ARL-TR-9156 ● Mar 2021

DEVCOM ARMY RESEARCH LABORATORY

A Community-Based Code-Switch Discussion Pipeline

by Prarthana Padia, Michelle Vanni, Lucia Falzon,
Aaron Harwood, Shanika Karunasekera, and Sue Kase

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9156 ● Mar 2021

A Community-Based Code-Switch Discussion Pipeline

Prarthana Padia, Aaron Harwood, and Shanika Karunasekera
University of Melbourne

Michelle Vanni and Sue Kase
Computational and Information Sciences Directorate,
DEVCOM Army Research Laboratory

Lucia Falzon
Australia Defence Science & Technology Group

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
March 2021	Technical Report	15 March 2018–2 November 2020

4. TITLE AND SUBTITLE
A Community-Based Code-Switch Discussion Pipeline

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Prarthana Padia, Michelle Vanni, Lucia Falzon, Aaron Harwood, Shanika Karunasekera, and Sue Kase

5d. PROJECT NUMBER
W911NF-18-2-0050

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLC-NC
Adelphi, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9156

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: Prarthana Padia: 0000-0002-6649-0968; Lucia Falzon: 0000-0003-3143-4351; Sue Kase 0000-0003-2732-629X;

14. ABSTRACT
Social media (SM) facilitates discussions within communities across the globe, and to communicate effectively, multilinguals will often alternate languages in a phenomenon known as code-switching (CW). Discussions in which CW is exhibited can, upon analysis, reveal a community's diversity and provide insight into evolving trends and opinions. Widespread use of SM allows for tracking and characterizing these discussions for cultural and linguistic analysis. Advanced algorithms for community detection, based on network structures of followers and friends, interactions of retweets and mentions, and patterns of hashtag occurrence largely ignore linguistic cues in the body of posts. For this reason, the applicability of these state-of-the-art approaches to problems involving CW analysis has been limited, as the resulting communities are dependent on attribute types used in the detection rather than on attributes characterizing the significance of the CW; that is, the connections among posters, the topics under discussion, and the social context in which it occurs. Here we develop a new framework to facilitate understanding and CW processing of high volumes of SM information by 1) detecting community-based multilingual SM discussions, 2) defining evaluation metrics and heuristics to obtain CW discussions, 3) developing word-level language ID algorithms, 4) visualizing user-discussion graphs where component types are extracted based on defined rankings, and 5) representing discussions as trees with first-order nodes as posts, and nonterminal and leaf nodes as responses.

15. SUBJECT TERMS
community detection, language identification, code-switching, CW, social media, SM

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sue Kase
Unclassified	Unclassified	Unclassified	UU	25	19b. TELEPHONE NUMBER (Include area code) (301) 802-3976

Contents

List of Figures

List of Tables

Acknowledgments

Many thanks are owed to Drs Elizabeth Bowman and Stephen LaRocca for their support at various stages of this effort. Thanks also to Sue Kase for ensuring this report was published.

1. Introduction

Social media (SM) facilitates large numbers of discussions involving multilingual communities across the globe. These individuals often alternate languages in a behavior known as code-switching (CW). Since CW is a community-specific phenomenon, determining its social meaning provides insight into the evolution of trends and opinions. The enormous spans of SM data currently available are valuable for linguistic analysis, but for two reasons it is a challenge to track CW SM discussions. First, state-of-the-art community-detection algorithms are limited in their applicability to this problem. These techniques are based on network structures of followers and friends, interactions of retweets and mentions, and hashtags occurring in discussion contexts. They thus result in communities heavily dependent on attribute types used in the detection, with the body of the postings largely ignored. Conversely, computational linguists analyze CW in SM outside of its networked context, training on boutique datasets and producing techniques with an unclear level of scalability.

Here we propose a novel framework for multilingual discussion analysis that mitigates these issues and facilitates understanding and processing of high volumes of linguistically dynamic SM information in five steps: 1) detecting community-based multilingual SM discussions, 2) defining the evaluation metrics and heuristics that identify SM discussions with CW, 3) developing language identification (LID) algorithms to detect the language(s) of texts, 4) visualizing a user-discussion graph with language clues, where components are extracted through defined rankings, and 5) representing public discussions as trees with nodes corresponding to microblog posts (tweets) and growing with responses.

Herein, the SM data under consideration was mined through the Real-time Analytics Platform for Interactive Data-mining (RAPID), a real-time topic-tracking and analytics platform developed at the University of Melbourne, resulting in a corpus of aggregated tweets focused on the 2018 Quebec election. Our innovation resulted in a language distribution mirroring that of a major SM platform, and it uncovered at least one network configuration with which CW could be associated in this dataset.

