belonging to its community or from *gender_women* and *disability_special_needs*. Combined to *gender_women* the cluster having *sexual_orientation_individual* as head highlights properties related to the *women* group and hate speeches based on the vocabulary used to discriminate or insult people based on their sexual orientation including 'faggot bitch' and 'suck dick'. *origin_african_descent* and

origin_asians are the ground-truth communities obtaining the lowest degrees of scatteredness with respectively 23.0% and 26.9%. Both of them occur only in the clusters having *origin_other* and *disability_special_needs* as head. From the generated clusters assimilating the ground-truth community *origin_asians*, different offensive speeches emerge based on slurs such as 'screeches high' and 'ching chang' when combined to *origin_other* and 'proceeds squint' and 'ching chong' when mixed to *disability_special_needs*' entries. Figure 7 allows to better visualise the mixing of communities' semantic spaces within the generated clusters. As previously reported, we can observe that both vocabularies related to the target *disability* and the group *specNeeds* are used in offensive comments targeting women. Other prominent mixing can be noticed including the community *disability_special_needs* whose the vocabulary is based on common slurs and demeaning expressions widely used to discriminate and insult the other communities.

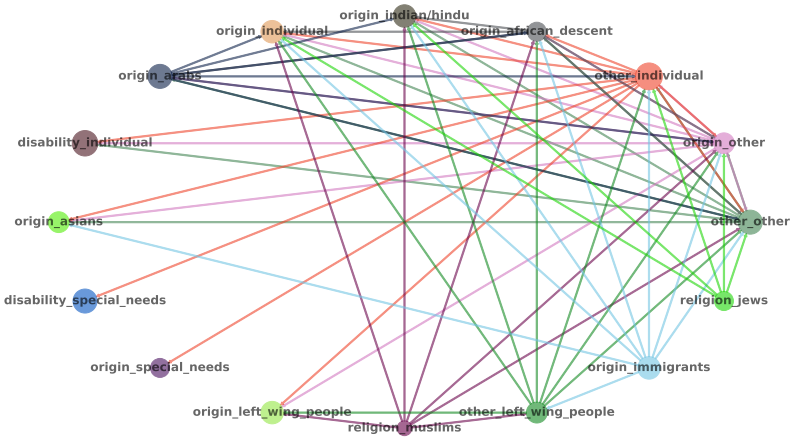


Fig. 7: Visualisation of the mixing of the communities within the generated clusters for the *EN hate speech target community* corpus. The bigger the nodes, the bigger the communities occur as a cluster-head in the generated partition. Nodes refer to ground-truth communities and edges link towards communities occurring in the given cluster-head.

Figure 8 describes the behaviours of the communities observed in the partitioning obtained from the *FR hate speech target community* corpus. Considering the equivocality, 9 ground-truth communities federate their own cluster(s).

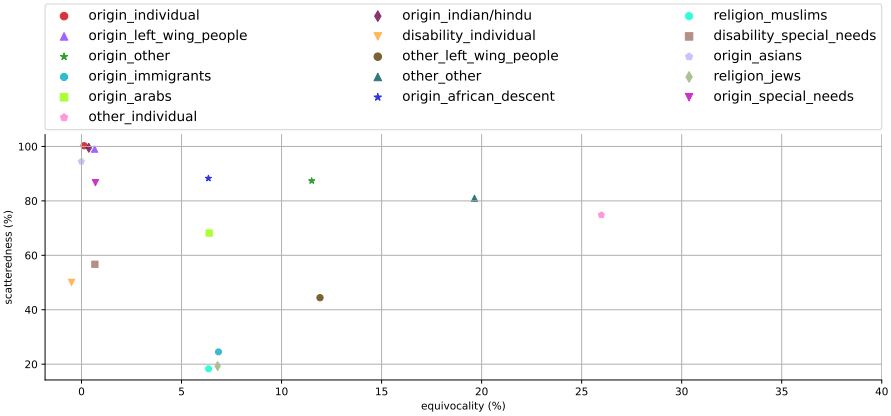


Fig. 8: Dispersion and equivocality of the communities in the *FR hate speech target community* corpus.

Among them 5 are the head of one cluster including *origin_immigrants*, *origin_arabs*, *origin_african_descent*, *religion_muslims* and *religion_jews* while the others are the head of multiple clusters. *other_individual* reaches the highest number of clusters being the head of 4 different clusters representing 25.0% of the total number of clusters. The second community obtaining the highest degree of equivocality is *other_other* representing respectively 18.7% of cluster-heads. Considering the scatteredness, we can observe that some communities being cluster-heads are also assimilated to other clusters. For instance, *origin_african_descent* and *origin_other* are above 70% of scatteredness. Entries composing both of these two clusters occur frequently in clusters having *other_individual* and *other_other* as head. 'les renois' (EN: niggers) and 'attardé mentaux' (EN: retarded) are the most frequent patterns extracted from the cluster mixing *origin_african_descent* and *other_individual* while 'cet attardé' (EN: this retarded) and '@user mongol' (EN: @user mogolian) are the most redundant ones in the generated cluster combining *other_other* and *other_individual*. These findings highlight that both ground-truth communities use frequently a vocabulary based on common slurs and demeaning expressions use to target individuals belonging to the *other_individual* community. In parallel, 7 ground-truth communities are fully assimilated into the generated clusters regarding various degrees of scatteredness. Among them, *origin_individual*, *origin_left_wing_people* and *origin_indian/hindu* reach 100% of scatteredness showing that tweets composing these ground-truth communities are also scattered throughout all the clusters. *origin_asians* and *origin_special_needs* are also highly dispersed with a degree of scatteredness above 80%. Considering both of these aspects, we can observe that the cluster-head *religion_muslims* and *religion_jews* are the less equivocal and scattered ground-truth community. Indeed, these communities assimilate only entries belonging to their communities or from *origin_arabs* for *religion_muslims* and from *religion_muslims* for

