Community based Discussion Detection on Code Switched Social Media Content

Prarthana Padia[1], Lucia Falzon[2], Aaron Harwood[1], Shanika Karunasekera[1] and Michelle Vanni[3]

[1] The University of Melbourne, Melbourne, Australia
[2] Defence Science and Technology Organisation, Adelaide, Australia
[3] CCDC Army Research Laboratory, Adelphi, MD USA
prarthana.padia@gmail.com, Lucia.Falzon@dst.defence.gov.au,
{aharwood,karus}@unimelb.edu.au, michelle.t.vanni.civ@mail.mil

Abstract. Social media (SM) facilitates large numbers of discussions involving communities across the globe. Multilingual individuals often communicate using alternate languages, in a behavior known as code-switching (CW). Characterizing SM discussions involving CW uncovers factors affecting cultural diversity and provides insight into the evolution of trends and opinions. Currently, enormous spans of SM usage—valuable data for linguistic behavior analysis—challenge such discussion-tracking efforts. Applicability of state-of-the-art community detection algorithms—based on network structure of followers and friends, interactions of retweets and mentions, and hashtags occurring in discussion contexts—has been limited. Resulting communities are heavily dependent on attribute types used in the detection and, linguistic characteristics are ignored in the analysis. We propose a novel framework for discussion analysis, which facilitates understanding and processing of the high volume of dynamic SM information currently available. Our research focuses on 1) detecting community-based multilingual discussions in SM; 2) defining evaluation metrics and heuristics to obtain CW discussions from the SM posts; 3) developing language identification algorithms to detect the language(s) of discussion texts; 4) visualizing the user-discussion graph with language clues, where each user component connects to its discussion components extracted through a defined ranking; 5) representing individual public discussions in the form of a tree, where each node corresponds to a microblog post (tweet) and grows with responses.
Keywords: social media, community detection, language identification, code-switching

1 Introduction

Social media (SM) facilitates a large number of discussions and debates that involve the participation of different communities from across the globe. Anybody can contribute to a discussion by responding to a piece of content that they come across in any depth. Individuals often utilize alternate languages in a single conversation, known as code switching (CW), for effective communication in a discussion. Characterizing these discussions helps to analyze the contextual factors directing a community's cultural diversity. Diversity in communities leads to multilingual discussions, which provide valuable insights into the evolution of public trends and opinions in communities.

The current span of enormous SM usage makes tracking and understanding these multilingual discussions a challenging task for linguistic behavior analysis. Recently, significant progress has been made by developing community detection algorithms based on the network structure of followers and friends, interactions of retweets and mentions, and hashtags occurring in the discussion context. However, the general applicability of these state-of-the-art approaches has been limited, because the resulting communities are heavily dependent on the type of attributes used in the detection and the linguistic characteristics of the discussion are ignored in the analysis.

We are developing a new framework of discussion analysis that aims to facilitate the understanding and processing of current high volumes of dynamic SM information. The research aims to perform the following tasks: 1) detecting community based multilingual discussions of SM involving multilingual users; 2) defining evaluation metrics and heuristics to obtain CW discussions from the SM posts; 3) developing language identification algorithms to detect the language(s) of discussion texts; 4) visualizing the user-discussion graph with language clues where each user component connects to its discussion components extracted through defined ranking; 5) representing individual public discussions in form of a tree, where each component corresponds to a SM tweet and grows with responses.

2 Background

Considering social studies as the earliest context, community detection was based on the concept of groups of people sharing interests or ideas. Recently, due to work from a variety of related perspectives, there is no one unique definition of community. A community can be formed on the basis of the network structure or the domain under study. With the increase in online modes of communication, the notion of SM communities came into existence. In the context of SM, a community can be defined as "a network subgraph comprising a set of [SM] entities associated with common elements of interest" [8]. A common element might be a topic, place, event, activity or cause. [9] uses network edge content to detect SM communities. It considers edge content as a source of information for characterizing the nature of interactions between participants effectively, making community detection more efficient. [8] further describes SM communities as explicit or implicit. Explicit SM communities are formed based on human decisions and acquire members on consent. Facebook and Twitter can be the examples of such communities. Our research targets dynamically formed Twitter-based communities for discussion detection and analysis.