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, as a result of work being done 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. Online modes of communication brought the notion of SM communities 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".[1] A common element might be a topic, place, event, activity, or cause. Qi et al.[2] use network edge content to detect SM communities: Edge content is considered a source of information for characterizing the nature of interactions between participants effectively, making community detection more efficient. Papadopoulos et al.[1] further specify SM communities as explicit or implicit. Explicit SM communities are formed based on human decisions and acquire members on consent, as with Facebook and Twitter. Considerations of this nature are beyond the scope of our research, which targets implicit SM communities dynamically formed around cultural topics for discussion detection and analysis.

The abundance and variety of such online communities has led to multilingual discussions, which provide insights into the evolution of trends in communities. Multilingual discussions involve the use of multiple languages or language alternation, also known as CW. In the era of online communication, CW has become a challenge for the automatic LID task and has been a topic of formal research since the 1970s.[3] Several research works focusing on CW and its impact date back to the late 1900s.[4–6] Before SM or access to sophisticated or computationally feasible algorithms, CW analysis was carried out by observing human conversations and hypothesizing linguistic characteristics.[4,5] The aim was to observe and analyze the impacts of CW on the community's linguistic adaptation behavior.

Recently, significant progress has been made in the field of LID algorithms for SM content based on various machine learning techniques ranging from a simple approach involving frequencies of character n-grams to more-complicated approaches using word embeddings, extended Markov models, and conditional random field (CRF) auto-encoders.[7–11] Barman et al.[8] used supervised classification and sequence labelling to devise an automatic LID mechanism for CW in SM. On the other hand, in Jain et al.[10] the problem of named entity recognition (NER) in code-switched tweets is addressed using simple features and a CRF classifier.

Diverse language-modeling approaches for word-level LID tasks in the context of CW in SM text are proposed in Baheti et al.[7] and Mave et al.,[11] a deep neural network architecture for training models in Baheti et al.,[7] and a character n-gram-based CRF model using CW metrics like multilingual index, integration index, and code mixing index in Mave et al.[11] A big picture of the first shared LID task for CW data was provided in Soloriol et al.,[9] which dealt with analyzing the performance of various techniques adopted by participating teams to accomplish

the task. The evaluation showed that LID 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 CW data based on datasets consisting of a limited number of languages. 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 preprocessed amounts of language data. This leaves the field open for developing novel generalized and scalable approaches to CW data identification in a realistic SM context.

3. Algorithm Development Methodology

Our five-step approach addresses the following research question: To what extent can current SM community detection technology, algorithms, and research address the task of automating the prediction of the social meanings of CW in multilingual SM community discussion contexts?

3.1 Language Identification Using Natural Language Processing

Dataset. Social media data was collected through the RAPID resulting in a corpus of aggregated tweets focused on the 2018 Quebec election. Data was stored as JavaScript Object Notation files, which facilitates the training of our algorithm, and results are reported here. The RAPID platform is extendable for dynamic processing and analysis of discussions.

Method. The LID technique consists of filtration, vector encoding, and tie-breaking.

Filtration: Text words are matched against the words in current standard language-specific word dictionaries,[*] filtering out the nonpruned words.

Vector encoding: The order-wise occurrences of pruned words (words that match in Aspell) in each dictionary are marked as "1", with the rest marked as "0".

Tie-breaking: We specify the following two mechanisms:

- *Maximum run length*: maximum length of consecutive occurrence of words of the text in a dictionary

[*] We used the 56 dictionaries in Aspell, a standard free software spell-checker for the GNU operating system, which also compiles for other Unix-like operating systems and Windows (https://ftp.gnu.org/gnu/aspell/dict/0index.html).

- *Percentage of language*: proportion of words occurring in a single dictionary to 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. Furthermore, the training stores the occurrences in the trainable tweets of pairs of consecutive words (bigrams), used while labeling, in a specialized database.* The count of language bigrams in the database used for training is shown in Fig. 1. It depicts the proportion of data that is taken up by a given language in the training set, according to our calculations.

