Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter?

Koustav Rudra
IIT Kharagpur, India
koustav.rudra@cse.iitkgp.ernet.in

Shruti Rijhwani*
Carnegie Mellon University,
Pittsburgh, Pennsylvania
srijhwan@cs.cmu.edu

Rafiya Begum
Microsoft Research Labs,
Bangalore, India
t-rafbeg@microsoft.com

Kalika Bali
Microsoft Research Labs,
Bangalore, India
kalikab@microsoft.com

Monojit Choudhury
Microsoft Research Labs,
Bangalore, India
monojitc@microsoft.com

Niloy Ganguly
IIT Kharagpur, India
niloy@cse.iitkgp.ernet.in

Abstract

Linguistic research on multilingual societies has indicated that there is usually a preferred language for expression of emotion and sentiment (Dewaele, 2010). Paucity of data has limited such studies to participant interviews and speech transcriptions from small groups of speakers. In this paper, we report a study on 430,000 unique tweets from Indian users, specifically Hindi-English bilinguals, to understand the language of preference, if any, for expressing opinion and sentiment. To this end, we develop classifiers for opinion detection in these languages, and further classifying opinionated tweets into positive, negative and neutral sentiments. Our study indicates that Hindi (i.e., the native language) is preferred over English for expression of negative opinion and swearing. As an aside, we explore some common pragmatic functions of code-switching through sentiment detection.

1 Introduction

The pattern of language use in a multilingual society is a complex interplay of socio-linguistic, discursive and pragmatic factors. Sometimes speakers have a preference for a particular language for certain conversational and discourse settings; on other occasions, there is fluid alteration between two or more languages in a single conversation, also known as *Code-switching* (CS) or *Code-mixing*[1]. Understanding and characterizing language preference in multilingual societies has been the subject matter of linguistic inquiry for over half a century (see Milroy and Muysken (1995) for an overview).

Conversational phenomena such as CS were observed only in speech and therefore, all previous studies are based on data collected from a small set of speakers or from interviews. With the growing popularity of social media, we now have an abundance of conversation-like data that exhibit CS and other speech phenomena, hitherto unseen in text (Bali et al., 2014). Leveraging such data from Twitter, we conduct a large-scale study on language preference, if any, for the expression of opinion and sentiment by Hindi-English (Hi-En) bilinguals.

We first build a corpus of 430,000 unique India-specific tweets across four domains (sports, entertainment, politics and current events) and automatically classify the tweets by their language: English, Hindi and Hi-En CS. We then develop an opinion detector for each language class to further categorize them into opinionated and non-opinionated tweets. Sentiment detectors further classify the opinionated tweets as positive, negative or neutral. Our study shows that there is a strong preference towards Hindi (i.e. the native language or L1) over English (L2) for expression of negative opinion. The effect is clearly visible in CS tweets, where a switch from English to Hindi is often correlated with a switch from a positive to negative sentiment. This is referred to as the *polarity–switch* function of CS (Sanchez, 1983). Using the same experimental technique, we also explore other pragmatic functions of CS, such as *reinforcement* and *narrative–evaluative*.

* This work was done when the author was a Research Fellow at Microsoft Research Lab India.

[1] Although some linguists differentiate between Code-switching and Code-mixing, this paper will use the two terms interchangeably.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1131–1141,
Austin, Texas, November 1-5, 2016.

Apart from being the first large-scale quantitative study of language preference in multilingual societies, this work also has several other contributions: (a) We develop one of the first opinion and sentiment classifiers for Romanized Hindi and CS Hi-En tweets with higher accuracy than the only known previous attempt (Sharma et al., 2015b). (b) We present a novel methodology for automatically detecting pragmatic functions of code-switching through opinion and sentiment detection.