The abundance and variety of online communities has led to multilingual discussions that provide interesting insights on the evolution of public trends in communities. Multilingual discussions involve the use of multiple languages or language alternation, also known as code-switched content. CW is often practiced by multilingual speakers and, in the era of online communication, has become a challenge for the language identification task. CW has been a topic of formal research since the 1970s [5]. There are several research works focusing on CW and its impact, dating back to late 1900s [1, 2, 10]. Before SM or access to sophisticated or computationally feasible algorithms, CW analysis was carried out by observing human conversations and hypothesizing linguistic characteristics [1, 2]. The aim was to observe and analyze the impacts of CW on the community's linguistic adaptation and behavior.

Recently, significant progress has been made in the field of language detection algorithms for SM content based on various machine learning techniques ranging from a simple approach based on frequencies of character n-grams to more complicated approaches using word embedding, extended Markov Models and Conditional Random Field (CRF) auto-encoders [3, 4, 6, 7, 11]. Several researches aim at modelling and proposing solutions for the same [3, 4, 6, 7, 11]. [4] used supervised classification and sequence labelling to device an automatic language identification mechanism for the code-switched languages of SM. On the other hand, [6] address the problem of Named Entity Recognition (NER) in code-switched tweets using simple features and Conditional Random Fields (CRF) classifier. To dive in detailed language modelling, [3, 7] proposes different approaches for word-level language identification tasks in the context of code-switched SM text. [3] employs DNN architecture for training models for word-level language identification and [7] trains a character n-gram based CRF model which is investigated using interesting CW metrics like Multilingual index, Integration index and Code-mixing index. A bigger picture of the first shared language identification task for code-switched data was provided by [11]. It dealt with analyzing the performance of various techniques adopted by the participating teams to accomplish the task. The evaluation showed that language identification at the token level became more difficult when the languages present were closely related.

The current state-of-the-art approaches aim to classify/identify languages of code-switched data based on datasets consisting of a limited number of languages. However, the scalability of these approaches to real time data volumes and variety has been limited. Most are applied to annotated or filtered datasets as it is feasible to apply complex algorithms to definite pre-processed amounts of language data. This leaves a good scope of developing a furthermore generalized and scalable approach for code-switched data identification in a realistic SM context.

3 Experiment Methodology

3.1 Language Identification using NLP

Dataset. Social media data has been collected through RAPID (Real-time Analytics Platform for Interactive Datamining, a real-time topic tracking and analytics platform developed at the University of Melbourne) with aggregated tweets focused on the 2018 Quebec Election, among others, with datasets stored in JSON files, which facilitates the training of our algorithm. Reported results are on Quebec Election data. The framework is extendable to process dynamic real time tweets to analyze latest discussions.

Method. The technique consists of filtration, vector encoding, and tie-breaking.
*Filtratio*n**.** Text words are matched against the words in all (56 aspell) dictionaries, filtering out the non-pruned words.
Vector encoding. The order-wise occurrences of pruned words in each dictionary are marked as 1, with the rest marked as 0.
Tie-breaking mechanism. This step relies on maximum run length, or the maximum length of consecutive occurrences of words of the text in the dictionary along with the percentage of language, that is, the proportion of words occurring in a single dictionary to the total number of words in the text**.**

Training. The system is trained on tweets that have all the pruned words belonging to the same language dictionary and the number of pruned words in the text is at least three. The training stores in a 'mongo' database the occurrences in the trainable tweets of pairs of consecutive words (bi-grams), which are utilized while labeling. The count of language bigrams in the database used for training is shown in Figure 1.