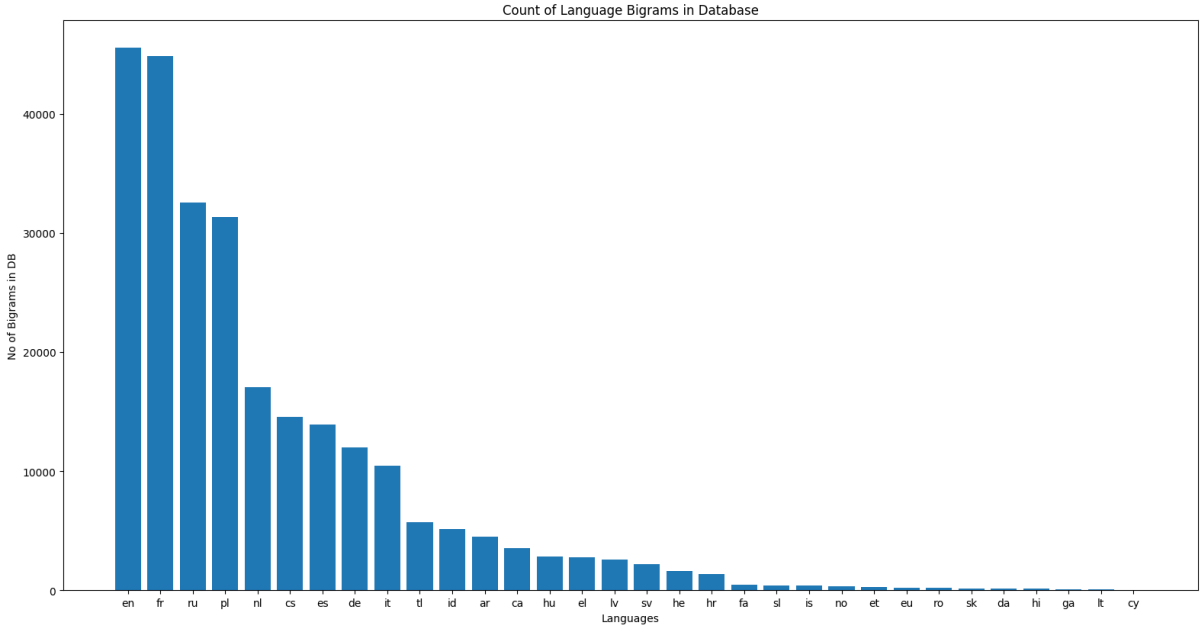

Fig. 1 Count of language bigrams in the database

Figure 2 is a temporal graph that shows the percentage addition of the bigrams in the database over time. It depicts the expansion of the training data with additional bigrams over time. It is evident that with time there is no significant increase in training with bigrams. This indicates the completion of the training set including all possible bigrams for the purpose of this experiment.

*MongoDB is a database with advanced design for the handling of large unstructured data, flexible querying, and results sampling (https://www.mongodb.com/).