The rest of the paper is organized as follows: Sec. 2 introduces language preference, functions of CS and Hindi-English bilingualism on the web. Sec. 3 formulates the problem and presents the fundamental questions that this paper seeks to answer. Sec. 4 and 5 discuss dataset creation and opinion and sentiment detection techniques respectively. Sec. 6 evaluates the hypotheses in light of the observations on the tweet corpus. We conclude in Sec. 7, and raise some interesting sociolinguistic questions for future studies.

2 Background and Related Work

In order to situate the questions addressed in our work in existing literature, we present a brief overview of the past research in pragmatic and discursive analysis of code-switching, and specifically, on language preference for emotional expression. A primer to Hi-En bilingualism and its presence in social media shall follow.

2.1 CS Functions and Language Preference

In multilingual communities, where there are more than one linguistic channels for information exchange, the choice of the channel depends on a variety of factors, and is usually unpredictable (Auer, 1995). Nevertheless, linguistic studies point out certain frequently-observed patterns. For instance, certain speech activities might be exclusively or more commonly related to a certain language choice (e.g. Fishman (1971) reports use of English for professional purposes and Spanish for informal chat for English-Spanish bilinguals from Puerto Rico). Apart from association between such conversational contexts and language preference, language alteration is often found to be used as a signaling device to imply certain pragmatic functions (Barredo, 1997; Sanchez, 1983; Nishimura, 1995; Maschler, 1991; Maschler, 1994) such as: (a) reported speech (b) narrative to evaluative switch (c) reiterations or emphasis (d) topic shift (e) puns and language play (f) topic/comment structuring etc. Attempts of predicting the preferred language, or even exhaustively listing such functions, have failed. However, linguists agree that language alteration in multilingual communities is not a random process.

Of specific interest to us are the studies on language preference for expression of emotions. Through large-scale interviews and two decades of research, Dewaele (2004; 2010) argued that for most multilinguals, L1 (the dominant language, which is often, but not always, the native or mother tongue) is the language preference for emotions, which include emotional inner speech, swearing and even emotional conversations. Dewaele argues that emotionally charged words in L1 elicit stronger emotions than those in other languages, and hence L1 is preferred for emotion expression.

2.2 Hindi-English Bilingualism

Around 125 million people in India speak English, half of whom have Hindi as their mother-tongue. The large proportion of the remaining half, especially those residing in the metropolitan cities, also know at least some Hindi. This makes Hi-En code-switching, commonly called *Hinglish*, extremely widespread in India. There is historical attestation, as well as recent studies on the growing use of Hinglish in general conversation, and in entertainment and media (see Parshad et al. (2016) and references therein). Several recent studies (Bali et al., 2014; Barman et al., 2014; Solorio et al., 2014; Sequiera et al., 2015) also provide evidence of Hinglish and other instances of CS on online social media such as Twitter and Facebook. In a Facebook dataset analyzed by Bali et al. (2014), almost all sufficiently long conversation threads were found to be multilingual, and as much as 17% of the comments had CS. This study also indicates that on online social media, Hindi is seldom written in the Devanagari script. Instead, loose Roman transliteration, or Romanized Hindi, is common, especially when users code-switch between Hindi and English.

While there has been some effort towards computational processing of CS text (Solorio and Liu, 2008; Solorio and Liu, 2010; Vyas et al., 2014; Peng

et al., 2014), to the best of our knowledge, there has been no study on automatic identification of functional aspects of CS or any large-scale, data-driven study of language preference. The current study adds to the growing repertoire of work on quantitative analysis of social media data for understanding socio-linguistic and pragmatic issues, such as detection of depression (De Choudhury et al., 2013), politeness (Danescu-Niculescu-Mizil et al., 2013), speech acts (Vosoughi and Roy, 2016), and social status (Tchokni et al., 2014).