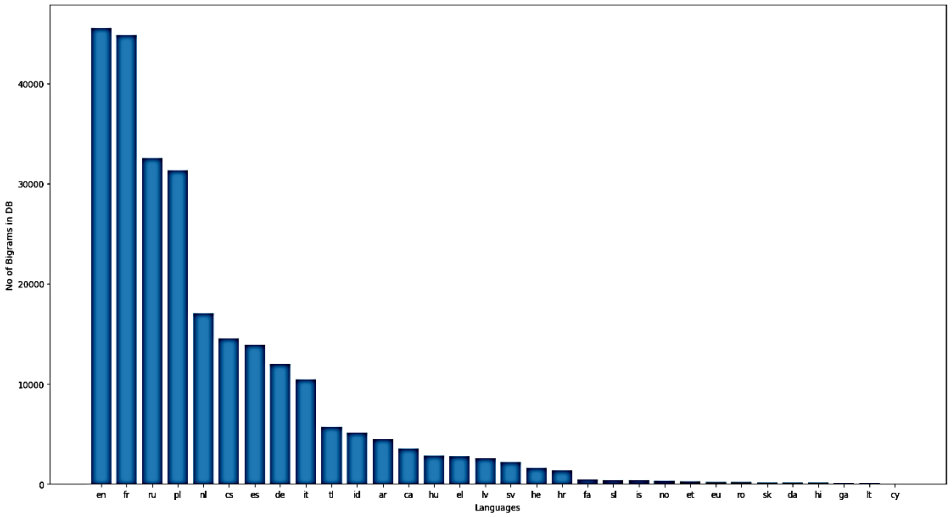

Fig. 1: Count of language bigrams in the database

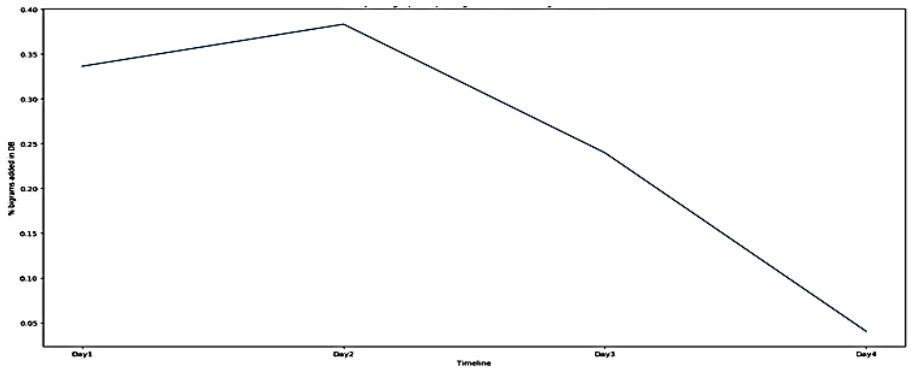

Fig. 2: Percentage addition of bigrams in database with respect to time.

The temporal graph in Figure 2 shows percentage addition of bigrams to the database. It depicts the progress and peak stage of inclusion of all possible bigrams from the training data. Note that, as time progresses, there is no significant increase in bigrams, despite further training.

Labeling. Labeling associates languages with tweets. Languages in the text are identified using the three techniques described here, maximum run length of the words in a dictionary, percentage of a language in the text and bi-gram score of the words. The bigram score is obtained from the trained mongo database bigram set.

The algorithm also applies a tag indicating mixed/multi/mono-language(s) with the 'mixed' tag specifying use of one major language and 1-2 words from a different language, while the 'multi' tag denotes use of more than one language in a text.

3.2 Scoring Users and Public Discussions

Definite scoring metrics rank users and corresponding discussions as indicated here.

User score metric. Users are ranked on maximum code switching/language alternation's harmonic mean scores over multiple tweets in a discussion.
Score function.
For a user *u*, let X be the set of languages used by *u*.
For a tweet *v* of user *u*, let *Lang(v)* be the language(s) of the tweet.
Let,

$$Lang(V) = \bigcup (Lang(v)) \quad \forall\, v \in V \tag{1}$$

$$count(V, x) = |\{v | x \in Lang(v)\}| \tag{2}$$

$$\frac{1}{HarmonicMean(V)} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{count(V, x_i)}, \quad \forall\, x \in Lang(V) \tag{3}$$

The score for a user *u* is given as follows:

$$Score(u) = HarmonicMean(V) \tag{4}$$

where *V* is a set of all tweets of user *u*.

