Graph-based Approach for Semantic Text Clustering: Topic Detection in the Context of a Large Multilingual Public Consultation

Anonymous ACL submission

Abstract

We present a novel algorithm for multilingual text clustering built upon two well studied techniques: multilingual aligned embedding and community detection in graphs. The aim of our algorithm is to discover underlying topics in a multilingual dataset using clustering. We present both a numerical evaluation using silhouette and V-measure metrics, and a qualitative evaluation for which we propose a new systematic approach. Our algorithm presents robust overall performance and its results were empirically evaluated by an analyst. The work we present was done in the context of a large multilingual public consultation, for which our new algorithm was deployed and used on a daily basis.

1 Introduction

Making sense of vast amounts of text from different authors is a typical Natural Language Processing task which dates back from the early days of the Social Networks (Fu et al., 2008), with use cases as diverse as crisis response management (De Longueville et al., 2009), collection of scientific survey data (Pierce et al., 2009) or political elections outcome predictions (Chung and Mustafaraj, 2011). In this context, mining opinions and arguments on a given online conversation has emerged as a research field on its own (Liu and Zhang, 2012), to which this paper aims to contribute.

The Conference on the Future of Europe (CoFE) is a pan-European democratic exercise that contributes shaping the future of the European Union by collecting and debating proposals on the evolution of the European Union from citizens[1]. The exercise includes a system of Plenary Meetings, which involve a mix of parliaments members, government representatives and citizens, as well as Citizens Panels - each with a dedicated topic such as protection of the environment or social justice-, which are real-life debates of 200 randomly selected citizens representing the EU diversity (in terms of country of origin, age, gender, urban/rural living environment, education level, ...). Last but not least, CoFE provides a Multilingual Digital Platform[2] where any European can share proposals, as well as endorse or comments other's proposals. The contents of this online platform provides the multilingual text studied in this paper (see section 3). To be noted: the text corpus used in this paper does not contain any personal data as defined in the CoFE platform's Privacy Statement. In other words, only publicly available text - without any reference to author's identification - is considered in this research.

In order to make sense of the overall conversation, involving at time writing several tens of thousands of individual contributions in 24 languages, analysts need then:

- to have some form of overview of proposals showing how they relate to each other;
- to identify recurring proposals and comments posted;
- to characterise such proposals (how significant are they in the overall conversation? are they widely supported/opposed to by other participants?).

The overall approach to address this need is the one of semantic clustering of sentences. The choice to work at sentence level comes from the nature of the text posted on the CoFE platform ; although different by nature, both proposals and comments can contain relevant 'idea-laden' sentences. In the

[1] https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/conference-future-europe_en

[2] https://futureu.europa.eu/

case of comments, it can be counter-proposals (e.g. "I think that instead of banning plastic packaging, EU should promote truly sustainable recycling") or further proposals (e.g. "EU Army is fine, but then it needs a commander in chief - a President - elected through direct universal suffrage"). Sentences with other rhetorical functions are also widely present in the text (examples, supporting arguments, judgements, expressions of agreement/disagreement...), so it is expected that working at sentence level would allow identify more easily the core proposal within any contribution.

Grouping similar proposals posted in CoFE proposals and comments poses however a double challenge. Firstly, different contributors would convey the same meaning in different ways ("EU should have a common language", "EU should adopt a single official language"), and a similar proposal can have infinite (important) nuances (is this common language meant to be English, Esperanto or another one? what status would it have such common/single language vis-à-vis national/local languages? ...). It is thus important to measure proposals similarity on a continuous relative scale, rather than as a classification task. By organising text based on their semantic similarity, it is expected that similar proposals will emerge as clusters that can be further analysed to get a nuanced picture of the related conversations present on the CoFE platform.