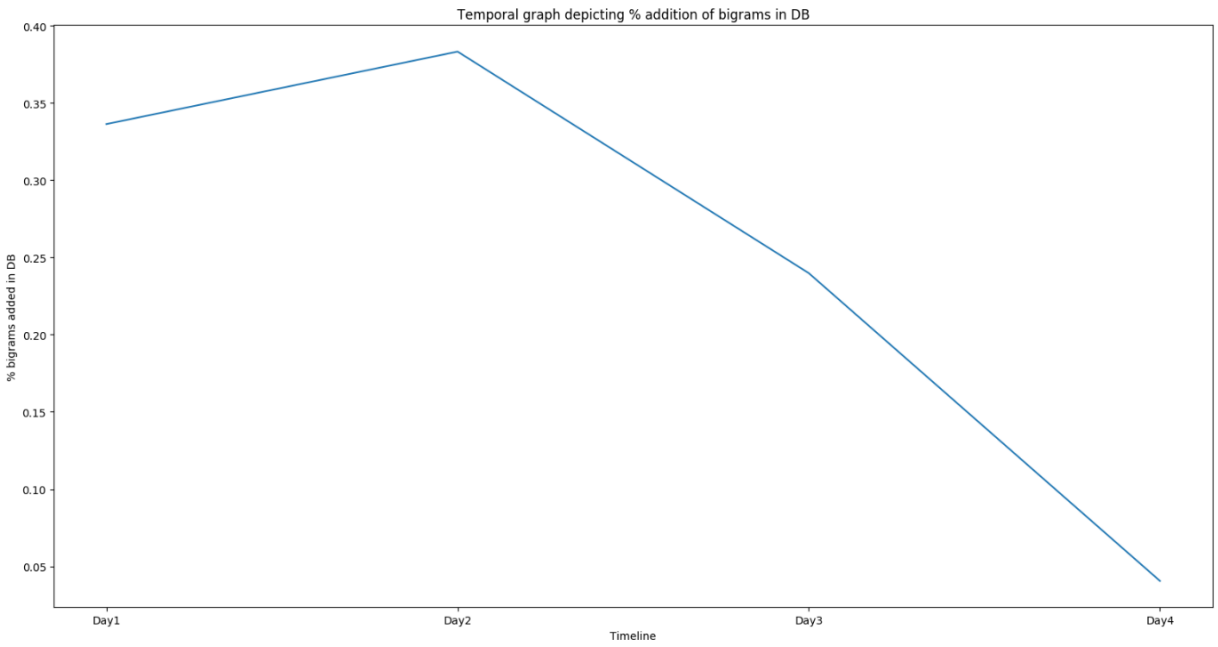

Fig. 2 Percentage addition of bigrams in database with respect to time

Labeling: Labeling associates languages with tweets. Languages in the text are identified using the three techniques: 1) Maximum Run Length of the words in a language-specific dictionary, 2) Percentage of Language in the text, and 3) Bigram Score of the words. The Bigram Score is obtained from the trained database of bigram set. Besides classifying the languages, the algorithm also associates a tag with the tweet to determine the use of mixed/multi-/mono-language(s) in the classification as described:

- "Mixed" tag stands for the use of one major language and one or two words from a different language in a text.
- "'Multi" tag stands for the use of more than one language in a text.

3.2 Scoring Users and Public Discussions

Definite scoring metrics are employed to rank users and their corresponding discussions. In the given context, a discussion is considered to be a nontransactional online exchange of information between or among two or more SM users on a topic of shared interest.

User Score metric. Users are ranked on maximum CW/language alteration'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.

$$Lang(V) = \cup \left(Lang(v)\right) \forall\, v \in V \tag{1}$$

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

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

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*.

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

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

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

The score for a discussion *d* is given as follows:

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

Discussions containing at least two languages are considered for analysis in the user discussion graphs in Figs. 3 and 4. Furthermore, the top 20 users with common discussions (discussion with participation of more than 1 top user) are represented in the common discussion graph in Fig. 4. The top 10 user-specific discussions, in which the user has used more than one language, are the source for the discussion-tree representations.