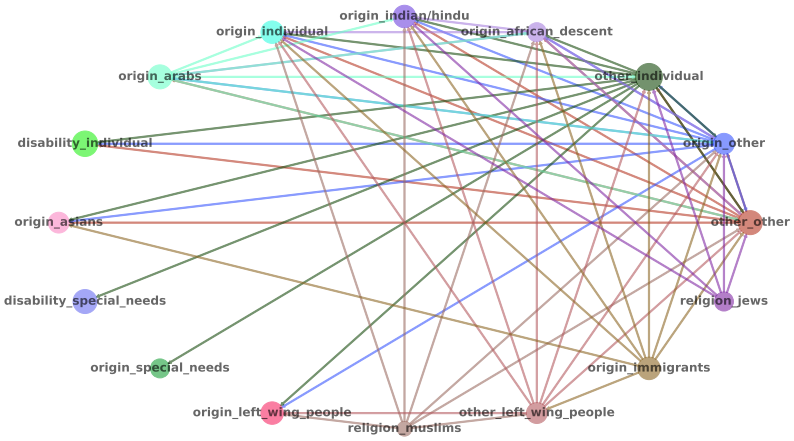


Fig. 9: Visualisation of the mixing of the communities within the generated clusters for the *FR hate speech target community* corpus. The bigger the nodes, the bigger the communities occur as a cluster-head in the generated partition. Nodes refer to ground-truth communities and edges link towards communities occurring in the given cluster-head.

religion_jews. From these clusters, we can observe common slurs specific to each religion. For instance, 'danger de l'islam' (EN: danger of Islam) and 'source du terrorisme' (EN: source of terrorism) for the cluster corresponding to the ground-truth community *religion_muslims* and 'sale juif' (EN: kike) and 'antisémitisme et complotiste' (EN: antisemitism and conspiracy) for the cluster corresponding to the ground-truth community *religion_jews*. From the visualisation of the mixing of the communities' semantic spaces presented in Figure 9, we can observe that both vocabularies related to the communities *origin_immigrants* and *other_left_wing_people* are used to convey hate against people based on their origin. Other prominent mixing can be noticed including the community *other_individual* targeting individuals using vocabularies and slurs related to the target *disability* and the group *special_needs*.

Whilst data are highly imbalanced, explaining the assimilation of some ground-truth communities in favor of the multiplication of others, applying the proposed pipeline has unveiled new insights improving the understanding on how hate is conveyed. From this study, we can confirm the relevance of the MLMA facets to unveil sub-semantic spaces reflecting the nature of offensive comments expressed towards the defined attributes. However, from the observed

behaviours of the ground-truth communities within the generated clusters we notice difficulties to discriminate among victim-groups and target-types. Indeed vocabularies related to victim-groups, target-types or both can hold specific properties allowing to discriminate between each other or conversely they can share common properties leading to increase the bias. Communities resulting from the partitioning rely on two kinds of phenomenons: clusters assimilating ground-truth communities relying on vocabularies mostly composed of common slurs and demeaning expressions or clusters mixing facets based on complementary vocabularies. For instance, the target-type *disability* and the victim-group *specNeeds* composed of offensive messages targeting people with special needs constitute a fully-fledged sub-semantic space specific to these attributes. However, this vocabulary is mostly composed of slurs and demeaning expressions used to discriminate or insult victim-groups, target-types or both such as the target-type *origin* in both languages and more particularly the victim-group *arabs* in French (e.g.: '@user c'est pas à toi que je parle gros mongol. t'as pris le melon sale arabe', (EN: '@user I am not talking to you retard. you are big-headed raghead')) and the community *gender_women* in English (e.g.: 'smh women really retarded @url'). Concerning the complementarity between sub-semantic spaces, we have observed bridges built between some spaces leading to a precise characterisation of the type of hate speech expressed frequently towards victim-groups, target-types or both. For instance, clusters have emerged in English combining vocabularies from the community *sexual_orientation_special_needs* and *gender_women* (e.g.: 'keep your mouth shut you retard bitch') or *other_left_wing_people* and *religion_muslims* in French (e.g.: '@user ferme ta gueule islamo gauchiste', (EN: 'shut the fuck up you libtard')). In addition, we have also observed that communities based on victim-groups, target-types or both referring to *other* (i.e., the facets gathering tweets which do not correspond to the specific target-types / victim-groups defined by MLMA) federate multiple clusters in both languages. This finding allows to establish that from these facets either new ones emerged allowing to capture other target-types / victim-groups or a finer-grained target community taxonomy exists based on subdivisions of the given ground-truth communities.