User scores are sorted to find users with maximum harmonic means score.

Discussion score metric. Discussions are ranked on maximum language users within a discussion. Briefly, discussions are sorted based on count of users using more than one language over multiple tweets within a discussion.

Score function.
Let *Y* be the set of all users in a discussion *d*.

$$Y = \bigcup (y) \quad \forall\, y \in Y \tag{5}$$

Let *X(y)* be the set of languages used by user *y* in *d*, such that $y \in Y$
Let,

$$count(Y, x) = |\{y | x > 1, x \in X(y)\}| \tag{6}$$

The score for a discussion d is given as follows:

$$Score(d) = max(count(Y)) \tag{7}$$

Discussions containing at least 2 languages are considered for further analysis. Note as well that the top 20 users with common discussions (discussion with participation of more than 1 top user) plus the top 10 (max 10) user-specific discussions, in which the user has used more than 1 language, have been considered for further analysis.

3.3 Visualizing Representation

Subsequent to labeling, for better visual clues, algorithm output is used to represent the relationship between the extracted users and discussions in graph structure. The following are the graph structure specifications useful for understanding the result files.

User-Discussion Graph Structure. There are two color-coded components.
Graph components. The user-discussion graph structure includes top user components (sorted by maximum CW score) with directed edges towards their corresponding public Twitter discussions (sorted by maximum language users score). The user component is represented as a square whereas the discussion component has a circular shape.
Color code for graph components. Several color combinations are used:
1. In user component: Proportion of language(s) used overall, i.e., in all user's tweets.
2. In discussion component: Proportion of language(s) used by all users in discussion.
Each user component includes the user's 'screen name' label besides the node. Whereas the discussion components include the globally generated discussion ID as label besides the node. A language legend is plotted for ease of analysis.

Public Discussion Graph Structure. There are two color-coded components.
Graph components. The public discussion graph illustrates a twitter discussion scored by maximum language users within a discussion. Each node corresponds to a tweet. The arrow of the directed edge points towards the tweet being replied to by a tweet on the other end. The length of the edge is proportional to the relative time taken to reply.
Color code for graph components. Several color combinations are used to depict participating users, classified tweet languages and extent of classified language clarity. Each node is divided into two colors as described here.
1. The upper half color corresponds to the dominant language in the tweet.
2. The lower half of the node represents a dominant user who occurred more frequently than others. Each top user gets a distinct color. Legends of both users and languages are displayed alongside the figure.

Lastly, the black circle around the node stands for 'unclear' language classification of the tweet, e.g., when the tweet contains all non-pruned words, such as emoticons.

4 Results and Discussion

4.1 Language Percentages

Results were generated on 2017 Quebec Elections data with the framework described. While Table 1 describes the percentages of the individually classified languages, it should also be noted that 3% of tweets were classed as multilingual or mixed.

Language	Abbreviation	%	Language	Abbreviation	%
French	fr	82.017	Indonesian	id	0.006
English	en	13.183	Breton	br	0.005
Spanish	es	2.303	Croatian	hr	0.005
Catalan	ca	2.133	Hungarian	hu	0.004
Norwegian	no	0.077	Estonian	et	0.004
German	de	0.044	Esperanto	eo	0.003
Italian	it	0.041	Welsh	cy	0.003
Lithuanian	lt	0.023	Russian	ru	0.003
Dutch	nl	0.018	Swedish	sv	0.002
Romanian	ro	0.015	Czech	cs	0.002
Arabic	ar	0.015	Icelandic	is	0.002
Afrikaans	af	0.014	Irish	ga	0.002
Kazakh	kk	0.014	Latvian	lv	0.001
Basque	eu	0.012	Greek	el	0.001
Polish	pl	0.01	Slovak	sk	0.001
Danish	da	0.007	Tagalog	tl	0.001

Table 1. Percent languages in Quebec Elections dataset labeled by algorithm.