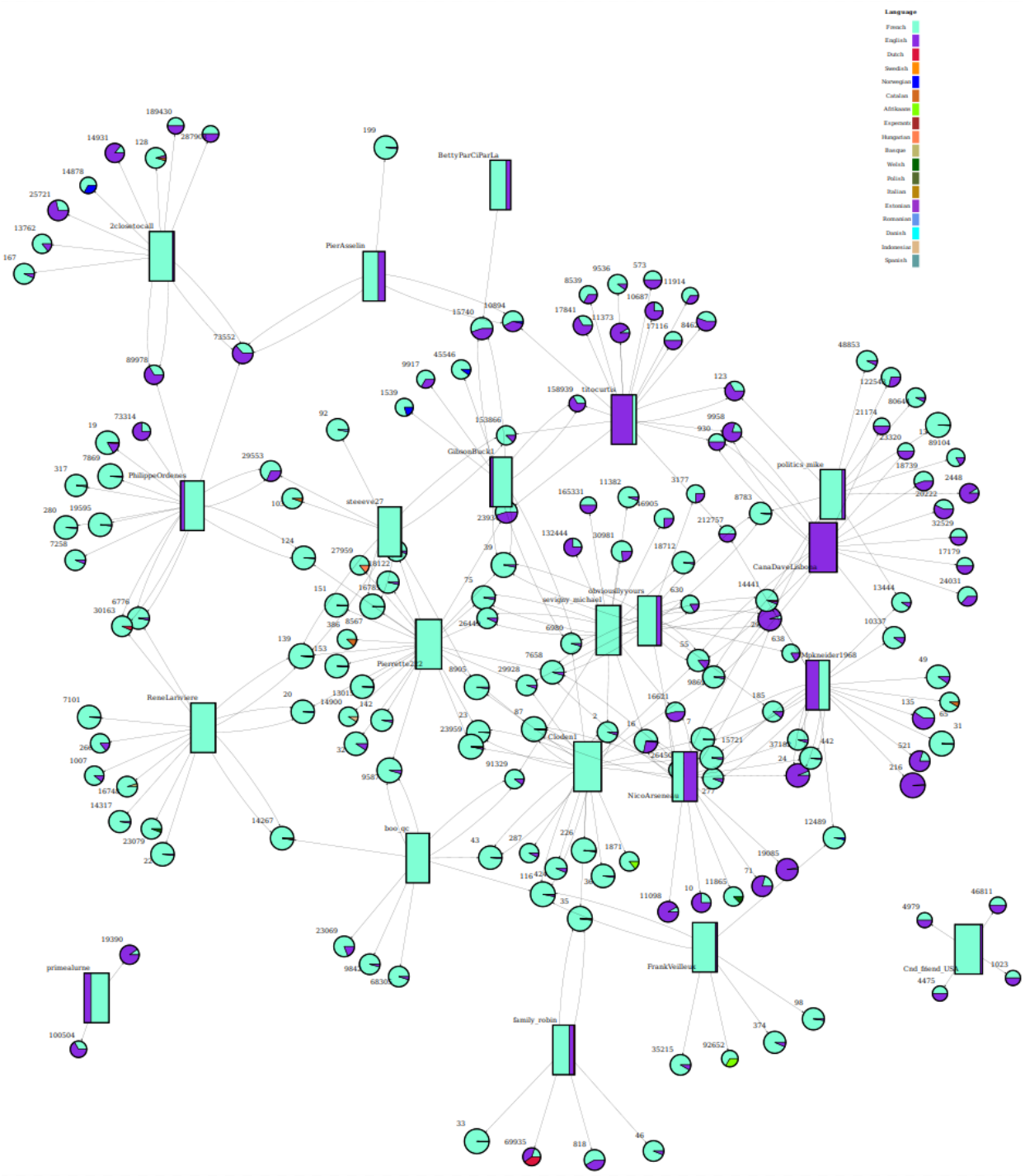

Fig. 3 User-discussion graph with top user components linked to corresponding discussion components