8 Conclusion

In this paper, we have presented a complete pipeline addressing the task of fine-grained hate speech target community detection adopting a clustering approach. Leveraging the last advances in clustering, the proposed pipeline is based on the MVSC-CEV algorithm which performs a simultaneous clustering of multiple data views resulting in a consensus partition of the data. We have explored the use of different language modelling resources to derive different types of data views exhibiting different properties (i.e., syntactic/semantic and relationship information). In total 69 experiments were conducted, both on the *FR hate speech target community* corpus and on the *EN hate speech target community* corpus, evaluated against state-of-the-art clustering techniques. As a result, we

showed that the majority of the models relying on the MVSC-CEV algorithm achieves the best results on both languages and on the tested evaluation metrics. More particularly, the {mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET} model outperforms all the other models reaching the best balance considering all the evaluation metrics on both languages. Besides that, through the use of clustering techniques we enabled the study of mesoscopic structures of the data to unveil insights on how the hate is conveyed against target communities through the textual messages. Our results on the MLMA ground-truth communities showed the ability of the proposed pipeline to generate clusters corresponding to sub-semantic spaces reflecting the nature of offensive comments related to the defined facets. Although the generated clusters do not correspond exactly to the MLMA communities, they unveil new information allowing to investigate hate speech properties related to specific victim-groups, target-types or both. To conclude, this study has proven the possibility to transpose the task of fine-grained hate speech detection into a clustering problem by its ability to address current challenges in the field, i.e., capturing complex hate speech phenomena using unsupervised methods which are more appropriate to deal with social data. An API allowing to test the different view configurations on both languages is available online[7].

The findings resulting from the conducted study open also multiple research directions: first, developing improved clustering-oriented solutions to address this task and, (2) leveraging communities to derive auxiliary features aiming at supporting downstream tasks (e.g., classification and misogyny detection). More generally, we hope our efforts based on unsupervised learning clustering techniques will pave the road to achieving an unsupervised pipeline estimating optimal cluster number. Expectations from automating this step include to establish whether the MLMA multi-aspect analysis reflects all the facets allowing to describe accurately hate speech phenomena on social media. In future work, we intend to evaluate the proposed pipeline on the task of *fine-grained hate speech detection* by including non-hostile tweets in order to establish its ability to deal with *real-world* data.

Acknowledgments. This work is supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and the EFELIA Côte d'Azur project ANR-22-CMAS-0004.

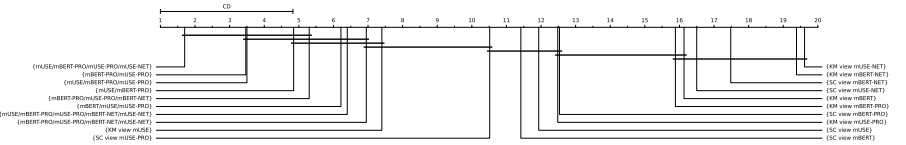
Appendix A Experiments

[7] http://134.59.134.227/demo/index.html Date of access: 19th November 2021.

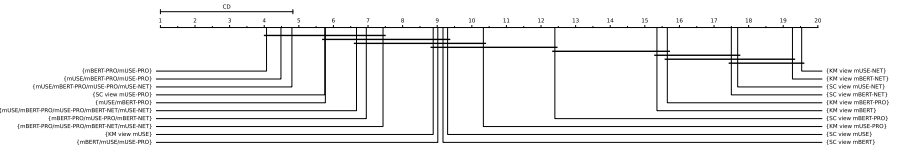
Table A1: Detailed results w.r.t. Purity, ARI and NMI. The mean and standard deviation from 31 independent runs are reported for each language. According to the *post hoc Nemenyi test* the best values reported for each metric are in bold while the highlighted rows refer to the best models considering all the metrics for each language.

