

Understanding Multilingual Communities through Analysis of Code-switching Behaviors in Social Media Discussions

Aaron Harwood

*School of Computing and Info. Sys.
The University of Melbourne
Melbourne, Australia
aharwood@unimelb.edu.au*

Shanika Karunasekera

*School of Computing and Info. Sys.
The University of Melbourne
Melbourne, Australia
karus@unimelb.edu.au*

Michelle Vanni

*Combat Capabilities Development Command
Army Research Laboratory
Washington D.C., U.S.
michelle.t.vanni.civ@mail.mil*

Lucia Falzon

*Information Analytics Branch
Defence Science and Technology
Adelaide, Australia
Lucia.Falzon@dst.defence.gov.au*

Prarthana Padia

*School of Computing and Info. Sys.
The University of Melbourne
Melbourne, Australia
ppadia@student.unimelb.edu.au*

Amila Silva

*School of Computing and Info. Sys.
The University of Melbourne
Melbourne, Australia
amila.silva@student.unimelb.edu.au*

Abstract—Currently, the enormous span of social media usage – while providing valuable resources for linguistic behavior analysis – makes tracking and understanding these multilingual discussions a challenging task. We have undertaken a multi-disciplinary comprehensive study of multilingual discussions via the development of specialized data collection techniques that discover and track multilingual users of social media, and their associated discussions, within a defined geographical region. To facilitate automatic discussion analysis of large numbers of discussions we generated a machine learning model based on ground truth data obtained from Amazon Turk. Our approach goes beyond analyzing social media posts in isolation, by analyzing them in the context of the discussion in which they appear. We show a selection of example discussions found using our approach which reveals a number of interesting socio-linguistic interactions in the communities that we sampled, in support of approach as a general methodology for multilingual community analysis.

Index Terms—code-switching, machine learning

I. INTRODUCTION

Social media facilitates large numbers 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 of any depth that they encounter. Multilingual individuals often tend to utilize alternate languages in a single conversation – code switching – for effective communication in a discussion. Characterizing these discussions helps to analyze the contextual factors affecting the cultural diversity of a community. The variety and diversity of communities engender multilingual discussion, which provides valuable insight into the evolution of trends and public opinion in communities.

Funded in part by US Army Research Office grant W911NF-18-2-0050, 978-1-7281-0858-2/19/\$31.00 © 2019 IEEE

Currently, the enormous span of social media usage – while providing valuable resources for linguistic behavior analysis – makes tracking and understanding these multilingual discussions a challenging task. Significant progress has recently been made by developing community detection algorithms based on the network structure of followers and friends, the interactions of retweets and mentions and the 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 linguistic characteristics of the discussions are ignored in the analysis.

To code-switch is to change from one natural language to another within a conversational turn at talk. There are diverse inventories of the functions of this behavior, which overlap to varying degrees; there are related but contrasting notions, e.g. “code-mixing” as interspersing of individual second-language words into one’s stream of talk; and there are measures of its levels of complexity. For example, Figure 1 shows both code-switching and code-mixing.

How the linguistic and interactional patterns that occur in a discussion are associated with its social meaning or function is a phenomenon whose study requires large datasets comprised of naturally-occurring examples of code-switching. This is particularly important for modeling in order to build technologies for automatic recognition of that meaning or function. In the age of social media, global communication, and big data, assembling corpora of this type of material should be within reach.

A. Big data challenges

Barriers to the collection of high quality multilingual discussions in social media include:

```

1 @user1 @user2 My cousin is coming today. I asked her to buy u one and she said yes.
2 @user2 @user1 are you kidding meeee
3 @user1 @user2 Il y a des news qui rendent le smile. Tell me im amazing. Gonna meet
   her later.
4 @user2 @user1 you are so amazing. pick me up after work
5 @user1 @user2 What times tu finis ? Je les vois à 16h à priori
6 @user2 @user1 and tell me how much i owe her

```

Fig. 1. Example code switching and mixing. @user1 changes from English (line 1) to French (line 3), which is code switching, and uses English words in a French sentence (line 3), which is code mixing.

- *veracity of user generated content* – social media posts are largely or entirely un-moderated, incomplete or inadequately solicited, and so we cannot substantially rely on reported meta-data such as country or language;
- *social media idiosyncrasies and short text* – the constraints and structured interactions of some social media platforms invite a wide variety of creative changes, to otherwise standard language expression, that need to be understood for proper analysis;
- *language identification* – the above aspects combine to make accurate language identification, which is a core aspect of multilingual analysis, considerably difficult, particularly in the face of abbreviations, misspellings, and novel textual expressions;
- *discussion mining* – while the definition of a discussion, in the sense of a set of posts that reply to each other, is easily defined, the process of collecting “high quality” discussions from an overwhelming amount of data is significantly hampered by the large amount of noisy posts and aberrant user social media activity, including spam, trolling and bot activity;
- *high level discussion analysis* – analysis of a social media post can not always be oblivious to its context.