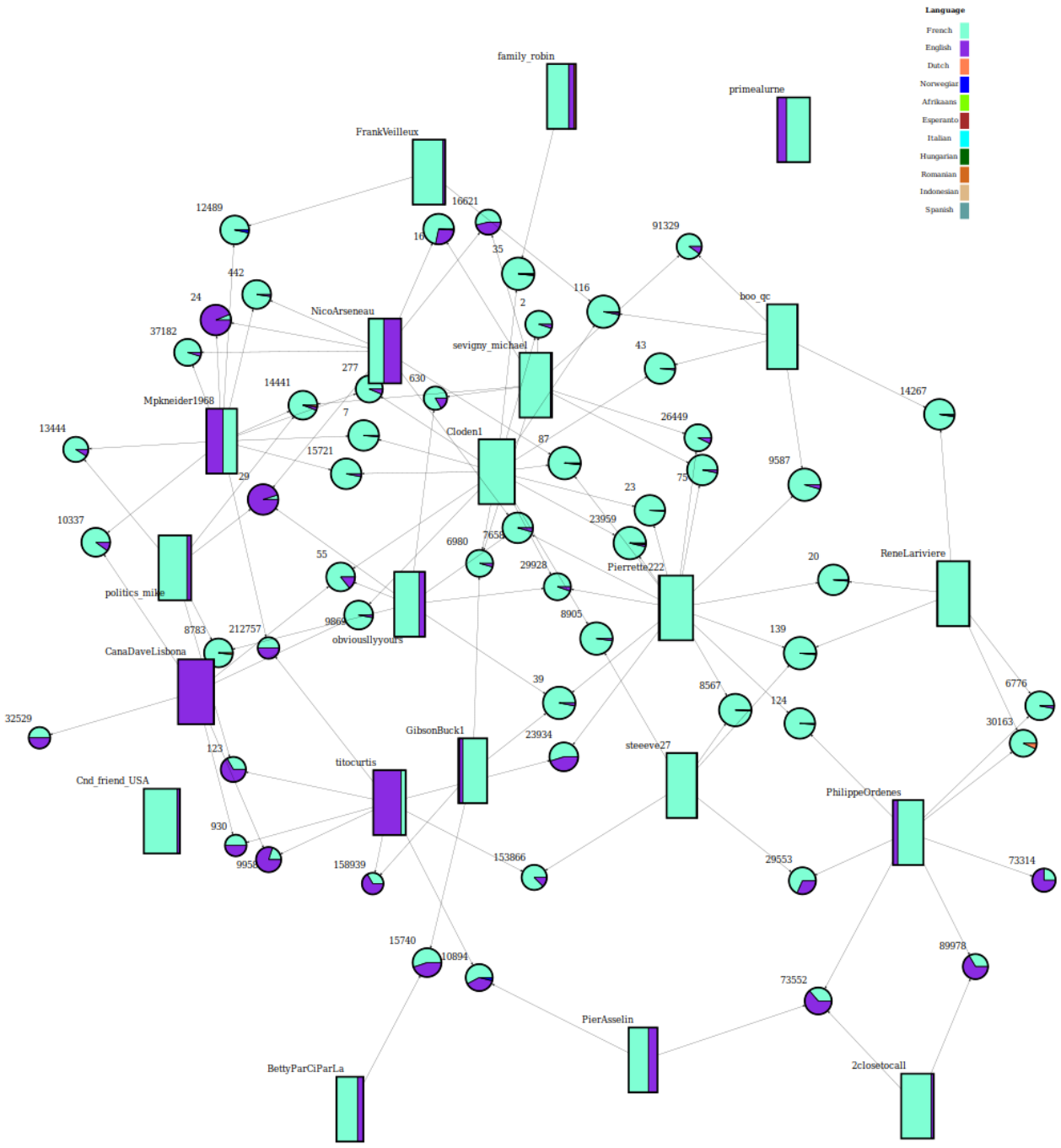

Fig. 4 Common discussion graph of top users

3.3 Output Visualization

After labeling for better visual clues, algorithm output is hosted in a graph structure representing the relationship between extracted users and discussions. The graph structure specifications presented here are useful for understanding the result files.

User Discussion Graph Structure.

Graph components: The user discussion graph structure includes top user components (sorted by maximum CW score) with directed edges toward their corresponding public Twitter discussions (sorted by maximum language users score). The user component is represented as a rectangle, whereas the discussion component has a circular shape.

Color code for graph components: Several color combinations are used to depict the following:

- In user component: proportion of various language(s) used overall, in all posts made by a given user
- In discussion component: proportion of various language(s) used by all users in a 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.

Public Discussion Graph Structure.

Graph components: The public discussion graph illustrates a Twitter discussion scored by maximum language users within a discussion. Each node corresponds to a tweet in the discussion. The arrow of the directed edge points toward the tweet being replied to by 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 follows:

- The upper-half colors correspond to posts' dominant classified languages.
- The lower half of the node represents a dominant user who posted more frequently than others. Each of these users gets a distinct color. Legends for users and languages are displayed alongside the figure.

The black circle around the node stands for "unclear" language classification, as happens when a posting contains all nonpruned words such as emoticons.

4. Results and Commentary

4.1 Language Percentages

The framework described generated results from a 2018 Quebec election dataset. While Table 1 describes the percentages of the classified languages, note that 3% of postings were classed as multilingual or mixed.

Table 1 Percentage of languages classified by the labeling algorithm in 2018 Quebec election dataset

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

4.2 Conflicting Classifications

Languages classified for 4.58% tweets are in disagreement with the language tag provided by Twitter. This is due to Twitter's limitation of single language tags. As Twitter does not provide multilingual tags, all the tweets classified as multilingual or mixed by our algorithm disagree with the Twitter's single language labels. Table 2 gives examples of tweets in disagreement.

Table 2 Tweets language labels (generated by the labeling algorithm on 2018 Quebec elections dataset) in disagreement with Twitter-provided language labels

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'

4.3 User-Discussion Graph

Figure 3 is a user-discussion graph representing the relationships among the top 20 users and their corresponding scored discussions. Figure 4 is a refinement that represents the relationships among the top 20 users, limited to only the scored discussions in which more than one of these users participates; that is, the shared or common 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.

Figure 5 displays a word cloud of all hashtags, which occur at least twice in the common discussions displayed in Fig. 4. This verifies the consistency of the discussions' content as majority of the hashtags correspond to the Quebec elections.

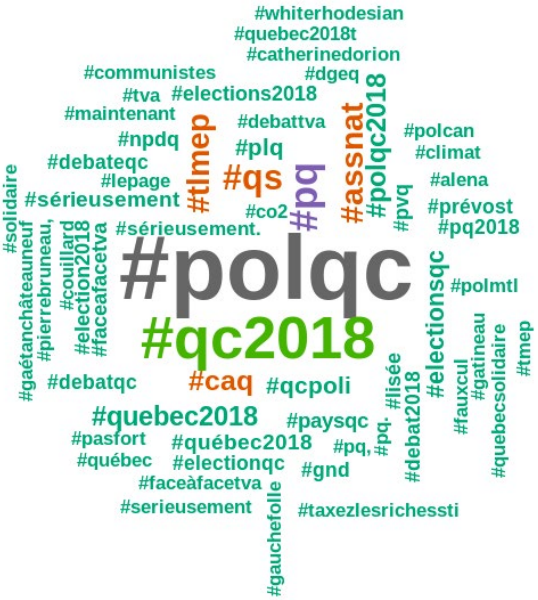

Fig. 5 Word cloud of hashtags of common discussions

We can see from the graphs in Figs. 3 and 4 that most of the discussion content is in French or English. For the top discussions displayed, language dominance on the graphs is fairly similar to that of the language percentages in Table 1. More important, the results of the scoring functions for both users and discussions tend to indicate significant and proportional language alternation by the users. This feature can be used not only to discover multilingual user communities, but also to analyze their detected discussions. Discovery of CW communities can enable or enhance analysis of factors related to cultural diversity, which drive public trends. For instance, the impact of a user's linguistic behavior on other users' opinions or response can be evaluated by checking the influence of an individual user's CW on other users' responses and language in the same discussion.