Single-view configs.	EN			FR		
	K-Means & Spectral clustering					
	Purity	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**
{KM view mBERT}	0.286±0.006	0.053±0.005	0.128±0.008	0.258±0.003	0.039±0.003	0.121±0.003
{KM view mBERT-PRO}	0.299±0.007	0.045±0.007	0.139±0.008	0.254±0.010	0.040±0.007	0.113±0.009
{KM view mBERT-NET}	0.191±0.004	0.002±0.004	0.045±0.011	0.177±0.005	0.000±0.000	0.026±0.007
{KM view mUSE}	0.372±0.007	0.096±0.006	0.259±0.004	0.348±0.004	0.093±0.007	0.229±0.003
{KM view mUSE-PRO}	0.336±0.012	0.088±0.010	0.218±0.011	0.324±0.014	0.091±0.011	0.197±0.013
{KM view mUSE-NET}	0.187±0.003	0.001±0.002	0.033±0.009	0.183±0.007	0.000±0.001	0.032±0.009
{SC view mBERT}	0.352±0.003	0.106±0.009	0.234±0.004	0.306±0.000	0.071±0.000	0.178±0.000
{SC view mBERT-PRO}	0.347±0.000	0.094±0.000	0.238±0.000	0.300±0.001	0.065±0.001	0.182±0.001
{SC view mBERT-NET}	0.251±0.000	0.013±0.000	0.151±0.000	0.237±0.000	0.011±0.000	0.118±0.000
{SC view mUSE}	0.344±0.000	0.090±0.001	0.242±0.000	0.339±0.000	0.101±0.000	**0.237±0.000**
{SC view mUSE-PRO}	0.345±0.000	0.101±0.000	0.234±0.000	0.342±0.002	0.103±0.001	0.231±0.003
{SC view mUSE-NET}	0.232±0.000	0.004±0.000	0.131±0.000	0.291±0.000	0.031±0.000	0.166±0.000
Multi- view configs.	MVSC					
	Purity	**ARI**	**NMI**	**Purity**	**ARI**	**NMI**
{mBERT/mUSE}	0.348±0.004	0.080±0.003	0.238±0.002	0.347±0.002	0.094±0.001	0.217±0.001
{mBERT/mBERT-PRO}	0.320±0.004	0.065±0.003	0.181±0.003	0.302±0.003	0.069±0.001	0.155±0.002
{mBERT/mUSE-PRO}	0.357±0.003	0.084±0.002	0.245±0.003	0.353±0.001	0.093±0.002	0.217±0.000
{mBERT/mBERT-NET}	0.226±0.002	0.012±0.000	0.087±0.001	0.260±0.003	0.037±0.001	0.119±0.003
{mBERT/mUSE-NET}	0.233±0.003	0.009±0.003	0.077±0.001	0.290±0.004	0.056±0.002	0.151±0.003
{mUSE/mBERT-PRO}	0.382±0.003	0.098±0.003	0.263±0.003	0.360±0.002	0.107±0.004	0.227±0.002
{mUSE/mUSE-PRO}	0.361±0.003	0.087±0.003	0.243±0.003	0.358±0.001	0.103±0.005	0.218±0.001
{mUSE/mBERT-NET}	0.361±0.004	0.090±0.003	0.250±0.003	0.341±0.001	0.086±0.002	0.210±0.002
{mUSE/mUSE-NET}	0.359±0.003	0.088±0.005	0.246±0.003	0.344±0.003	0.092±0.003	0.219±0.002
{mBERT-PRO/mUSE-PRO}	0.385±0.002	0.101±0.003	0.267±0.002	0.362±0.002	**0.110±0.001**	0.227±0.002
{mBERT-PRO/mBERT-NET}	0.335±0.006	0.078±0.004	0.199±0.004	0.281±0.002	0.048±0.001	0.138±0.001
{mBERT-PRO/mUSE-NET}	0.345±0.006	0.080±0.006	0.206±0.005	0.300±0.002	0.058±0.002	0.157±0.002
{mUSE-PRO/mBERT-NET}	0.362±0.003	0.092±0.004	0.253±0.003	0.339±0.002	0.084±0.001	0.207±0.001
{mUSE-PRO/mUSE-NET}	0.358±0.003	0.087±0.005	0.245±0.003	0.342±0.004	0.082±0.002	0.215±0.001
{mBERT-NET/mUSE-NET}	0.251±0.003	0.016±0.001	0.137±0.002	0.259±0.002	0.016±0.001	0.141±0.003
{mBERT/mUSE/mBERT-PRO}	0.371±0.002	0.092±0.002	0.256±0.002	0.345±0.002	0.094±0.003	0.205±0.002
{mBERT/mUSE/mUSE-PRO}	0.360±0.002	0.085±0.004	0.247±0.002	0.362±0.001	0.104±0.004	0.222±0.002
{mBERT/mUSE/mBERT-NET}	0.360±0.003	0.089±0.005	0.243±0.002	0.337±0.002	0.083±0.001	0.198±0.001
{mBERT/mUSE/mUSE-NET}	0.360±0.004	0.094±0.004	0.246±0.002	0.339±0.002	0.086±0.002	0.214±0.002
{mBERT/mBERT-PRO/mUSE-NET}	0.373±0.003	0.094±0.002	0.260±0.001	0.360±0.002	0.098±0.005	0.218±0.002
{mBERT/mUSE-PRO/mBERT-NET}	0.327±0.004	0.065±0.004	0.179±0.003	0.274±0.006	0.047±0.002	0.134±0.004
{mBERT/mBERT-PRO/mUSE-NET}	0.314±0.003	0.069±0.004	0.173±0.003	0.305±0.001	0.063±0.001	0.164±0.002
{mBERT/mUSE-PRO/mBERT-NET}	0.367±0.002	0.091±0.004	0.247±0.002	0.338±0.005	0.083±0.001	0.199±0.002