B. Our approach and contribution

We have undertaken a comprehensive study of multilingual discussions via the development of specialized data collection techniques that discover and track multilingual users of social media, and their associated discussions, within a defined geographical region. To this end we made use of RAPID, a real-time stream processing system for both real-time and retrospective analysis of social media. After data collection we obtained ground truth on a sample, for a variety of discussion analysis questions, including language identification for both code switching and mixing, using Amazon Turk. From this ground truth we developed a basic machine learning model that could be applied to the entire data set for discussion mining. Our paper contributes an overall methodology for undertaking multilingual analysis of geographically defined communities.

II. CODE-SWITCHING

Code-switching behaviors carry important sociolinguistic, social and sociocultural meaning. This section discusses the background, importance and recent research targeting code-switching as a means of language development.

A. Code-switching and its impact on language development

Code-switching refers to the process of alternation from one language to another in a single conversation. According to Nilep (2006) code-switching is “defined as the practice of selecting or altering linguistic elements so as to contextualize talk in interaction”.

1) *Code-switching before social media*: Code-switching is common among bilingual speakers and has been identified as a part of bilingual conversation since 1970s. Crystal suggests various possible reasons for speakers to shift between languages. First of these is to compensate for deficiency in a language, which may occur when the speaker is tired or distracted. Another reason is the desire to express unity or support to a specific community. It influences the listeners in both positive and negative ways. A rapport is established between the listener and the speaker when the listener understands and responds with the similar switch. On the other hand, switching may exclude others from the conversation who are not the speakers of the second language.

Skiba (1997) interprets code-switching as a means of language development and suggests that it comes naturally to the bilingual speakers. Moreover, it increases the impact of speech and provides continuity in speech rather than interference, when used as a sociolinguistic tool. Skiba (1997) used theoretical framework (“Think tank” theory) and syntactic framework (Lexical Functional Grammar). Two syntactic structures used are as follows: constituent-structure that represents word order and phrasal groupings and functional-structure that represents grammatical functions like subject and object. Using the above theories on the examples from Efik – English language texts, it was deduced that code alteration provides linguistic advantages rather than obstruction in communication. Skiba (1997) also suggests that code-switching supplements speech in cases of inability of expression and allows the speaker to convey attitude and emotions to connect to the listeners in a better way.

2) *Code-switching in conversation*: Auer (1988) analyzed the code-switching and transfer carried out in conversations among the second generation of Italian migrants, with Southern Italian background. The study was based on corpus built on (non)spontaneous speech used by those children interacting with people around them. 400 code alternation instances in the conversations of 19 children (6-16 years) were analyzed. Not necessarily occurring independently, code-switching based on competence and preference were deduced. With three

supporting hypothesis, it was inferred that the sociolinguistic situation of the Italian migrant's second generation may lead to complete linguistic adaptation (including loss of Italian language) or stabilization as a bilingual community.

In further research, [Auer \(2013\)](#) modelled the situation of the conversation as an interactively achieved phenomenon, suggesting that interaction continuously produces frames for succeeding activities. Code-switching in discussions may lead to creation of new frames or change in some features of the situation and maintain or re-establish a different discussion. The choice of the language influences on the succeeding language choices by the same or the other speakers.

B. Code-switching in language identification task

Code-switching is often practiced by multilingual speakers on social media and has become a challenge for language identification task. Several research papers/projects aim at modelling and proposing solutions for the same [Baheti et al. \(2017\)](#); [Barman et al. \(2014\)](#); [Jain et al. \(2018\)](#); [Mave et al. \(2018\)](#); [Solorio et al. \(2014\)](#). [Barman et al. \(2014\)](#) used supervised classification and sequence labelling, whereas [Jain et al. \(2018\)](#) 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, [Baheti et al. \(2017\)](#); [Mave et al. \(2018\)](#) proposes different approaches for word-level language identification tasks in the context of code-switched social media text. [Baheti et al. \(2017\)](#) employs DNN architecture for training models for word-level language identification and [Mave et al. \(2018\)](#) trains a character n-gram based CRF model which is investigated using interesting code-switching metrics. A bigger picture of the first shared language identification task for code switched data was provided by [Solorio et al. \(2014\)](#). It dealt with analyzing the performance of various techniques adopted by the participating teams to accomplish the task.

Following is the detailed review and summary on above mentioned research based on Twitter or/and Facebook data.

- 1) An automatic language identification mechanism for the code switched languages of social media was proposed by [Barman et al. \(2014\)](#). The dataset composed of Facebook posts and comments that exhibited code mixing between Bengali, English and Hindi languages. Following machine learning techniques were applied and compared:
 - A simple unsupervised dictionary-based approach
 - Supervised word-level classification with and without contextual clues
 - Sequence labelling using Conditional Random Fields
 The results suggested that supervised classification and sequence labelling approach surpassed the dictionary based approach, and that it is important to take contextual clues into consideration.
- 2) [Baheti et al. \(2017\)](#) adopted DNN (Deep Neural Network) architectural strategies (curriculums) for training the code-switched model, defined on combination of monolingual languages - En and Es tweets. The best

curriculum involved first training a network with monolingual training instances, where each mini-batch has instances from both languages and subsequently training the resulting network on the code switched data. The two types of code-switching (1. inter-sentential code-switching (language change across sentences) and 2. intra-sentential code-switching (language change within sentences)) were handled separately using sequential language identification tasks and sentence boundary detection.