4.4 Public Discussion Graph

Figure 6 is a public discussion graph, which includes participation from three top users. Associated postings in the discussion displayed are described in detail in Table 3.

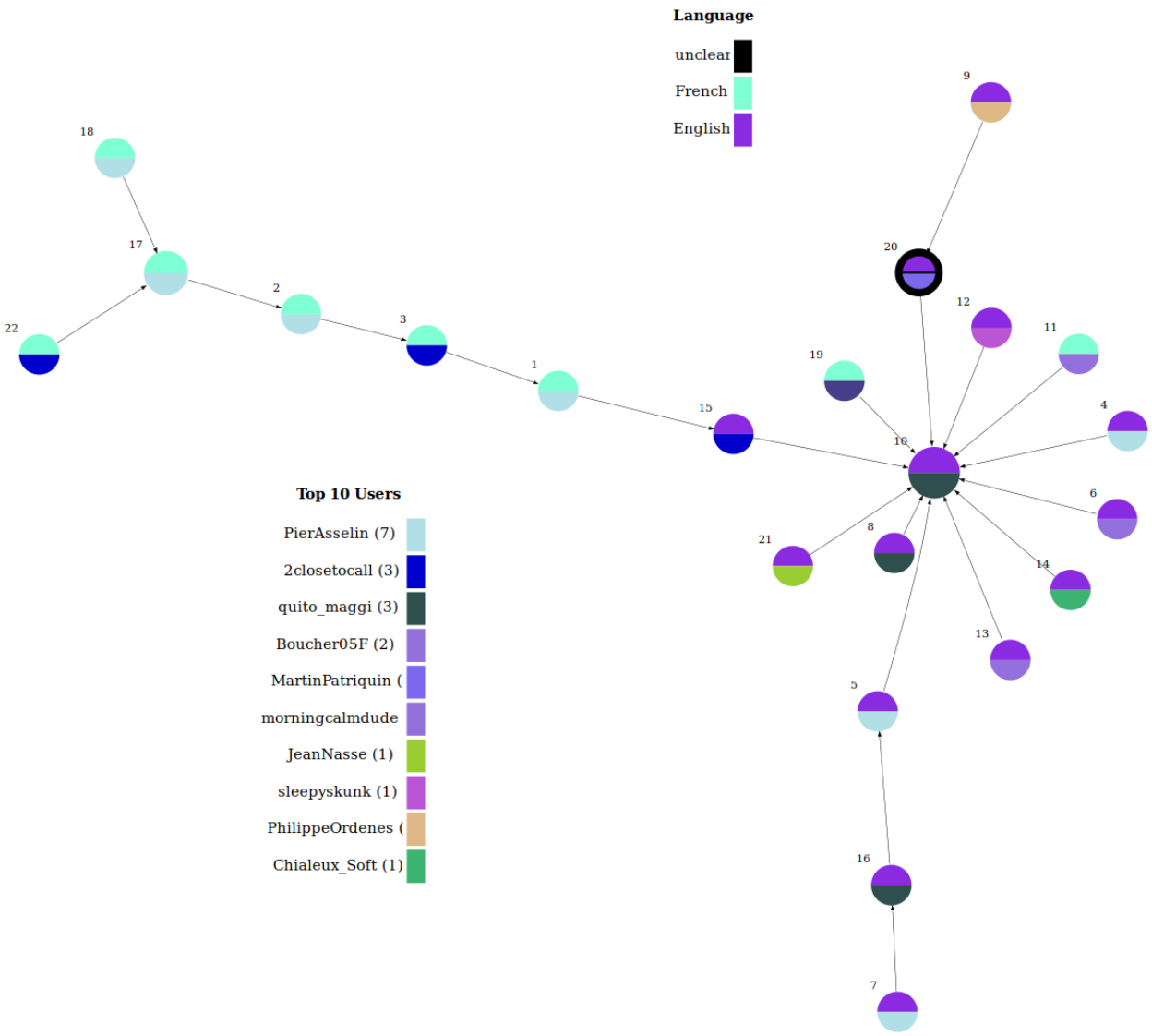

Fig. 6 Public discussion graph example with participation from top users

Table 3 Associated text component table for public discussion graph in Fig. 6

ID	Tweet text	Classified lang	Twitter lang
1	"@2closetocall @quito_maggi @partiquebecois @QuebecSolidaire @coalitionavenir Est-ce que les autres maisons perçoivent le même signal?"	'fr'	'fr'
2	"@2closetocall @quit_maggi @partiquebecois @QuebecSolidaire @coalitionavenir 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 @coalitionavenir 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 inter-esting 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 @coalitionavenir 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'
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 @coalitionavenir Par curiosité, tou-jours 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 @coalitionavenir 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 @coalitionavenir 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 @coalitionavenir Le poll tracker est inutile cette élection car il n'inclut quasiment aucun sondage Mainstreet"	'fr'	'fr'