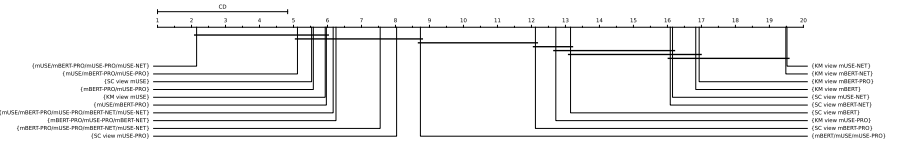
{mBERT/mUSE-PRO/mUSE-NET}	0.363±0.004	0.092±0.005	0.246±0.002	0.340±0.004	0.084±0.003	0.208±0.002
{mBERT/mBERT-NET/mUSE-NET}	0.238±0.001	0.011±0.000	0.099±0.001	0.271±0.004	0.046±0.002	0.136±0.005
{mUSE/mBERT-PRO/mUSE-PRO}	0.384±0.003	0.101±0.004	0.268±0.003	0.364±0.003	0.109±0.003	0.227±0.002
{mUSE/mBERT-PRO/mBERT-NET}	0.378±0.004	0.099±0.004	0.263±0.003	0.351±0.004	0.089±0.002	0.213±0.002
{mUSE/mBERT-PRO/mUSE-NET}	0.380±0.004	0.101±0.005	0.267±0.003	0.357±0.003	0.095±0.002	0.226±0.003
{mUSE/mUSE-PRO/mBERT-NET}	0.362±0.004	0.088±0.003	0.250±0.002	0.359±0.003	0.096±0.005	0.220±0.002
{mUSE/mUSE-PRO/mUSE-NET}	0.360±0.004	0.088±0.006	0.249±0.003	0.355±0.001	0.100±0.005	0.218±0.000
{mUSE/mBERT-NET/mUSE-NET}	0.362±0.004	0.088±0.004	0.247±0.004	0.340±0.003	0.090±0.002	0.216±0.001
{mBERT-PRO/mUSE-PRO/mBERT-NET}	0.391±0.004	**0.108±0.005**	0.273±0.003	0.343±0.005	0.088±0.001	0.216±0.003
{mBERT-PRO/mUSE-PRO/mUSE-NET}	0.380±0.005	0.103±0.004	0.268±0.003	0.351±0.005	0.092±0.001	0.220±0.004
{mBERT-PRO/mBERT-NET/mUSE-NET}	0.361±0.003	0.087±0.003	0.243±0.003	0.285±0.004	0.050±0.001	0.147±0.003
{mUSE-PRO/mBERT-NET/mUSE-NET}	0.331±0.005	0.073±0.003	0.192±0.003	0.344±0.005	0.088±0.001	0.211±0.002
{mBERT/mUSE/mBERT-PRO/mUSE-PRO}	0.372±0.002	0.096±0.004	0.262±0.002	0.361±0.001	0.099±0.004	0.225±0.001
{mBERT/mUSE/mBERT-PRO/mBERT-NET}	0.369±0.004	0.094±0.004	0.255±0.003	0.341±0.003	0.085±0.001	0.198±0.003
{mBERT/mUSE/mBERT-PRO/mUSE-NET}	0.362±0.004	0.097±0.005	0.253±0.002	0.343±0.005	0.086±0.001	0.207±0.003
{mBERT/mUSE/mUSE-PRO/mBERT-NET}	0.360±0.002	0.090±0.003	0.246±0.002	0.345±0.004	0.087±0.001	0.212±0.001
{mBERT/mUSE/mUSE-PRO/mUSE-NET}	0.366±0.003	0.091±0.004	0.252±0.001	0.348±0.004	0.090±0.002	0.214±0.004
{mBERT/mUSE/mBERT-NET/mUSE-NET}	0.363±0.003	0.094±0.005	0.244±0.002	0.328±0.002	0.081±0.001	0.196±0.002
{mBERT/mBERT-PRO/mUSE-PRO/mBERT-NET}	0.363±0.004	0.089±0.003	0.248±0.002	0.335±0.002	0.081±0.001	0.194±0.001
{mBERT/mBERT-PRO/mUSE-PRO/mUSE-NET}	0.368±0.006	0.097±0.004	0.257±0.003	0.337±0.003	0.082±0.001	0.204±0.003
{mBERT/mBERT-PRO/mBERT-NET/mUSE-NET}	0.324±0.004	0.069±0.005	0.176±0.003	0.288±0.004	0.053±0.001	0.146±0.003
{mBERT/mUSE-PRO/mBERT-NET/mUSE-NET}	0.366±0.003	0.094±0.004	0.248±0.002	0.330±0.004	0.081±0.002	0.202±0.003
{mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET}	0.379±0.002	0.099±0.005	0.265±0.003	0.348±0.003	0.095±0.005	0.221±0.002
{mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}	**0.390±0.002**	0.108±0.005	**0.276±0.001**	**0.367±0.001**	0.099±0.005	0.231±0.001
{mUSE/mBERT-PRO/mBERT-NET/mUSE-NET}	0.379±0.005	0.100±0.005	0.266±0.004	0.333±0.005	0.086±0.002	0.206±0.002
{mUSE/mUSE-PRO/mBERT-NET/mUSE-NET}	0.362±0.005	0.089±0.005	0.251±0.004	0.349±0.004	0.093±0.001	0.217±0.002
{mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.383±0.004	0.106±0.006	0.270±0.004	0.343±0.003	0.088±0.001	0.208±0.002
{mBERT/mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET}	0.374±0.003	0.094±0.004	0.260±0.002	0.355±0.002	0.093±0.003	0.218±0.002
{mBERT/mUSE/mBERT-PRO/mUSE-PRO/mUSE-NET}	0.374±0.003	0.098±0.005	0.262±0.002	0.355±0.003	0.094±0.001	0.225±0.001
{mBERT/mUSE/mBERT-PRO/mBERT-NET/mUSE-NET}	0.369±0.005	0.097±0.004	0.257±0.002	0.329±0.002	0.078±0.002	0.194±0.001
{mBERT/mUSE/mUSE-PRO/mBERT-NET/mUSE-NET}	0.364±0.003	0.090±0.003	0.249±0.002	0.343±0.002	0.089±0.002	0.213±0.001
{mBERT/mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.366±0.003	0.097±0.004	0.254±0.002	0.332±0.003	0.081±0.001	0.197±0.002
{mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.380±0.003	0.105±0.005	0.271±0.002	0.350±0.003	0.093±0.003	0.221±0.002
{mBERT/mUSE/mBERT-PRO/mUSE-PRO/mBERT-NET/mUSE-NET}	0.371±0.005	0.098±0.004	0.261±0.005	0.348±0.004	0.090±0.001	0.218±0.002