- 3) In contrast to [Baheti et al. \(2017\)](#) (which employed neural network architecture for word-level classification), [Mave et al. \(2018\)](#) approached the problem using character n-gram based CRF model training. The dataset included code switched Hindi-English and Spanish-English data extracted from Facebook public posts and Twitter. Various code-switching metrics for investigation were deployed as follows:
 - Multilingual index (M-index): A word count based index that quantifies the inequality of the language tags distribution in the corpus.
 - Integration index: An approximate probability of any given token to be a switch point in the corpus
 - Code mixing index: Finding the most frequent language and then calculating the frequency of words of all other languages present.
- 4) Named Entity Recognition (NER) in code switched Twitter data was performed by [Jain et al. \(2018\)](#). A performance comparison of the model built with simple features (sans multilingual features) using CRF classifier against few sophisticated machine learning models was carried out. The simplest NER model, trained without using any language identification or translations worked best. The other more sophisticated experiments improved the recall, but damaged the precision too much.
- 5) Combining different machine learning approaches, [Solorio et al. \(2014\)](#) provided a bigger picture of the first shared task on language identification on code switched data. The experiment involved 42 participating teams and 4 language pairs: Modern Standard Arabic- Dialectal Arabic (MSA-DA), Mandarin-English (MAN-EN), Nepali-English (NEPEN) and Spanish-English (SPA-EN). Each team designed its own system architectures for language identification task on twitter data. The architectures ranged from a simple approach based on frequencies of character n-grams to more complicated and sophisticated approaches using word embedding, extended Markov Models and CRF autoencoders. The evaluation showed that language identification at the token level became more difficult when the languages present were closely related (like MSA-DA).

C. Language Detection in posts

The automatic detection of language from text is a well-studied problem, which was solved with near-perfect scores in some of the previous works ([McNamee, 2005](#)). However,

most of the previous works assume that the documents are monolingual. There are only limited works on language detection from multilingual documents. Prager (1999) proposes a vector space model, in which word-frequency based feature vectors are constructed for each language. For a given test document, it uses the cosine similarity between the feature vector of the test document and the feature vector for each language to predict the proportions of languages in the text document. Lui et al. (2014) models the distribution of words in each language using a generative approach and this model is able to reliably estimate the relative language proportions of a document even if the training dataset consists of both monolingual and multilingual documents. However, none of the aforementioned methods can perform accurate word-level language prediction for the multilingual documents. Instead of that, they predict the relative proportions of languages in the documents. One trivial approach to extend existing models to make word-level prediction is that detecting language for each monolingual text segment after segmenting the document into multiple monolingual segments. Teahan (2000) proposes a supervised approach to detect monolingual text segments in a multilingual text document. However, such an approach is not suitable for many social media posts, which can be short in length and includes noisy tokens such as emojis, hashtags, and mentions.

Word-level language prediction can also be considered as a sequence prediction task, which is a quite popular modelling task in NLP (e.g., Part-Of-Speech tagging and Named-Entity recognition). The models with Markov assumption such as Hidden Markov Model (HMM) (Baum and Petrie, 1966), Maximum Entropy Markov Model (MEMM) (McCallum et al., 2000) and Conditional Random Field (CRF) (Lafferty et al., 2001) were quite popular in early days to address sequence prediction task. Generally, CRF is the superior out of the above models due to its relaxed independence assumptions and its potential to accommodate flexible feature space. However, all these models were generally outperformed by recurrent neural networks such as LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) as they can capture the long-distance relationships of sequences. More recently, Huang et al. (2015) proposes BiLSTM-CRF neural architecture to exploit the benefits of both LSTM and CRF models. However, such neural models require a lot of training data to generalize well due to the large number of parameters.

III. DISCUSSION MODEL

Consider a set of social media posts \mathcal{R} where $r = \langle u_r, p_r, l_r \rangle \in \mathcal{R}$ is a post with u_r being the author of the post and either $p_r \in \mathcal{R}$ is a replied to post or $p_r = \emptyset$, meaning the post does not reply to another post, and l_r is the language of the post (given by the social media platform). Consider a discussion graph $G = (V, E)$ where $v \in V$ iff $v \in \mathcal{R}$ and there is some $u \in \mathcal{R}$ such that $p_u = v$ or $p_v = u$ (remove all posts that are not replied to and do not reply to anything), and for $i, j \in V$, $(i, j) \in E$, iff $p_i = j$ (edges exist from a post to the post that it replies to). Generally, G is then a forest of