Secondly - and this is one of the key specificities of the data at hand -, the contents posted on the CoFE platform is highly multilingual, with the use of all 24 EU official languages being promoted and supported by the user interface, and instant machine translation to any other EU language is being offered to allow some form of interactions between participants using different languages. The multilingual aspect of the text corpus studied is thus an important driver for the choice of methods, although the outcomes of this paper are also relevant in a mono-lingual setting.

2 Related works

Our contribution repose on original combination of two well know techniques: embedding based representation of sentences, and community detection over graphs. In the field of Natural Language Processing (NLP) community detection algorithms have already been used. It has been frequently used at the level of words-based graph, such as in (Jurgens, 2011) it has been used to perform word sense induction or in (de Arruda et al., 2016; Gerlach et al., 2018; Hamm et al., 2021) where it is used for topic detection. However, our work distinguishes itself from these in that it consider embedding-based distances and is at the sentence level. In (Boltužić and Šnajder, 2015) in the context of online debates try to identify salient arguments by using the text similarity between arguments by considering both the bag-off-word and embeddings based semantic distance followed by a clustering step, however the embedding used are monolingual, and they do not build a graph representation. Similarly, (Sawhney et al., 2017) described the uses of averaged word embeddings, which is a crude representation of sentences as well compare against Louvain community detection which is found the best performing. Our approach is different in its use of more advanced multilingual aligned embedding and in its cluster filtering step.

Work around cluster evaluation metrics revolves around numerical evaluation (Chakraborty et al., 2017) and finding empirically good parameters (Arinik et al., 2021), while we propose a new subjective evaluation methodology from the point of view of end-users needs.

3 Data: COFE data specificity

There are several ways citizens can interact with the CoFE platform : by submitting an proposal (i.e. a text of about 2000 characters explaining and justifying a desirable evolution of the European Union on a given topic), by commenting other people's proposals, by endorsing other people's proposals. CoFE-affiliated events (both held online or in real-life) can also be registered on the platform, in which case the event description and final report are also available as full text on the platform. For the clarity of the argument, we will only refer to proposals'- and comments' text in this paper, although events-related text is also considered in the overall analysis of the CoFE outcomes. This has a direct impact on the data analysis with a strong multilingual aspect as a crucial specificity. Any user can write a proposal, or a comment on an existing proposal in one of the 24 EU languages (some even used languages out of these 24 languages, like Esperanto), and thanks to the automatic translation of each proposal and each comment, any user can read any proposal or comment in her/his own languages (if part of the 24 EU languages). Less that 40%

of the proposals are written in English, together with the other main languages used: French and German, they amount to 2/3 of all the proposals, indicating a strong need to multilingual solutions. The CoFE data contains about 41 000 textual contributions, including 18 800 proposals and 22 200 comments. A publicly available version of this dataset is accessible[3]

4 New Approach for Multilingual Sentence Clustering

The clustering approach we propose is a generic multistep process that does not depend on the specific technology used to implement each of these step, as these could be freely replaced by other implementation without changing the core of the approach with propose:

1. SPLITING: The text of every proposal is split in actual sentences, taking into account quotations marks and polysemy of punctuation in order to always process full sentences

2. EMBEDDING: The embedding of each sentence is computed, using aligned multilingual embeddings

3. GRAPH BUILDING: The pairwise similarity between every pair of sentence is computed, an edge is added in the graph between two sentences only if the similarity is above a given threshold;

4. CLUSTERING: The sentence graph is clustered using a community detection algorithm

5. FILTERING: Each cluster, or community, is further processed in the following way:

   (a) The centroid vector of the cluster is computed

   (b) Sentences belonging to the cluster are ranked according to distance to the centroid

   (c) Sentence bellow a given threshold are discarded from the cluster

   (d) For each language the sentence closest to the centroid is used to represent that cluster when summarizing the cluster information

[3] https://data.europa.eu/data/datasets/conference-on-the-future-of-europe

The technological choice we did for each steps of the approach are the following, along with their parameters when relevant:

- splitting is done using an in-house sentence splitter respecting the described requirements, as we did not found existing libraries supporting them;