Public discussion graphs are important for inferring user-related details, to include linguistic behavior, from the reply structure. Detailed discussion analysis can lead to discovery of factors affecting community trends and concerns based on response characteristics of participating users. Graph expansion—an increase in the number of nodes—indicates greater user participation. Current topics can spawn new threads, the different but related content of which can initiate further discussions. In these cases, analysis of raw tweet content, which is made possible with public discussion graphs such as Fig. 6, is advantageous.

5. Conclusion

We have presented a novel approach to discovery of CW SM discussions in a community context. Our work is unique in that it unifies two problem areas: one is the detection of multilingual SM communities and the other is the detection of CW in the body of SM texts. When the task is to determine the social meaning of a CW occurrence, it is important to consider 1) the participants and their relationship, 2) the topic under discussion, and 3) the social setting of the linguistic exchange. By recognizing the importance of, and simultaneously addressing, the perspectives of community context and CW expression in SM, we can give a positive response to the research question. By providing proxies for the three driving factors, our unified five-step strategy of topic-based community detection, word-level LID within an SM posting to identify and rank CW discussions and their participants, culminating in discussion-tree representations, is a technique that allows for close-grain sociolinguistic analysis of the CW phenomenon.

This approach lays the groundwork for future investigations involving CW and social meaning, creating baselines for specific communities and language pairs, and developing ground truth for LID and social meaning for specific factors. Future work might also lead to studies of the social meaning and function of CW expected in social network intersections between communities speaking primarily different languages, important comparisons of opinion and facts shared within and between language communities, and exploration of the role of CW in bridging or enforcing cultural boundaries.

6. References

1. Papadopoulos S, Kompatsiaris Y, Vakali A, Spyridonos P. Community detection in social media. Data Mining and Knowledge Discovery. 2012;24(3):515–554.
2. Qi G-J, Aggarwal CC, Huang T. Community detection with edge content in social media networks. Proceedings of the IEEE 28th International Conference on Data Engineering; 2012.
3. Crystal D. The Cambridge encyclopedia of language. Cambridge University Press; 1987.
4. Auer P. Code-switching in conversation: language, interaction and identity. Routledge; 2013.
5. Auer P. A conversation analytic approach to code-switching and transfer. In: Heller M, editor. Codeswitching: Anthropological and Sociolinguistic Perspectives. Mouton de Gruyter; 1988;48:187–213.
6. Skiba R. Code-switching as a countenance of language interference. The Internet TESL Journal; 1997. http:/iteslj.org/.
7. Baheti A, Sitaram S, Choudhury M, Bali K. Curriculum design for code-switching: experiments with language identification and language modeling with deep neural networks. Proceedings of the 14th International Conference on Natural Language Processing; 2017.
8. Barman U, Das A, Wagner J, Foster J. Code mixing: a challenge for language identification in the language of social media. Proceedings of the First Workshop on Computational Approaches to Code-Switching; 2014. p. 13–23.
9. Solorio T, Blair E, Maharjan S, Bethard S. Overview for the first shared task on language identification in code-switched data. Proceedings of the First Workshop on Computational Approaches to Code-Switching; 2014. p. 62–72.
10. Jain D, Kustikova M, Darbari M, Gupta R, Mayhew S. Simple features for strong performance on named entity recognition in code-switched Twitter data. Proceedings of the Third Workshop on Computational Approaches to Code-Switching; 2018. p. 103–109.
11. Mave D, Maharjan S, Solorio T. Language identification and analysis of code-switched social media text. Proceedings of the Third Workshop on Computational Approaches to Code-Switching; 2018. p. 51–61.

List of Symbols, Abbreviations, and Acronyms

CRF	conditional random field
CW	code-switching
ID	identification
LID	language identification
NER	named entity recognition
RAPID	Real-time Analytics Platform for Interactive Data-mining
SM	social media

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLD DCI TECH LIB
1 (PDF)	DEVCOM ARL FCDD RLC NC S E KASE
2 (PDF)	UNIV MELBOURNE S KARUNASEKERA A HARWOOD
1 (PDF)	P PADIA
1 (PDF)	L FALZON