(a) *purity*



(b) ARI



(c) NMI

Fig. A1: The different hierarchies obtained from models' average ranking for each evaluation metric using the *post hoc Nemenyi test*

References

[1] Nockleby, J.T.: Hate speech. Encyclopedia of the American Constitution **2nd ed.**, 1277–1279 (2000)

[2] Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web **10**(5), 925–945 (2019). https://doi.org/10.3233/SW-180338

[3] Blaya, C.: Cyberhate: A review and content analysis of intervention strategies. Aggression and Violent Behavior **45**, 163–172 (2019). https://doi.org/10.1016/j.avb.2018.05.006. Bullying and cyberbullying: Protective factors and effective interventions

[4] Zhang, Z., Robinson, D., Tepper, J.A.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M. (eds.) The Semantic Web - 15th International Conference, ESWC 2018, Proceedings. Lecture Notes in Computer Science, vol. 10843, pp. 745–760. Springer, Heraklion, Crete, Greece (2018). https://doi.org/10.1007/978-3-319-93417-4_48. https://doi.org/10.1007/978-3-319-93417-4_48

[5] Liu, P., Li, W., Zou, L.: NULI at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, pp. 87–91. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/s19-2011. https://doi.org/10.18653/v1/s19-2011

[6] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: A multilingual evaluation for online hate speech detection. ACM Trans. Internet Techn. **20**(2), 10–11022 (2020). https://doi.org/10.1145/3377323

[7] Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. **51**(4), 85–18530 (2018). https://doi.org/10.1145/3232676

[8] Mossie, Z., Wang, J.: Vulnerable community identification using hate speech detection on social media. Inf. Process. Manag. **57**(3), 102087 (2020). https://doi.org/10.1016/j.ipm.2019.102087

[9] Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V., Patti, V.: Emotionally informed hate speech detection: A multi-target perspective. Cogn. Comput. **14**(1), 322–352 (2022). https://doi.org/10.1007/s12559-021-09862-5

[10] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.: Multilingual and multi-aspect hate speech analysis. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, pp. 4674–4683. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1474. https://doi.org/10.18653/v1/D19-1474

[11] Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Evaluation **55**(2), 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8

[12] Fortuna, P., da Silva, J.R., Wanner, L., Nunes, S., *et al.*: A hierarchically-labeled portuguese hate speech dataset. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 94–104 (2019)

[13] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019). https://aclweb.org/anthology/papers/N/N19/N19-1423/

[14] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Ábrego, G.H., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R.: Multilingual universal sentence encoder for semantic retrieval. In: Celikyilmaz, A., Wen, T. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, pp. 87–94. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-demos.12. https://doi.org/10.18653/v1/2020.acl-demos.12

[15] Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLOS ONE **15**(12), 1–32 (2021). https://doi.org/10.1371/journal.pone.0243300

[16] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., Shrivastava, M.: A dataset of hindi-english code-mixed social media text for hate speech detection. In: Nissim, M., Patti, V., Plank, B., Wagner, C. (eds.) Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL 2018, pp. 36–41. Association for Computational Linguistics, New Orleans, Louisiana, USA (2018). https://doi.org/10.18653/v1/w18-1105. https://doi.org/10.18653/v1/w18-1105

[17] Poletto, F., Basile, V., Bosco, C., Patti, V., Stranisci, M.: Annotating hate speech: Three schemes at comparison. In: Bernardi, R., Navigli, R., Semeraro, G. (eds.) Proceedings of the Sixth Italian Conference on Computational Linguistics. CEUR Workshop Proceedings, vol. 2481. CEUR-WS.org, Bari, Italy (2019). http://ceur-ws.org/Vol-2481/paper56.pdf