- embedding is done using LASER embeddings (Artetxe and Schwenk, 2019), aligned multilingual sentence embeddings, that had the best performance in our experiments. The `laserembeddings` Python library was used;

- the graph is built by making full pairwise comparison between all pairs of sentences, the threshold is fixed at 0.8 - which is a low threshold in that it can capture spurious similarities and not only actual ones. The `networkx` Python library was used;

- clustering is done using the Louvain clustering algorithm (Blondel et al., 2008), with the `resolution` parameter set to 1. The `community` Python library was used;

- filtering is done using a minimal similarity threshold fixed at 0.85;

Due to the high cost of evaluating the quality of clustering for different parameters and in the absence of training data, all the parameters were found by human trial and error by trying to find what seemed a good balance to the experts between cluster precision and cluster quality. The quality of the cluster obtain with this fixed set of parameters is then rigorously evaluated in section 5.

5 Evaluation

For illustrative purpose and in order to better understand the algorithm and the cluster, we provide two illustrations: on Figure 1 we illustrate how the sentence graph looks like in a simplified example containing the most highly similar sentences, by setting the similarity parameter at 0.95. The largest communities found by the Louvain algorithm are highlighted in different colours. In Figure 2 we provide an example of a multilingual cluster.

5.1 Quantitative Evaluation

In order to evaluate qualitatively the clusters, we performed an experimental evaluation on a random

subset of 8k proposals from which we extract the titles. The sentences of this dataset have an average length of 50 chars and of 7 words. Using the above specified parameters, our algorithm produces 116 clusters. In order to evaluate the quality of this clustering, we use two metrics: silhouette and the V-measure. We our going to compare our approach, that we will refer to as "semantic community", or "SC" for short, to two other approaches: 1) "kmeans" or "KM" for shorts: kmeans (Lloyd, 1982) will be used as a baseline; and 2) "user generated category" or "UGC" for short: we will use the label under which a proposal has been posted as the cluster to which belong all the sentences and the title of that proposal. As such for UGC there only 10 possible clusters. Because the order of magnitude of the number of clusters returned by SC is 100, we will run kmeans for K=10 and K=100 in order to compare it fairly to the two other approaches.

We evaluate SC and KM twice by considering the version with and without post-filtering. When we run SC, we evaluate UGC over the precise subset of sentences that where clustered by SC. Note that when KM is run KM with filtering it also clusters only a subset of the sentences. Finally, we evaluate UGC and KM on the whole dataset.

All results are reported in Table 1. It is to be noted that after our clustering proceeds, half of the sentences are not part of any cluster and are as such discarded; moreover the cluster filtering removes a further half of the remaining data point, yielding for SC clusters containing only a quarter of the original data.

Silhouette (Rousseeuw, 1987) is an unsupervised clustering metrics that consider only the cluster and a distance metrics between each data point, which is here the semantic distance between two sentences defined as the cosine similarity between the embedding representation of these sentence. Silhouette score goes from -1 to 1.

We can see that SC with filtering (SC-F), which is the final output of our algorithm, has the highest silhouette score of all compared algorithms, this is no surprise as the filtering stage actually optimize the very measure of the silhouette score. Using filtering improves the score by about 9 points with respect to the non filtered version (SC-NF). Filtering also improves the performance of KM but not as strongly as it does for SC: 5 and 7 points for K=100 and K=10 respectively. This indicates by itself filtering is an useful operation independently of the algorithm used. SC-NF is close to 0, which shows that filtering is crucial to get a better clustering.

Comparing SC and UGC over the same subsets ss1 and ss2 show consistently that UGC, the user generated label have a worst score than SC, showing that our algorithm is slightly better at grouping information together than the user hand made classification does. This effect could be a side effect that several proposals are actually at the intersection of two or more categories, but the user had to select only one under which to reference their post, while our algorithm goes at sentence level, able to put back together the parts related to other categories and consequently improving the silhouette score.