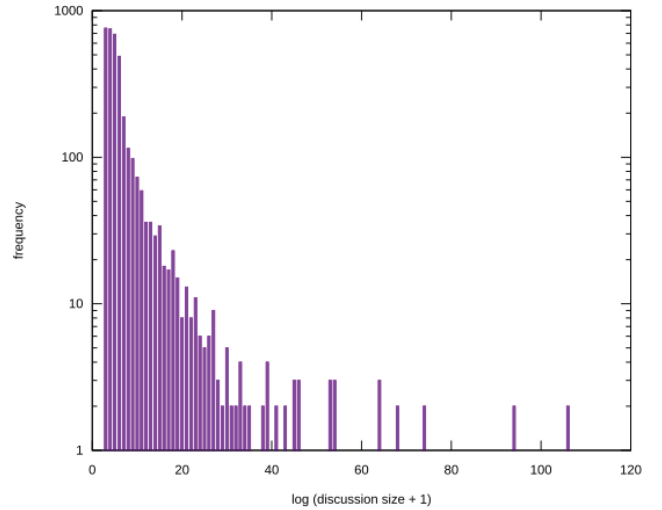


Fig. 2. Typical distribution of discussion size in a collection of discussions, with 21944 posts in total and 3504 discussions.

discussions, where each connected component of G is a single discussion of size at least 2. Each $r \in \mathcal{R}$ is either included in exactly one discussion or excluded.

If the social media posts are found in online forums, such as Reddit, then they naturally form discussions, since each forum thread with at least 2 posts fits the definition of a discussion, and the set of such forum threads from a forum is G exactly. However in some social media platforms, such as Twitter, discussions must be pieced together by tracing individual “reply to” chains until the root of the discussion is found, and then finding branches of the discussion based on those chains having a common root. Furthermore, when collecting social media posts as a third party, the platform may not provide all posts, e.g. private posts are not revealed, and asking for all posts that reply to a given post may not be allowed. This can lead to discussion fragmentation and missing information, especially with respect to the root of the discussion.

Figure 2 shows a typical distribution of discussion size in a set of 3504 discussions, with 21944 posts in total. At a high level we are interested in extracting a representative set of discussions for further detailed (manual) analysis. In the first instance this can be achieved by ranking the discussions in some way and examining the top N discussions. This leads to the question of how we score discussions for ranking. We have considered a range of different scoring functions, which generally take a connected component $g \in G$ and give a score $score(g)$, shown in Table I. Note that $|g|$ is the cardinality of set g , $\langle x, \cdot, \cdot \rangle$ is any tuple r with $u_r = x$, and similarly for other components of r . Also, we can consider scoring the authors over all the posts in the collection, with respect to how many languages they are observed using:

$$C_u = |\{x \mid \langle u, \cdot, x \rangle \in V\}|.$$

TABLE I
TYPICAL SCORING FUNCTIONS

Ranking approach	Discussion score
Discussion size	$ g $
Unique authors	$ \{x \mid \langle x, \cdot, \cdot \rangle \in g\} $
Unique replied to authors	$ \{x \mid \langle \cdot, y, \cdot \rangle \in g, y_p = \langle x, \cdot, \cdot \rangle\} $
Unique languages	$ \{x \mid \langle \cdot, \cdot, x \rangle \in g\} $
Multilingual users	$ \{x \mid \langle x, \cdot, \cdot \rangle \in g, C_x > 1\} $

Ranking on discussion size reveals discussions that have a star-like, or extremely shallow tree. Such discussions tend to be post from a well known user such as a politician, that invites significant numbers of replies, with very little actual discussion taking place. Ranking on unique authors is similar to ranking on discussion size, due to many of the discussions having posts from each author exactly once. On the otherhand ranking on the unique replied to authors leads to discussions that are more bushy with a reasonable degree of branching. Such discussions are more interesting for analysis, but they do not include other kinds of discussions, such as those between two authors that continue for many turns. When ranking with respect to languages we have found that the language labels given by the social media platform can be insufficient to provide accurate ranking, and more so, the language label is monolingual in that it assume a given post is always in a single language, which frustrates code mixing/switching analysis.

IV. DATA COLLECTION

We make use of the Twitter social media platform to obtain data for our research. Although Twitter provides Application Programming Interfaces (APIs) to track social media posts or tweets based on keywords, users, location, languages, etc., collecting multilingual, context-rich discussions using these APIs is not a straightforward task. This section describes the approach we used to collect multilingual discussions used in our analysis.

We used a real-time data tracking system, *Real-time Analytics Platform for Interactive Data mining* (RAPID) for data tracking and analysis. RAPID is a cloud-based platform capable of collecting data from social media services 24/7 based on user-defined queries (also referred to as topics). Figure 3 shows the high level architecture of RAPID. RAPID supports a real-time analytics capability as well as post-analysis of data collected through real-time tracking which is stored in a database. RAPID provides a rich set of data collection and analytics capabilities, which include multi-source dynamic tracking, query expansion, community detection, topic tracking, event detection and discussion analysis. For collecting multilingual discussions we used the query expansion and the discussion analysis capabilities in RAPID.