[18] Nascimento, G., Carvalho, F., da Cunha, A.M., Viana, C.R., Guedes, G.P.: Hate speech detection using brazilian imageboards. In: dos Santos, J.A.F., Muchaluat-Saade, D.C. (eds.) Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019, pp. 325–328. ACM, Rio de Janeiro, Brazil (2019). https://doi.org/10.1145/3323503.3360619. https://doi.org/10.1145/3323503.3360619

[19] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, pp. 54–63. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/s19-2007. https://doi.org/10.18653/v1/s19-2007

[20] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/s19-2010. https://doi.org/10.18653/v1/s19-2010

[21] Alrehili, A.: Automatic hate speech detection on social media: A brief survey. In: 16th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, pp. 1–6. IEEE Computer Society, Abu Dhabi, UAE (2019). https://doi.org/10.1109/AICCSA47632.2019.9035228. https://doi.org/10.1109/AICCSA47632.2019.9035228

[22] Themeli, C., Giannakopoulos, G., Pittaras, N.: A study of text representations in hate speech detection. CoRR **abs/2102.04521** (2021) https://arxiv.org/abs/2102.04521

[23] Malmasi, S., Zampieri, M.: Challenges in discriminating profanity from hate speech. J. Exp. Theor. Artif. Intell. **30**(2), 187–202 (2018). https://doi.org/10.1080/0952813X.2017.1409284

[24] MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: Challenges and solutions. PLOS ONE **14**(8), 1–16 (2019). https://doi.org/10.1371/journal.pone.0221152

[25] Vigna, F.D., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on facebook. In: Armando, A., Baldoni, R., Focardi, R. (eds.) Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). CEUR Workshop Proceedings, vol. 1816, pp. 86–95. CEUR-WS.org, Venice, Italy (2017). http://ceur-ws.org/Vol-1816/paper-09.pdf

[26] Tian, Z., Kübler, S.: Offensive language detection using brown clustering. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pp. 5079–5087. European Language Resources Association, ??? (2020). https://aclanthology.org/2020.lrec-1.625/

[27] Davidson, T., Warmsley, D., Macy, M.W., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017., pp. 512–515 (2017)

[28] Vogel, I., Meghana, M.: Profiling hate speech spreaders on twitter: SVM vs. bi-lstm. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 2936, pp. 2193–2200. CEUR-WS.org, Bucharest, Romania (2021). http://ceur-ws.org/Vol-2936/paper-196.pdf

[29] Asogwa, D.C., Chukwuneke, C.I., Ngene, C.C., Anigbogu, G.N.: Hate speech classification using SVM and naive BAYES. CoRR **abs/2204.07057** (2022) https://arxiv.org/abs/2204.07057. https://doi.org/10.48550/arXiv.2204.07057

[30] Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Kumar, R., Ojha, A.K., Zampieri, M., Malmasi, S. (eds.) Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, pp. 1–11. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://aclanthology.org/W18-4401/

[31] Indurthi, V., Syed, B., Shrivastava, M., Gupta, M., Varma, V.: Fermi at SemEval-2019 task 6: Identifying and categorizing offensive language in social media using sentence embeddings. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 611–616. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/S19-2109. https://aclanthology.org/S19-2109

[32] Koufakou, A., Pamungkas, E.W., Basile, V., Patti, V.: HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 34–43. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.alw-1.5. https://aclanthology.org/2020.alw-1.5

[33] Caselli, T., Basile, V., Mitrovic, J., Granitzer, M.: Hatebert: Retraining BERT for abusive language detection in english. CoRR **abs/2010.12472** (2020) https://arxiv.org/abs/2010.12472

[34] Magu, R., Luo, J.: Determining code words in euphemistic hate speech using word embedding networks. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 93–100. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/W18-5112. https://aclanthology.org/W18-5112

[35] Nobata, C., Tetreault, J.R., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pp. 145–153 (2016)

[36] Jain, M., Goel, P., Singla, P., Tehlan, R.: Comparison of various word embeddings for hate-speech detection. In: Khanna, A., Gupta, D., Pólkowski, Z., Bhattacharyya, S., Castillo, O. (eds.) Data Analytics and Management, pp. 251–265. Springer, Singapore (2021)

[37] Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10. Association for Computational Linguistics, Valencia, Spain (2017). https://doi.org/10.18653/v1/W17-1101. https://aclanthology.org/W17-1101

[38] Meena, P., Pawar, M., Pandey, A.: A survey on community detection algorithm and its applications. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**(6), 4807–4815 (2021)

[39] Mohamed, M.M.: Clustering halal food consumers: A twitter sentiment analysis. International Journal of Market Research **61**(3), 320–337 (2019) https://arxiv.org/abs/https://doi.org/10.1177/1470785318771451. https://doi.org/10.1177/1470785318771451

[40] Wang, A., Gao, X.: Multifunctional product marketing using social media based on the variable-scale clustering. Tehnički vjesnik **26**(1), 193–200 (2019)

[41] Kingston, C., Nurse, J.R.C., Agrafiotis, I., Milich, A.: Using semantic clustering to support situation awareness on twitter: the case of world

views. Hum. centric Comput. Inf. Sci. **8**, 22 (2018). https://doi.org/10.1186/s13673-018-0145-6