4.2 Tweets in Disagreement

Languages classified for 4.58% tweets are in disagreement with the language tag provided by Twitter. As Twitter does not provide multilingual tags, all the tweets classified as mixed/multilingual by our algorithm as well disagree with the Twitter language labels. Table 2 gives examples of tweets in disagreement.

Tweet text	Classified language(s)	Twitter language
"RT @mtlgazette: St. Petersburg Philharmonic appoints Charles Dutoit as guest conductor https://t.co/ojegSycJ9d https://t.co/h4R4hpUvcq"	'en'	'fr'
"RT @metromontreal: Le B'nai Brith s'invite dans la campagne du PQ #Quebec2018 https://t.co /AHevE7RYRi https://t.co/rKcCDiSLv3"	'fr' (mixed)	'fr'
"Canadiens Notebook: Nick Suzuki returned to his junior team https://t.co/68lbVNdwE2 https://t.co/PzB02wUBtc"	'en' (mixed)	'en'
"RT @marchetucaq: Appel au vote #CAQ de @liseravary aux anglophones du QuÃ©bec : "Anglos shouldn't spurns alternatives to the Liberals". #Nowâ"	'fr' and 'en'	'fr'

Table 2. Tweets language labels (generated by the labeling algorithm on Quebec Elections 2017 dataset) in disagreement with Twitter provided language labels.

4.3 User Discussion Graph

Figure 3 is a user-discussion graph representing the relationships between the top 20 users and their corresponding (common and non-common) scored discussions. Each node is filled with

colors corresponding to the languages on the legend. In this and the following section, graphs follow the structure described in Section 3.3.

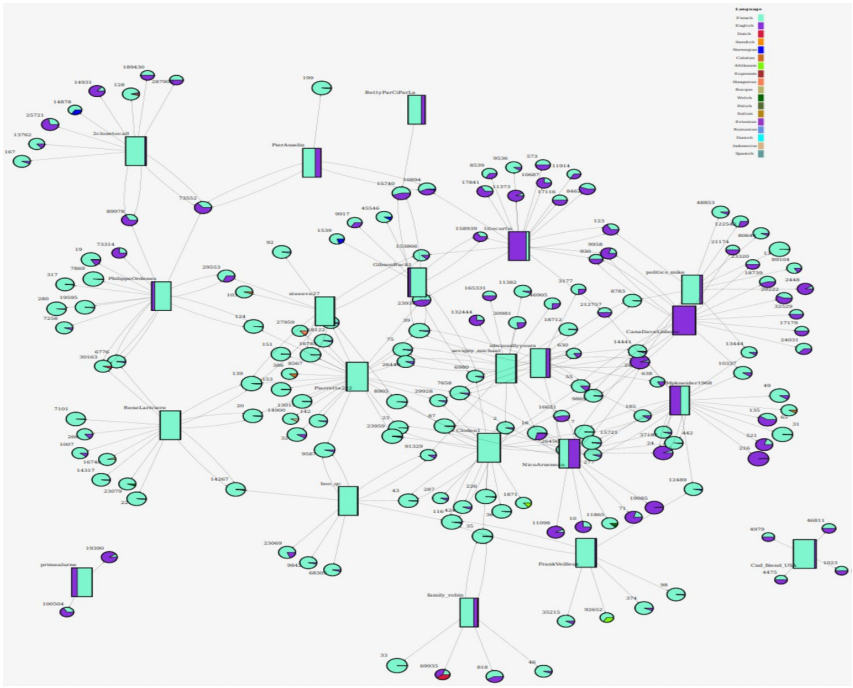

Fig. 3. Graph of top user components connecting to corresponding discussions.

Note that most of the data is in French or English. Although not all discussions are displayed, it can be observed that the language dominance bears a fair resemblance to that of the language percentages in Table 1. Moreover, the scoring functions for both users and discussions tend toward significant and proportional language alternation, a feature which assists in discovery of multilingual user communities and analysis of detected discussions. Such discovery facilitates the uncovering of factors driving cultural diversity, trends, and public opinion. For example, the impact of a user‘s linguistic behavior on other users’ responses can be evaluated by checking the influence of an individual user‘s CW on the language of responses in the same discussion.