A. Query expansion

RAPID offers a query expansion capability, through allowing user-defined expansion algorithms to be plugged to the real-time streaming track. The query expansion algorithm

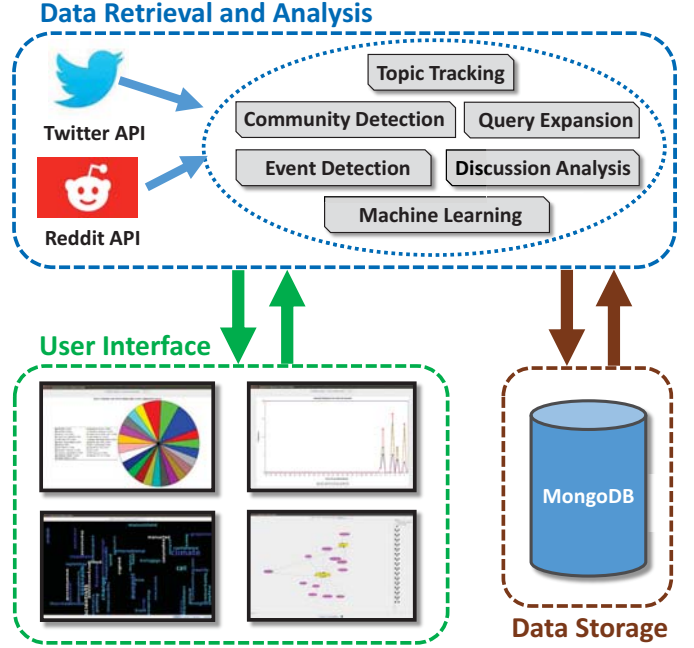


Fig. 3. Overview of RAPID System

takes the real-time data stream as input, and determines new keywords, users and/or locations that needs to be tracked, which in turn are used to dynamically change tracking (through changes that take place periodically). For multilingual user tracking we developed a Multilingual User Discovery algorithm as follows.

Let $\mathcal{R} = \{r_1, r_2, \dots, r_n, \dots\}$ be a stream of social media records from user posts, received through tracking a set of keywords \mathcal{K} , users \mathcal{U} , and geographical bounding boxes (locations) \mathcal{B} . Similarly to the previous section, each post $r \in \mathcal{R}$ is represented by a tuple $\langle u_r, p_r, l_r \rangle$ where u_r represents the posting user, p_r the replied to post (if any, but not needed in this section) and l_r the language of the post. Let $\mathcal{X} = \{x_1, x_2, \dots, x_M\}$ be a set of languages of interest of $(M = |\mathcal{X}|)$, where each record $x \in \mathcal{X}$ is represented by a tuple $\langle l_x, c_x \rangle$, where l_x represents a language of interest and c_x represents a count of occurrences of that language, which is used for thresholding the particular language. We categorize a user as a multilingual user if the user posts at least c_x posts in each language l_x , in $x \in \mathcal{X}$, during the observed period since the start of the stream \mathcal{R} . Algorithm 1 shows how up to $N = |\mathcal{U}|$ multilingual users are identified for tracking, while periodically adding users to the track every T minutes.

For the data sets used in this paper we collected tweets using three separate queries. For all three queries the language set \mathcal{X} was chosen as, English, French and Spanish, with each having a threshold of 5, and we selected $T = 5$ and $N = 400$. Any time a track is changed Twitter requires a 1 minute pause before reconnecting, and therefore we chose $T = 5$ to avoid re-connecting too frequently. When a user is chosen for tracking, RAPID allows the user to be added as a

Algorithm 1: Multilingual User Discovery Algorithm

input : \mathcal{K} , N , T , \mathcal{K}_0 (Initial set of keywords to track),
 \mathcal{B}_0 (Initial set of bounding boxes to track).

1 **begin**
2 $\mathcal{K} \leftarrow \mathcal{K}_0$, $\mathcal{B} \leftarrow \mathcal{B}_0$, $\mathcal{U} \leftarrow []$.
3 $t = 0$ Start timer.
4 $\mathcal{T} \leftarrow []$ Newly indentified multilingual users.
5 Track posts using \mathcal{K} , \mathcal{U} , \mathcal{B} .
6 **repeat**
7 **for** each tweet $r \in \mathcal{R}$ **do**
8 **if** language l_r exists in \mathcal{K} **then**
9 Increment user u_r language count for
 language l_r by 1
10 **if** user u_r is multilingual **then**
11 $\mathcal{T} \leftarrow \mathcal{T} \cup \{u_r\}$
12 **if** ($t == T \&\& |\mathcal{U}| > 0$) **||** ($|\mathcal{U}| + |\mathcal{T}| == N$)
 then
13 $\mathcal{U} \leftarrow \mathcal{U} \cup \mathcal{T}$
14 Track posts using \mathcal{K} , \mathcal{U} , \mathcal{B} .
15 $t = 0$, $\mathcal{T} \leftarrow []$
16 **until** $|\mathcal{U}| \leq N$;

keyword in addition to purely following the user, which what Twitter provides when a user is tracked. We found this feature extremely useful for collecting discussions because through tracking the user as a keyword we fetch all tweets that mention the user; that allows us to get the replies of the user because in the replied tweet the original user appears as a mention. However, because Twitter has a limitation of 400 keywords it can track, this meant that we had to set $N = 400$. Each of the queries seeded with no keywords ($\mathcal{K}_0 \leftarrow []$) and a single bounding box (\mathcal{B}_0) per query placed around Australia, California and France, respectively. These queries collected data over a period of three months starting February 2019.

V. GROUND TRUTH

To obtain ground truth upon which to train a machine learning model for discussion analysis, we devised an Amazon Turk experiment, based on 584 discussions selected based on a first-order approximation of their multilingual content from the language labels provided by the social media platform. For simplicity, we considered discussions that contained only English, French, Italian and Spanish, though some other languages appeared in the set due to errors of language identification by the social media platform.