V-measure (Rosenberg and Hirschberg, 2007) is a supervised clustering metrics requiring the use of a ground truth to compare a clustering against hat ground truth. Because it is not humanly feasible to cluster the vast quantity of data at hand in order to provide a ground truth, we consider the use of the following clustering: all sentences from all the proposal from a given category are considered belonging to the same cluster. As such, the UGC clusterings has the highest possible score, as it is the exact same clustering as the ground truth.

SC-NF has the highest V-measure, while SC-F has the second highest. This means that the small clusters eliminated at the filtering stage were more coherent with the ground truth than the remaining ones. However because they are so small and numerous (several thousands), they are not interesting from an analyst point of view that needs a synthetic view of the overall data.

SC-F has a V-score 10 points higher that KM-F K=100 for a similar number of clusters. SC-NF V-score is even 23 points higher than KM-NF K=100, its kmeans counterpart. This indicate that our approach is significantly better than kmeans in comparable settings. Applying filtering to kmeans improves its performance by a maximum of 10 points.

The algorithm SC-F provide the overall best performances in terms of silhouette score and of V-score.

5.2 Qualitative Evaluation

The qualitative evaluation aimed to assess the suitability of the clustering method to satisfy a real life Use Case, namely : Topic Mining a highly

subset	algorithm	filtering	tot. sent.	% sent.	tot. clust.	avg. silhhouette	V-measure
ss1	SC-F	yes	1924	24.0 %	116	0.134	0.350
ss1	UGC	no	1924	24.0 %	10	-0.076	1.0
ss2	SC-NF	no	5666	70.6 %	3673	0.042	0.384
ss2	UGC	no	5666	70.6 %	10	-0.012	1.0
ss3	KM-F K=10	yes	1013	12.6 %	9	0.078	0.115
ss4	KM-F K=100	yes	3251	40.5 %	97	0.050	0.254
full	UGC	no	8016	100.0 %	10	-0.023	1.0
full	KM-NF K=10	no	8016	100.0 %	10	-0.006	0.042
full	KM-NF K=100	no	8016	100.0 %	100	0.003	0.151

Table 1: Sentence coverage and clustering quality measures for 3 different clustering

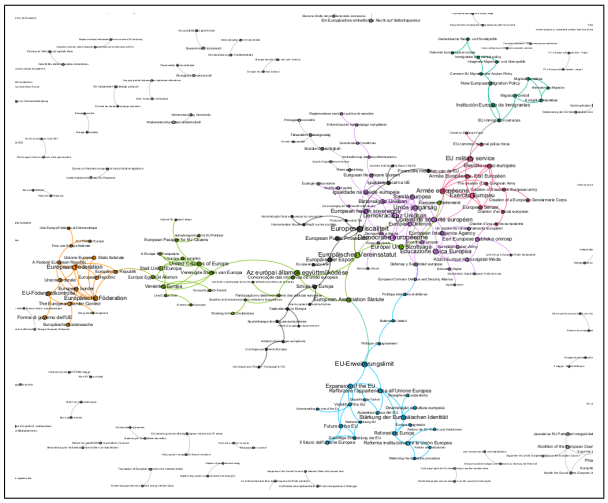

Figure 1: Example of a sentence-graph where the largest communities found by Louvain algorithm are also highlighted with different colours

multi-lingual citizen consultation. This is a good complement to the quantitative analysis. (Costa and Ortale, 2021) As explained in the introduction, we define in this context the Topic Mining task as: a) Grouping similar proposals, and b) provide a human-understandable summary for this group of similar proposals. To this end, we asked Analysts to review 116 clusters based on a set of input from a given month (October 2021). These clusters represent, for each CoFE Theme (see Table 2), all clusters with a size above 1% of the total of sentences. If there were more than 15 clusters above this size threshold, then only the 15 biggest ones were reviewed. For each cluster, the analyst was required to answer, the following questions:

Q.1 Do the sentences in the clusters are consistent with each other? (5 = very consistent, 1 = very inconsistent) (automated English translations were also provided to the analysts)

Q.2 Based on these sentences, how easy is it to draft a title for this cluster? (5 = very easy, 1 = very uneasy)

As we can see in Table 2, the qualitative analysis gives overall positive results: the analysts considered that the sentence clusters were quite consistent (score 4.28/5) and that was quite easy to tell what topic the cluster was about by giving it a title (score 4.03/05). Which tends to demonstrate the fitness-for-purpose of the method. To be noted, the title examples are provided for illustration purposes, and do not represent the outcome of the Conference on the Future of Europe[4].

Some caveats need nevertheless to be underlined. Firstly, the long tail of small clusters (those for which the size was less than 1% of the number of sentences to be clustered) was not reviewed ; we expect lower quality for those, so the scores reflect only the quality of clusters that an analyst will use in his Topic Mining task, and not the quality of the clustering process stricto-sensu. Secondly, there is no clear explanation about the relative differences between themes (ranging from 3.25 for Q1 for European Democracy to 5.00 for Climate Change and Environment, and from 3.00 for Q2 for health to 5.00 for Climate Change and Environment). The sample being relatively limited in size, comparing themes may not be significant in statistical terms ; nevertheless, it seems that the more clusters have been reviewed for a theme – and therefore, by the very selection method, the smoother is the distribution of clusters by size – the higher the quality is for an analyst.

As a conclusion, the qualitative evaluation confirmed the fitness-for-purpose of the method for

[4]Final outcome of CoFE can be found here : https://futureu.europa.eu/en/pages/reporting

Theme	# clstr	Q1 avg	Q2 avg	Titles examples
Health	5	4.4	3	" technology for quicker medical data collection", "lessons learnt from the pandemics", . . .
Migration	14	4.43	3.64	"stop illegal immigration", "right for migration of non-EU people", "concrete and fair economic migration policy", . . .
European Democracy	8	3.25	3.75	"direct election for all EU institutions", "reform of European parties", "moving from unanimity to qualified majority", . . .
Digital Transformation	15	4.47	4	"investing in Open Source Software", "security of personal data", "an European Digital Academy", . . .
EU in the World	15	4.27	4.33	"Integration of EU armed forces", "EU to promote multilateralism", "updating EU neighbourhood policies", ...
Values, Rights and Rule of Law	15	4.53	4.4	"rights of LGBTQ+ people", "EU to protect Western Values", "protect citizens from crime", ...
Education, Youth and Sports	9	3.33	3.33	"teaching European History at school", "improvement of skills for EU youth", "Europe day as holiday in all EU", ...
Climate Change and Environment	11	5	5	"agro-ecology", "sustainable mobility", "eco-friendly manufacture", "stop subsidies for fossil fuels", . . .
Stronger Economy and Social Justice	9	3.67	3.67	"harmonise taxes in the EU", "a Universal Basic Income for all EU citizens", "reform pension shemes", . . .
Other proposals	15	4.67	4.2	"a common European language", " Member States should be less selfish", "an EU passport", . . .
Total	116	4.28	4.03	

Table 2: Quality evaluation of clusters for each CoFE category: number of clusters, Q1 and Q2 average scores and title examples to describe these clusters

a real-life Topic Mining in multi-lingual public consultation purpose, but did not allow inferring a quality indicator from cluster size distribution. This could be subject for further works.