4.4 Public Discussion Graph

Here we present a public discussion graph snippet with associated tweets. Figure 6 is an example of a public discussion graph including participation from 3 top users. Tweets associated with the discussion are described in detail in Table 3.

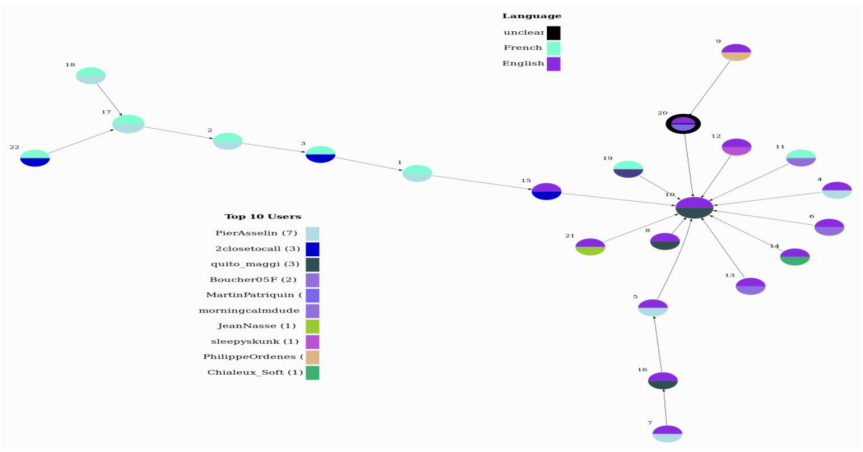

Fig. 4. Public discussion graph example with participation from 3 top users.

Public discussion graphs can be useful to infer user details, such as linguistic behavior from the reply structure. Discussion analysis can lead to discovery of interesting factors driving community trends or ideas based on response characteristics of the participating users. The expansion of the graph indicates increased user participation, suggesting the possibility of a new thread branching off from an existing one to become the content of further discussion. In case of such deviations, analysis of raw tweet content is advantageous.