3 Problem Formulation

Along the lines of (Dewaele, 2010), we ask the following question: *Is there a preferred language for expression of opinion and sentiment by the Hi-En bilinguals on Twitter?*

3.1 Definitions

More formally, let $\Lambda = \{h, e, m\}$ be the set of languages: Hindi (h), English (e) and Mixed (m), i.e., code-switched. Let $\Sigma = \{d, r\}$, be the set of scripts:[2] Devanagari (d) and Roman (r). Let us further introduce a set of sentiments, $\diamondsuit = \{+, -, 0, \otimes\}$, where $+$, $-$ and 0 respectively denote utterances with positive, negative and neutral opinions. $\otimes$ denote non-opinionated (like factual) texts.

Let $T = \{t_1, t_2, \ldots t_{|T|}\}$ be a set of tweets (or any text) generated by Hi-En bilinguals. We define:

- $\lambda(T)$, $\sigma(T)$ and $\diamond(T)$ as the subsets of T that respectively contain all tweets in language λ, script σ and sentiment $\diamond$.
- $\lambda\sigma\diamond(T) = \lambda(T) \cap \sigma(T) \cap \diamond(T)$. Likewise, we also define $\lambda\diamond(T) = \lambda(T) \cap \diamond(T)$, $\lambda\sigma(T) = \lambda(T) \cap \sigma(T)$ and $\sigma\diamond(T) = \sigma(T) \cap \diamond(T)$.

The preference towards a language-script pair $\lambda\sigma$ for expressing a type of sentiment $\diamond$ is given by the probability

$$pr(\lambda\sigma|\diamond; T) = \frac{pr(\diamond|\lambda\sigma; T)pr(\lambda\sigma|T)}{pr(\diamond|T)} \quad (1)$$

However, pr($\lambda\sigma$), which defines the prior probability of choosing $\lambda\sigma$ for a tweet is dependent on a large number of socio-linguistic parameters beyond sentiment. For instance, on social media, English is overwhelmingly more common than any Indic language (Bali et al., 2014). This is because (a) English tweets come from a large number of users apart from Hi-En bilinguals and (b) English is the preferred language for tweeting even for Hi-En bilinguals because it expands the target audience of the tweet by manifolds. The preference of $\lambda\sigma$ for expressing $\diamond$, therefore, can be quantified as:

$$pr(\diamond|\lambda\sigma; T) = \frac{|\lambda\sigma\diamond(T)|}{|\lambda\sigma(T)|} \quad (2)$$

We say $\lambda\sigma$ is the preferred language-script choice over $\lambda'\sigma'$ for expressing sentiment $\diamond$ if and only if

$$pr(\diamond|\lambda\sigma; T) > pr(\diamond|\lambda'\sigma'; T) \quad (3)$$

The strength of the preference is directly proportionate the ratio of the probabilities: $pr(\diamond|\lambda\sigma; T)/pr(\diamond|\lambda'\sigma'; T)$. An alternative but related way of characterizing the preference is through comparing the odds of choosing a sentiment type $\diamond$ to its polar opposite - $\diamond'$. We say, $\lambda\sigma$ is the preferred language-script pair for expressing $\diamond$, if

$$\frac{pr(\diamond|\lambda\sigma; T)}{pr(\diamond'|\lambda\sigma; T)} > \frac{pr(\diamond|\lambda'\sigma'; T)}{pr(\diamond'|\lambda'\sigma'; T)} \quad (4)$$

3.2 Hypotheses

Now we can formally define the two hypotheses, we intend to test here.

Hypothesis I: For Hi-En bilinguals, Hindi is the preferred language for expression of opinion on Twitter. Therefore, we expect

$$pr(\{+, -, 0\}|hd; T) > pr(\{+, -, 0\}|er; T) \quad (5)$$

$$\text{i.e.,} \quad pr(\otimes|hd; T) < pr(\otimes|er; T) \quad (6)$$

And similarly,

$$pr(\otimes|hr; T) < pr(\otimes|er; T) \quad (7)$$

Hypothesis II: For Hi-En bilinguals, Hindi is the preferred language for expression of negative sentiment. Therefore,

$$pr(-|hd; T) \approx pr(-|hr; T) > pr(-|er; T) \quad (8)$$

[2] Tweets in mixed script are rare and hence we do not include a symbol for it, though the framework does not preclude such possibilities.