6 Discussion

We didn't use kmeans for our system because a fixed set of cluster is a very important and risky choice when analysing a dataset containing an unknown number of topics. As such, we also wanted to give the analyst the possibility to zoom in and out in the clusters by changing the parameters of Louvain clustering, and being more or less conservative by adjusting the thresholds. While this possibility has ultimately not been implemented in the final system, these requirement guided us in the

"Accesibilitate la transport electric", "Zur Energiewende", "Coste real de las energías renovables y replanteamiento del sistema de fabricación de los conversores de energía", "Mehr Solarenergie !", "Veicināt ātru elektrovelosipēdu izmantošanu", "Mobilità elettrica", "Energia odnawialna", "Energías renovables planificadas", " Κινητικότητα για επαγγελματικούς σκοπούς*", "Umschichtung elektrischer Energie", "Transizione energetica", "Movilidad cotidiana", "Local returns from alternative energy installations", "Hasten the development of thermonuclear fusion", "Energiewende", "Incentivize the use of electric vehicles across the EU"*

Figure 2: Example of the top 15 sentences of a multilingual cluster of 45 sentences, ranked by network centrality.

fr: Apprentissage basique des langues de l'UE

en: The EU needs improved language learning

es: Facilitar el aprendizaje de idiomas

de: Die EU muss sich zum Ziel setzen, das Erlernen von Fremdsprachen radikal zu verbessern

Figure 3: Example of top sentences per language from one cluster: the closest sentence to the centroid for each of the languages of the cluster

design of the algorithm.

In Figure 2 we give an excerpt of a multilingual cluster whose topic revolves around green energy. We can see that the sentences are highly related to the broad topic, but also that it has different precise subtopics about mobility, transition and costs. The choice of the parameters has not only an impact over the number of clusters produced, but also the number of clustered sentences. Depending on the application one may want to have absolutely all the sentences clusters, or on the opposite to only extract topics out of these sentences. This was our case, and there is no optimal set of parameters as these depends on the subjective aim pursued by the analyst: only a small number of high level topic, or a collection of small but very precise topics?

While we could have produced clustering with even higher quality score by filtering even more the number of sentence making it to the final clusters. Nevertheless, in our context the only sentences that matters at the end are the one representing the cluster heads. Consequently, it can not have significant impact over the topic detection to filter the cluster from their most fringe elements. Moreover, an analyst has always the possibility to navigate through all the related sentences of the cluster.

When assessing the quality of semantic similarities computed with the LASER embeddings, a conservative threshold would have been 0.9 or even better 0.95. Such conservative thresholds would however have had high precision but low recall, and as such would have left out similar but still relevant sentences. Our algorithm use a threshold of 0.8 to build the graph, but it is able compensate for this lack in precision by using the filtering step. A such, it gives the possibility to sentences to be part of a cluster, by being close to its center, without having to be almost perfectly similar to any particular point from that cluster. Symmetrically, if a sentence is both far from both all data points and cluster center, it will rightfully be left out. In this way our algorithm is able to tackle with the lower precision threshold of 0.8 used to build the graph.

In Figure 3 we give the example of the "top sentences" which are the sentences of each language of the cluster that are closest to the centroid of the cluster. Such sentences are the most refined form of information our system is able to produce in order to describe clusters.

It has to be noted that our algorithm has the capacity to perform hierarchical clustering, by relaunching community detection on the clusters produced (with or without filtering). While this approach is promising, we did not explore it further due to lack of time for a thorough evaluation. We give an example of such a recursive decomposition by our algorithm in Figure 4. Such a feature is also interesting for analysts in order to better control the level of granularity of the information they extract.

6.1 Performances

The computation of embeddings is faster than with transformers models as LASER embeddings are BiLSTM. As such embeddings computation is not a bottleneck, and it is possible to compute them efficiently with respect to the other steps of the algorithm. In the experiments reported in this paper the machine used had 32 CPU and one RTX 8000 GPU. Computing the embeddings took less than a minute with an unoptimised code.