For each post in each discussion that replied to a post that had identifiable text, we issued a Human Intelligence Task (HIT) to:

- highlight and label fragments of text within the text post,
- indicate true/false for whether the post contains at least one full sentence or meaningful clause,
- categorize the relevance of the post on a five point Likert scale,

- indicate true/false for whether:
 - the post provides a response or additional information to the replied to post,
 - the post opposes the replied to post,
 - the post supports the replied to post,
 - the post questions the replied to post,
- indicate on a five point Likert scale how sure they are of their answer to the above question.

A. Language Identification

The Workers were asked to highlight a fragment of text by selecting it and clicking a language label button. The highlighted text will be shown as labelled in the language that was clicked. They were asked to label as many identifiable fragments as they could, and to select whether the text post contains at least one meaningful sentence or clause. The following additional requirements were provided:

- make use of the Other (oth) label to indicate a language that is not listed,
- for names of places/persons/organizations, label with a language if that would be considered the usual use of the name in that language,
- the following fragments should not be labelled: screen names, hashtags, urls, emoticons, numbers and any other text that is not a dictionary word or discussed in these instructions,
- misspelled words, abbreviations, slang, filler expressions like “Ummmm” and acronyms (e.g. “lol”) should be assumed to be in the language that they likely represent.

B. Discussion Analysis

The Workers were asked to read the discussion presented, and consider the yellow highlighted post within the context of the discussion that it appears. They were then asked to select a “relevance” category that best represents the text post within the context of the discussion:

- 5 The post is clearly relevant to the discussion
- 4 The post is probably relevant to the discussion
- 3 It is unclear if the post is or is not relevant to the discussion
- 2 The post is probably not relevant to the discussion
- 1 The post is clearly not relevant to the discussion

They were ask to select as applies from the remaining questions and to indicate how sure their answers to question 4.

1) *Definition of Relevance:* A text post was asked to be considered relevant to the discussion if its content relates to and/or provides meaning that contributes to the discussion. Examples of posts that are not relevant include spam, trolling, posts that have no discernible meaning, etc.

C. Reaching consensus

A post would have multiple HITs if the discussion had more than one of Spanish, Italian, French in it, and was therefore sent for evaluation to more than one country. As a result, some ground truth was computed only on the basis of 3 replicates (Assignments), while other ground truth was computed on the basis of 6 or possibly 9 replicates.

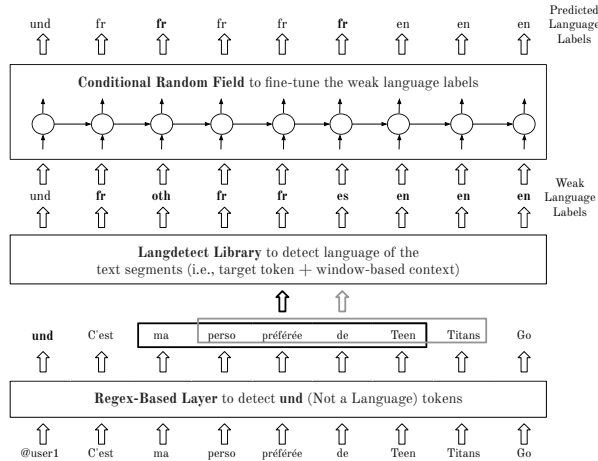


Fig. 4. Proposed Framework for Language Detection in Multilingual Posts

To reach consensus for a true/false type selection (Questions 2 and 4), more than half of the replicates needed to be identical responses. There had to be more than one response, else consensus was not considered to be reached. For two responses, they had to be identical. For three responses, two of them had to be identical. For four responses, 3 of them had to be identical, and so on.

To reach consensus for language label on a word was the same as true/false, but with more categories. As above, consensus was only reached for each word language label if there was a language label that was identical for more than half of the replicate responses. The number of disagreements, or cases when there was no consensus, was counted.

For the 5 point scales, the mean value was used. No consensus was required. Hypothesizing that these kinds of questions have natural biases for each turker, the mean value seems appropriate rather than the most frequent category.

Overall we generated 1873 HITs, of which we obtained consensus for 1560 posts, the remaining 313 being in doubt. These 1560 posts were used for training a machine learning model explained next.

VI. MACHINE LEARNING MODEL

A. Language Detection in Multilingual Posts

Although there exist well known software solutions¹ to predict the language of a given text document, they do not perform accurately when multiple languages are present in the given text. Also, we have empirically found that the existing systems are not accurate enough to make word-level language predictions to detect code-switching behaviors. Hence, this work proposes a framework to make word-level language predictions in multilingual posts (depicted in Figure 4), which consists of three layers:

- 1) **Detection of und tokens.** This layer detects *mentions*, *hashtags*, *urls*, *emojis*, *numbers*, and *punctuations*

¹<https://code.google.com/archive/p/language-detection/>

present in the given post using regular expressions and publicly available python libraries. The punctuation list in Natural Language Tool Kit (NLTK)² and emoji list in emoji³ library are used to detect *punctuations* and *emojis* respectively. All other aforementioned *und* tokens are detected using regular expressions. For instance, *hashtags* are detected as the words starting from '#' symbol.