[42] Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J.: An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. Inf. Process. Manag. **57**(2), 102034 (2020). https://doi.org/10.1016/j.ipm.2019.04.002

[43] Bickel, S., Scheffer, T.: Multi-view clustering. In: Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), pp. 19–26. IEEE Computer Society, Brighton, UK (2004). https://doi.org/10.1109/ICDM.2004.10095. https://doi.org/10.1109/ICDM.2004.10095

[44] de A.T. de Carvalho, F., Lechevallier, Y., Despeyroux, T., de Melo, F.M.: In: Guillet, F., Pinaud, B., Venturini, G., Zighed, D.A. (eds.) Multi-view Clustering on Relational Data, pp. 37–51. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-02999-3_3. https://doi.org/10.1007/978-3-319-02999-3_3

[45] Chao, G., Sun, S., Bi, J.: A survey on multi-view clustering. CoRR **abs/1712.06246** (2017) https://arxiv.org/abs/1712.06246

[46] Yang, Y., Wang, H.: Multi-view clustering: A survey. Big Data Min. Anal. **1**(2), 83–107 (2018). https://doi.org/10.26599/BDMA.2018.9020003

[47] Fu, L., Lin, P., Vasilakos, A.V., Wang, S.: An overview of recent multi-view clustering. Neurocomputing **402**, 148–161 (2020). https://doi.org/10.1016/j.neucom.2020.02.104

[48] Kanaan-Izquierdo, S., Ziyatdinov, A., Perera-Lluna, A.: Multiview and multifeature spectral clustering using common eigenvectors. Pattern Recognit. Lett. **102**, 30–36 (2018). https://doi.org/10.1016/j.patrec.2017.12.011

[49] Rout, N., Mishra, D., Mallick, M.K.: Handling imbalanced data: A survey. In: Reddy, M.S., Viswanath, K., K.M., S.P. (eds.) International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications, pp. 431–443. Springer, Singapore (2018)

[50] Ramírez-Gallego, S., Krawczyk, B., García, S., Wozniak, M., Herrera, F.: A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing **239**, 39–57 (2017). https://doi.org/10.1016/j.neucom.2017.01.078

[51] Pradha, S., Halgamuge, M.N., Vinh, N.T.Q.: Effective text data preprocessing technique for sentiment analysis in social media data. In: 11th International Conference on Knowledge and Systems Engineering, KSE 2019, Da Nang, Vietnam, October 24-26, 2019, pp. 1–8. IEEE, ??? (2019).

https://doi.org/10.1109/KSE.2019.8919368. https://doi.org/10.1109/KSE.2019.8919368

[52] Ollagnier, A., Williams, H.T.P.: Sequential transfer learning for event detection and key sentence extraction. In: Wani, M.A., Luo, F., Li, X.A., Dou, D., Bonchi, F. (eds.) 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020, pp. 1023–1027. IEEE, Miami, FL, USA (2020). https://doi.org/10.1109/ICMLA51294.2020.00166. https://doi.org/10.1109/ICMLA51294.2020.00166

[53] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. CoRR **abs/2003.08271** (2020) https://arxiv.org/abs/2003.08271

[54] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., Kurzweil, R.: Universal sentence encoder. CoRR **abs/1803.11175** (2018) https://arxiv.org/abs/1803.11175

[55] Vijaymeena, M., Kavitha, K.: A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal **3**(2), 19–28 (2016)

[56] Fortunato, S.: Community detection in graphs. Physics Reports **486**(3), 75–174 (2010). https://doi.org/10.1016/j.physrep.2009.11.002

[57] Trendafilov, N.T.: Stepwise estimation of common principal components. Comput. Stat. Data Anal. **54**(12), 3446–3457 (2010). https://doi.org/10.1016/j.csda.2010.03.010

[58] MacQueen, J., *et al.*: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA

[59] Park, H., Jun, C.: A simple and fast algorithm for k-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009). https://doi.org/10.1016/j.eswa.2008.01.039

[60] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, pp. 849–856. MIT Press, Vancouver, British Columbia, Canada (2001). https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[61] Manning, C., Raghavan, P., Schütze, H.: Introduction to information

retrieval. Journal of the American Society for Information Science and Technology **1**, 496 (2008)

[62] Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)

[63] Levene, H.: Robust tests for equality of variances. Stanford University Press, 278–292 (1960)

[64] Wilcoxon, F.: In: Kotz, S., Johnson, N.L. (eds.) Individual Comparisons by Ranking Methods, pp. 196–202. Springer, New York, NY (1992). https://doi.org/10.1007/978-1-4612-4380-9_16. https://doi.org/10.1007/978-1-4612-4380-9_16

[65] Milton, F.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association **32**(200), 675–701 (1937) https://arxiv.org/abs/https://www.tandfonline.com/doi/pdf/10.1080/01621459.1937.10503522. https://doi.org/10.1080/01621459.1937.10503522

[66] Nemenyi, P.: Distribution-free Multiple Comparisons. Princeton University, ??? (1963). https://books.google.fr/books?id=nhDMtgAACAAJ