In particular, we would like to hypothesize that the odds of choosing Hindi for negative over positive is really high compared to the odds for English. I.e.,

$$\frac{pr(-|hd;T)}{pr(+|hd;T)} \approx \frac{pr(-|hr;T)}{pr(+|hr;T)} > \frac{pr(-|er;T)}{pr(+|er;T)} \quad (9)$$

A special case of the above hypotheses arise in the context of code-mixing, i.e., for the set $mr(T)$. Since the mixed tweets certainly come from proficient bilinguals and have both Hi and En fragments, we can reformulate our hypotheses at a tweet level. Let $m^hr(T)$ and $m^er(T)$ respectively denote the set of Hi and En fragments in $mr(T)$.

Hypothesis Ia: Hindi is the preferred language for expression of opinion in Hi-En code-mixed tweets. Therefore, we expect

$$\text{i.e.,} \quad pr(\otimes|m^hr;T) < pr(\otimes|m^er;T) \quad (10)$$

Hypothesis IIa: Hindi is the preferred language for expression of negative sentiment in Hi-En code-switched tweets. Therefore,

$$pr(-|m^hr;T) > pr(-|m^er;T) \quad (11)$$

$$\frac{pr(-|m^hr;T)}{pr(+|m^hr;T)} > \frac{pr(-|m^er;T)}{pr(+|m^er;T)} \quad (12)$$

Likewise, the above hypotheses also apply for the Devanagari script, though for technical reasons, we do not test them here.

Besides comparing aggregate statistics on $mr(T)$, it is also interesting to look at the sentiment of $m^hr(t_i)$ and $m^er(t_i)$ for each tweet t_i. In particular, for every pair of $\diamond \neq \diamond'$, we want to study the fraction of tweets in $mr(T)$ where $m^hr(t_i)$ has sentiment $\diamond$ and $m^er(t_i)$ has $\diamond'$. Let this fraction be $pr(h\diamond \leftrightarrow e\diamond'; mr(T))$. Under "no-preference for language" (i.e., the null) hypothesis, we would expect $pr(h\diamond \leftrightarrow e\diamond'; mr(T)) \approx pr(h\diamond' \leftrightarrow e\diamond; mr(T))$. However, if $pr(h\diamond \leftrightarrow \diamond'; mr(T))$ is significantly higher than $pr(h\diamond' \leftrightarrow e\diamond; mr(T))$, it means that speakers prefer to switch from English to Hindi when they want to express a sentiment $\diamond$ and vice versa.

Pragmatic Functions of Code-Switching: When native speakers tend to switch from Hindi to English when they switch from an expression with sentiment $\diamond$ to one with $\diamond'$, or in other words $\diamond \leftrightarrow \diamond'$, we say this is an observed pragmatic function of code-switching between Hindi and English (note that the order of the languages is important), if and only if

$$\frac{pr(h\diamond \leftrightarrow e\diamond'; mr(T))}{pr(h\diamond' \leftrightarrow e\diamond; mr(T))} > 1 \quad (13)$$

Topic (# tweets): Hashtags
Sports (188K): #IndvsPak, #IndvsUae, #IndvsSa
Movies (82K): #MSG3successfulweeks, #MSGincinemas, #BlockbusterMSG, #Shamitabh, #PK
Politics (92K): #DelhiDecides, #RahulonlLeave, #AAPStorm, #AAPSweep
Current Events (68K): #RailBudget2015, #Beefban, #LandAcquisitionBill, #UnionBudget2015

Table 1: Hashtags used and number of tweets collected

3.3 A Note on Statistical Significance

All the statistics defined here are likelihoods; Equations 9, 12 and 13, in particular, state our hypothesis in the form of the *Likelihood Ratio Test*. However, the true classes λ and $\diamond$ are unknown; we predict the class labels using automatic language and sentiment detection techniques that have non-negligible errors. Under such a situation, the likelihoods cannot be considered as true *test statistics*, and consequently, hypothesis testing cannot be done per se. Nevertheless, we can use these as descriptive statistics and investigate the status of the aforementioned hypotheses.