- 2) **Prediction of weak language labels.** To detect weak language labels of the words in a given post, the publicly available langdetect⁴ library is used. However, It is empirically observed that labeling word by word using langdetect is not reliable enough. Because such an approach neglects the context of the target word when detecting its language. Hence, our approach defines the context for each word using a window (size was set to 10 after tuning with a development set) centered around the word. Then, langdetect uses the sequence of words in the context window (as they appear in the sentence) of a word to predict its language.

- 3) **Fine-tuning the weak language labels.** We formulate the fine-tuning step as a sequence prediction problem. Conditional Random Field (CRF) (Lafferty et al., 2001) is used as the model due to its higher modeling potential compared to the other underlying sequence prediction models such as HMM and its simplicity (i.e., ability to generalize with small training datasets) compared to the sophisticated recurrent neural models (e.g., LSTM, GRU, and BiLSTM-CRF). CRF++⁵ library is used to train the CRF model. For each target word, weak language labels and words appeared within a window (size is set to 5, which is the default setting in CRF++) around the target word and their bigrams are considered as features (Please refer <https://taku910.github.io/crfpp/> for the detailed description of the feature template).

The proposed framework shows $\approx 93\%$ accuracy (using 3 fold cross validation of the training dataset) for the task of predicting word-level language labels in multilingual posts.

B. Answering Questions about Posts

This task asks 6 different questions about each test instances: (1) does the post contain at least one meaningful sentence or a clause?; (2) does the post provide a response or additional information as a reply?; (3) does the post oppose as a reply?; (4) does the post support as a reply?; (5) does the post question as a reply?; and (6) is the post relevant to the discussion?. The first 5 questions are TRUE/FALSE questions and the answer for the last question should be in the range from 1 to 5. We have empirically observed that the distribution of the answers for each question in the training datasets is skewed significantly. For example, $\approx 98\%$ of labeled posts in the training dataset are full sentences and have TRUE as the label for the first question. Following the light of the aforementioned

²<https://www.nltk.org/>

³<https://pypi.org/project/emoji/>

⁴<https://pypi.org/project/langdetect/>

⁵<https://taku910.github.io/crfpp/>

observation, our model uses the majority baseline to make predictions for the test instances. Besides, a few heuristically motivated conditions are used in some of the tasks to fine-tune their labels. For example, if "???" punctuation mark is present in a post, our model returns TRUE as the predicted answer for the fifth question despite the majority baseline for the fifth question is FALSE.

VII. DISCUSSION ANALYSIS

We examined the highest scoring discussions on the different dimensions scored and here we provide some of our observations. Overall, all of the discussions seem to fall into 4 categories:

- 1) Monolingual with just an odd word or two from another language. We consider these to be just normal conversations – similar to when one uses a term like *uber* or *magnifique* in conversation. A single word might play a role in reinforcing the “support” “oppose” or “question” function of a posting. For example:
 - Figure 5, Post 3 supports Post 2, *Ah d'accord:FR, nice person:EN*;
 - Figure 6, Post 3 questions Post 2 with texting slang (exaggerated for enthusiasm);
 - Figure 7, Post 6 supports Post 5 and it's an intra-discussion code switching (CS).
 This category is also important for pointing out where texting slang or function words (*ouiiiiiii*) are modified for emphasis, where entity names can be identified, for possible pre-processing, tagging, or other special handling, as in Figure 8.
- 2) Bilingual, and sometimes multilingual. These are much better examples of CS tend to show how some bilingual people speak to each other, e.g. for the purpose of demonstrating shared history/identity as in Figure 11.
- 3) Some discussions were essentially monolingual except for the very last post – somewhat like a parting comment, perhaps to impress.
- 4) There are several examples of users posting some greetings (*good night, good morning*) or some compliment (*you're beautiful*) to several users and in a few different languages, e.g. shown in Figure 12.

For the last 3 categories above it is useful to consider the concept of Adjacency Pair (AP) [Schegloff and Sacks \(1973\)](#), a type of conversational turn-taking which consists of one utterance by one speaker (the first turn) which directly provokes a responding utterance from another speaker (the second turn), e.g. Greeting-Answer, Greeting-Greeting, Compliment-Compliment and Request-Acceptance/Refusal. In multilingual conversation, CS seems to have a role in establishing the social bond created by the AP, which functions to convey the first speaker's willingness to acknowledge the feelings of the second speaker. We can see this for example in Figure 1, the example given in the Introduction, Post 3 contains the first turn of the AP, a CS, *Il y a des news qui rendent le smile.:FR* >> *Tell me im amazing.<< Gonna meet her later.:EN*. In

Post 4 there is the second turn, acceptance, along with the first turn of a second AP, >> (AP1, T2) *you are so amazing.<< >> (AP2, T1) pick me up after work<<*. Post 5's second turn, acceptance, consists of a CS, *What times:EN tu finis.:FR ? [...]*. What may turn out to be significant here is that the CS renders both APs as understandable as possible, a phenomenon which invites further study. The first speaker uses French to start their post, then, with a CS, uses English to take the first turn of AP1, which elicits a second turn response also in English, followed by the first turn of AP2 also in English and a post which starts in English to formulate the second turn, the Response, then, with a CS, is completed in French, the language of the first speaker in the AP-initiating Post.