ID	Tweet text	Classified lang	Twitter lang
1	"@2closetocall @quito_maggi @partiquebecois @QuebecSolidaire @coalition-avenir Est-ce que les autres maisons perçoivent le même signal?"	'fr'	'fr'
2	"@2closetocall @quit_maggi @partiquebecois @QuebecSolidaire @coalition-avenir Une baisse de 8 points entre deux coups de sonde m'apparaît excessive, pour ne pas dire plus. J'ai du mal à croire que l'opinion fluctue aussi rapidement dans de telles proportions. CBC poll tracker du 21 sept. https://t.co/6Yti72Q1Qc https://t.co/RHNFlwdLvx"	'fr'	'fr'
3	"@PierAsselin @quito_maggi @partiquebecois @QuebecSolidaire @coalition-avenir On n'a pas assez de sondages de leur part. On verra"	'fr'	'fr
4	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir 8 points in one poll sounds a bit too much. I have my doubts."	'en'	'en'
5	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir That number got my attention yesterday, as you know. We published this from Steve Pinkus in Le Soleil: "Le Parti québécois a chuté de plus de 3% depuis le débat..." So I'm trying to understand where the 8% you quoted comes from?"	'en'	'en'
6	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir too close to call yesterday pq as 21.4% wrong for pq loose 8 pts . qs already in 4 th place ."	'en'	'en'
7	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir Thanks. This should be an interesting week."	'fr'	'fr'
8	"We also had a new high in the change vote, almost 76% of Quebec voters want a change of Government approaching the final week #quebec2018 #qcpoli"	'en'	'en'
9	"@MartinPatriquin @quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir Libs were done since the beginning with more than 70% of Quebecers hoping for a change. What is new however, is the split between PQ voters between CAQ and QS. It may be the end of an historical cycle."	'en'	'en'
10	"Last night we saw the biggest shift in #quebec2018 vote intentions. .@partiquebecois dropped 8points, .@QuebecSolidaire passed in to 3rd place & the .@coalitionavenir was first. If this trend holds over the weekend, by Monday #PQ will be at a near wipe out level Stay tuned"	'en'	'en'
11	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir je crois pas ça perte de 8 pts too close too call hier le pq à 21.4% 21 sept 2018 .qs est toujours 4eme ."	'fr'	'fr'
12	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir It's Calgary all over again."	'en'	'en'
13	"@quito_maggi @Anine_79 @partiquebecois @QuebecSolidaire @coalition-avenir This shift is suspect at best. Voters simply do not shift this fast. I will wait for more in depth surveys and polls later in the week to identify the real trends"	'en'	'en'
S 14	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir We'll see"	'en'	'en'
15	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir So my prediction that a debate victory could push QS to 20% wasn't crazy"	'en'	'en'
16	"@PierAsselin @partiquebecois @QuebecSolidaire @coalitionavenir 25% on Thursday night, 17% on Friday night, of course, the 3 days rolling poll doesn't reflect that"	'en'	'en'
17	"@2closetocall @quito_maggi @partiquebecois @QuebecSolidaire @coalition-avenir Par curiosité, toujours avec Poll Tracker, voici dans le carré bleu du haut la courbe de la CAQ depuis son sommet, et dans celui du bas j'ai dessiné une flèche qui représente une baisse de 8 pts pour le PQ. J'ai hâte de voir les prochaines données. https://t.co/EfqAbv60fh"	'fr'	'fr'
18	"@2closetocall @quito_maggi @partiquebecois @QuebecSolidaire @coalition-avenir C'est d'autant plus déroutant que la récente tendance pour le PQ était une hausse progressive."	'fr'	'fr'
19	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir Ah bon. #qc2018 #polqc" "	'fr'	'fr'
20	"@quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir Dueling vote splits. How much will QS eat into PQ? How much will PQ eat into CAQ? If answers are "more than in 2014" and "less than 2014" respectively then the Libs are indeed in trouble. #polqc"	'en'	'en'
21	"@quito_maggi @2closetocall @partiquebecois @QuebecSolidaire @coalition-avenir Sounds like a whole lot of manipulative hogwash. I'd like to see/hear one of your polls. #massManipulation #PROPAGANDA #QC2018 #POLQC"	'en'	'en'
22	"@PierAsselin @quito_maggi @partiquebecois @QuebecSolidaire @coalition-avenir Le poll tracker est inutile cette élection car il n'inclut quasiment aucun sondage Mainstreet"	'fr'	'fr'

Table 3: **Tweet components' table corresponding to Figure 4 public discussion graph.**

References

1. Peter Auer. A conversation analytic approach to code-switching and transfer. Codeswitching: Anthropological and sociolinguistic perspectives, 48:187–213, 1988.
2. Peter Auer. Code-switching in conversation: Language, interaction and identity. Routledge, 2013.
3. Ashutosh Baheti, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. Curriculum design for codeswitching: Experiments with language identification and language modeling with deep neural networks. Proceedings of ICON, pages 65–74, 2017.
4. Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. Code mixing: A challenge for language identification in the language of social media. In Proceedings of the first workshop on computational approaches to code switching, pages 13–23, 2014.
5. David Crystal. The Cambridge encyclopedia of language, vol. 1. Cambridge Univ. Press Cambridge.
6. Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. Simple features for strong performance on named entity recognition in code-switched twitter data. In Proc. of Third Workshop on Computational Approaches to Linguistic Code-Switching, pp.103–109, 2018.
7. Deepthi Mave, Suraj Maharjan, and Thamar Solorio. Language identification and analysis of codeswitched social media text. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 51–61, 2018.
8. Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. Community detection in social media. Data Mining and Knowledge Discovery, 24(3):515–554, 2012.
9. Guo-Jun Qi, Charu C Aggarwal, and Thomas Huang. Community detection with edge content in social media networks. In Data Engineering (ICDE), 2012 IEEE 28th International Conference on, pages 534–545. IEEE, 2012.
10. Richard Skiba. Code switching as a countenance of language interference. The internet TESL journal, 3(10):1–6, 1997.
11. Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 62–72, 2014.