4 Datasets

We collected tweets with certain India-specific hashtags (Table 1) using the Twitter Search API (Twi, 2015b) over three months (December 2014 – February 2015). In this paper, we use tweets in Devanagari script Hindi (hd), and Roman script English (er), Hindi (hr) and Hi-En Mixed (mr). English and mixed tweets written in Devanagari are extremely rare (Bali et al., 2014) and we do not study them here. We filter out tweets labeled by the Twitter API (Twi, 2015a) as German, Spanish, French, Portuguese, Turkish, and all non-Roman script languages (except Hindi).

We experiment on the following different corpora:

$\boldsymbol{T_{All}}$: All tweets after filtering. This corpus contains 430,000 unique tweets posted by 1,25,396 unique users.

$\boldsymbol{T_{BL}}$: Tweets from users who are certainly Hi-En bilinguals, which are approximately 55% (240,000) of the tweets in $\boldsymbol{T_{All}}$. We define a user to be a Hi-En bilingual if there is at least one mr tweet from the user, or if the user has tweeted at least once in Hindi (hd or hr) and once in English (er).

$\boldsymbol{T_{spo}, T_{mov}, T_{pol}, T_{eve}}$: Topic-wise corpora for sports, movies, politics and events (Table 1).

$\boldsymbol{T_{CS}}$: Tweets with inter-sentential CS. We define these as tweets containing at least one sequence of 5 contiguous Hindi words and one sequence of 5 contiguous English words. The corpus has 3,357 tweets.

SAC: 1000 monolingual tweets (er, hr, hd) and 260 mixed (mr) tweets manually annotated with sentiment and opinion labels. These were annotated by two linguists, both fluent Hi-En speakers. The annotators first checked whether the tweet is opinionated or $\otimes$ and then identified polarity of the opinionated tweets ($+$, $-$ or 0). Thus, the tweets are classified into the four classes in the set $\diamond$. If a tweet contains both opinion and $\otimes$, each fragment was individually annotated. The inter-annotator agreement is 77.5% ($\kappa = 0.59$) for opinion annotation and 68.4% ($\kappa = 0.64$) over all four classes. A third linguist independently corrected the disagreements.

$\mathbf{LLC_{Test}}$: 141 er, 137 hr, and 241 mr tweets annotated by a Hi-En bilingual form the test set for the Language Labeling system (Sec. 5.1).

SAC and $\mathbf{LLC_{Test}}$ can be downloaded and used for research purposes[3].

Note that apart from **SAC** and $\mathbf{LLC_{Test}}$, all corpora are subsets of $\boldsymbol{T_{All}}$. For generalizability of our observations, it is important to ensure that the tweets in $\boldsymbol{T_{All}}$ come from a large number of users and the datasets do not over-represent a small set of users. In Figure 1, we plot the minimum fraction of users required (x-axis) to cover a certain percentage of the tweets in $\boldsymbol{T_{All}}$ (y-axis). Tweets from at least 10%, i.e., 12.5K users are needed to cover 50% of the corpus. As expected, we do observe a power-law-like distribution, where a few users contribute a large number of tweets, and a large number of users contribute a few tweets each. We believe that 12.5K users is sufficient to ensure an unbiased study.

Further, we classify the users into three specific groups (i) news channels, (ii) general users (having $\leq$ 10,000 followers), (iii) popular users or celebrities (having $>$ 10,000 followers). Interestingly, for both $\mathbf{T_{All}}$, and $\mathbf{T_{BL}}$ corpora, we observe that around 98% of all users are general, and 96% of all tweets come from such users. Hence, most observations from these corpora are expected to be representative of the average online linguistic behavior of a Hi-En bilingual.