The Louvain clustering algorithm has taken only a few seconds to compute. The performance of how approach is dominated by the need for pairwise

- topic "developing solar panels":
- sub-topic 1 "sonar panels in the desert":
  - *Solar Panels in the Sahara Desert about Environmentally Friendly Energy for the World*
  - *Panouri solare in desertul Sahara*
- sub-topic 2 "sonar panels in the EU":
  - *Développer la production des panneaux solaires dans l'UE*
  - *Massiver Ausbau Solaranlagen in südlichen Mitgliedstaaten*

Figure 4: example of recursive decomposition of a cluster into relevant subtopics. In this case the topic names are given a posteriori

comparison between all sentences, which results in an complexity of O(n^2) for the algorithm. In order to mitigate this effect we used two strategies: parallelisation and incremental computation. Parallelisation had a very important impact over performance, bringing the running time to only 10 minutes using all the CPUs for the reported experiments. In order to tackle with the large quantity of data in the consultation, about a million sentences, we designed the graph computation to be stateful. As such in the production system deployed, it was incremental updated with daily new batches of documents, leveraging the fact that new data came regularly but in relatively small quantity, and handled the large quantity of data without problem.

7 Conclusion

We presented a novel algorithm for text clustering and topic detection in large sets of multilingual sentences. Our approach relies on two key aspects: the use of aligned multilingual embedding in order to compute a similarity graph between sentences written in different languages, on top of which community detection is applied, as well as a post-processing step in order to refine the cluster. The algorithm was developed for the need of a large public consultations with tens of thousands of sentences in 24 languages. The aim was to allow analysts to easily extract the underlying topics in this massive dataset. We provide a numerical evaluation that shows that our algorithm performs overall the best with respect to baselines. We also described a new methodology to empirically evaluate cluster quality and applied it to the output of our system on a random sample of sentences. The results shows that the algorithm is able to correctly group the topics together and that cluster centroids provide good names to the automatically extracted clusters.

Future works will involve measuring the quality of clustering for different parameters and study the stability of centroids for different set of parameters. The use of different aligned multilingual embeddings could be compared. We also want to perform additional manual error analysis of the not clustered sentences in order to better quantify the precision and recall of the algorithm. It would be also interesting to investigate the use of approximate distance comparison, possibly using approximate vector databases, in order to avoid pair-wise comparison between all sentences and, as such, reduce the algorithmic complexity of our approach.

References

Nejat Arinik, Vincent Labatut, and Rosa Figueiredo. 2021. Characterizing and comparing external measures for the assessment of cluster analysis and community detection. *IEEE Access*, 9:20255–20276.

Henrique F de Arruda, Luciano da F Costa, and Diego R Amancio. 2016. Topic segmentation via community detection in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):063120.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.

Tanmoy Chakraborty, Ayushi Dalmia, Animesh Mukherjee, and Niloy Ganguly. 2017. Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, 50(4):1–37.

Jessica Elan Chung and Eni Mustafaraj. 2011. Can collective sentiment expressed on twitter predict political elections? In *AAAI*, volume 11, page 1770–1771.

Gianni Costa and Riccardo Ortale. 2021. Jointly modeling and simultaneously discovering topics and clusters in text corpora using word vectors. *Information Sciences*, 563:226–240.

Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. 2009. "omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, page 73–80. ACM.

F. Fu, L. Liu, and L. Wang. 2008. Empirical analysis of online social networks in the age of web 2.0. *Physica A: Statistical Mechanics and its Applications*, 387(2–3):675–684.

Martin Gerlach, Tiago P Peixoto, and Eduardo G Altmann. 2018. A network approach to topic models. *Science advances*, 4(7):eaaq1360.

Andreas Hamm, Jana Thelen, Rasmus Beckmann, and Simon Odrowski. 2021. Tecominer: Topic discovery through term community detection. *arXiv preprint arXiv:2103.12882*.

David Jurgens. 2011. Word sense induction by community detection. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 24–28.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, page 415–463. Springer US.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

M.E. Pierce, G.C. Fox, J.Y. Choi, Z. Guo, X. Gao, and Y. Ma. 2009. Using web 2.0 for scientific applications and scientific communities. *Concurrency Computation Practice and Experience*, 21(5):583–603.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Kartik Sawhney, Marcella Cindy Prasetio, and Suvadip Paul. 2017. Community detection using graph structure and semantic understanding of text. *SNAP Stanford University*.