Figure 9 shows a Question-Answer and Compliment-Compliment with Three Part Interchange. Here, Post 1 is an assertion, which prompts a CS Question-Answer AP then a Compliment-Compliment AP with intra-discussion CS Three Part Interchange. Post 1 is in English, *Sis was cute and skinny tonight*. Post 2 is the first turn, Question, in AP1, with intra-discussion CS, *tu restes jusqu'à quand loulou*. Post 3 is second turn, Response, reflecting the tacit bond, in the same language, *Je repars demain :(. Post 4 starts the second AP, with the Compliment attached to the Greeting, d'ailleurs bon retour ma puce*, followed by Post 5, a same-language Compliment attached to an expression of Thanks, *Merci ma chouchoutte*. Post 6 rounds out the discussion with last line in the Three-Part Interchange, a clarifying question, intra-discussion CS, *are you in Las Vegas*.

An AP with Three-part Interchange might be relevant for the function of the third unit, which can serve, among other functions, to evaluate, show comprehension, request clarification, or bound a topic. Discussions falling into category 3 may likely have a CS final Post, which performs one of these functions, especially if preceded by an AP, the two parts of which are in the same language.

Figure 10 shows an Opinion-Opinion AP. Post 3 expresses an Opinion, *she annoyed the f outta me with her fake knowledge* and “*uuuuuhhhh womens shoes suck*” echoed in Post 4's intra-discussion CS, *j'ai pas vu cette vidéo et j'irai pas la voir parce qu'à chaque fois que je l'entends parler je la trouve hyper fake deep et condescendante, c'est un truc de ouf*.

VIII. CONCLUSION

We have undertaken a comprehensive study of multilingual discussions, specifically the phenomenon of code-switching, via the development of specialized data collection techniques that discover and track multilingual users of social media. To facilitate discussion analysis we developed a machine learning model based on ground truth obtained through Amazon Turk. Our approach goes beyond assessment of social media posts in isolation to assess them within the context of the discussion they appear in. Our approach can readily identify discussions of interest from which we provided commentary on a small selection. We show, what may be relevant to those who study indicators of the formation and dissolution of on line communities, that language holds artifacts, sociolinguistic constructs

```

1 @user2 [TEXT UNAVAILABLE]
2 @user1 @user2 Non mais cherche pas cette meuf c'est une DC stan problématique qui
   crache sur Marvel pour tout et n'importe quoi...
3 @user2 @user1 Ah d'accord, nice person

```

Fig. 5. Post 3 supports post 2.

```

1 @user2 [TEXT UNAVAILABLE]
2 @user1 @user2 tu sors demain soir
3 @user2 @user1 mddrrrrrrr yees si je vais mieux pq pas et toi ?

```

Fig. 6. Post 3 questions post 2 with texting slang (exaggerated for enthusiasm).

```

4 @user1 @user2 Abus de position dominante j'avais compris. C'est l'autre aspect.
   Humm ouais en même temps en terme de moteur de recherche c'est chaud de
   rivaliser avec Google
5 @user2 @user1 C'est chaud mais gars, on ne doit pas leur imposer de se mettre
   en retrait. La vie c'est le combat
6 @user1 @user2 I agree
7 @user2 @user1 ... ne pas se mettre en avant Le système est " gratuit " ça
   devrait être suffisant.

```

Fig. 7. Post 6 supports post 5 (previous posts omitted).

```

1 @user1 Si vous voulez tenter de gagner vos places pour le concert de @user2 à
   Paris c'est par ici !!
2 @user3 @user1 IL A FAIT UNE CHANSON AVEC NINA
3 @user1 @user3 Ooh j'ai jamais écouté ???
4 @user3 @user1 la chanson ou nina ??? : ' ( la chanson c somebody !! et nina
   j'l'aime trop, elle a sorti un album y a pas hyper longtemps ( the sun
   will come up, the seasons will change ) !!
5 @user1 @user3 Les deux ! Du coup j'irai écouter ça
6 @user3 @user1 ouiiiiiii trop bien &lt; 3333

```

Fig. 8. Example texting slang modified for emphasis.

incorporating code choice, which are associated with and may reflect the strength of social bonds among multilinguals as well as its relative increase and decrease with respect to specific situations and over time.

ACKNOWLEDGMENT

We thank the Amazon Turk workers for their contribution to the ground truth.

REFERENCES

- Peter Auer. 1988. A conversation analytic approach to code-switching and transfer. *Codeswitching: Anthropological and sociolinguistic perspectives*, 48:187–213.
- Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.
- Ashutosh Baheti, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. *Proceedings of ICON*, pages 65–74.
- Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. 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.
- Leonard E Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- David Crystal. *The Cambridge encyclopedia of language*, volume 1. Cambridge University Press Cambridge.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. 2018. Simple features for strong performance on named entity recognition in code-switched twitter data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–109.