[3] http://www.cnergres.iitkgp.ac.in/codemixing

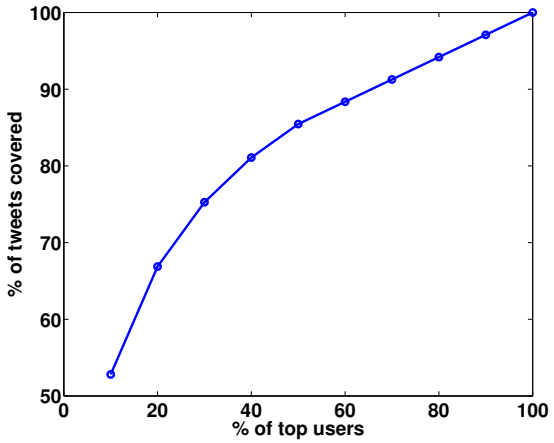


Figure 1: Distribution of cumulative % of tweets and # of users (sorted in descending order by number of tweets).

5 Method

Fig. 2 diagrammatically summarizes our experimental method. We identify the language used in each tweet before detecting opinion and sentiment.

5.1 Language Labeling

Tweets in Devanagari script are accurately detected by the Twitter API as Hindi tweets – we label these as hd, though a small fraction of them could also be md. To classify Roman script tweets as er, hr or mr, we use the system that performed best in the FIRE 2013 shared task for word-level language detection of Hi-En text (Gella et al., 2013). This system uses character n-gram features with a Maximum Entropy model for labeling each input word with a language label (either English or Hindi). We design minor modifications to the system to improve its performance on Twitter data, which are omitted here due to paucity of space.

5.2 Opinion and Sentiment Detection

Most of the existing research in opinion detection (Qadir, 2009; Brun, 2012; Rajkumar et al.,

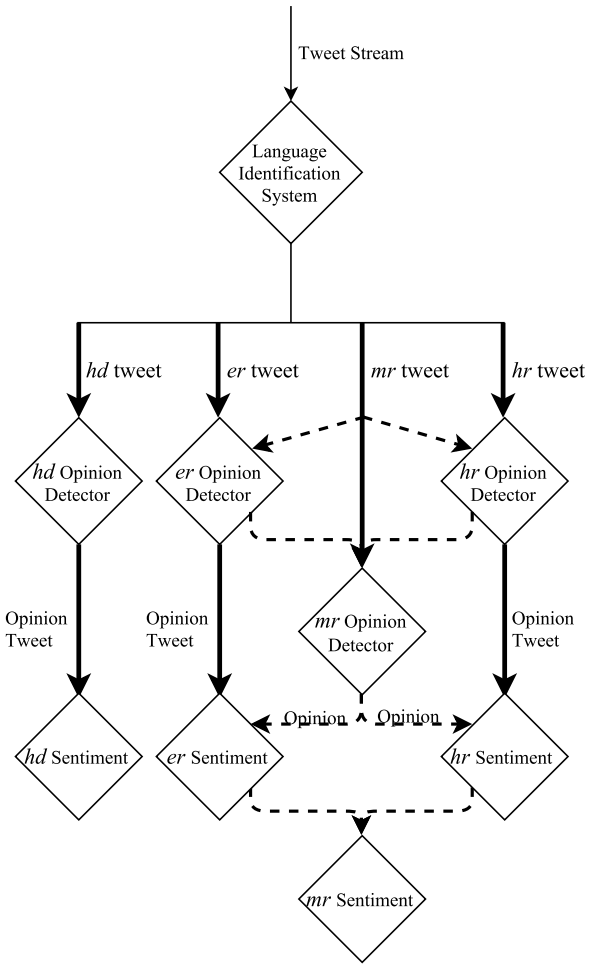


Figure 2: Overview of the experimental method.

2014) and sentiment analysis (Mohammad, 2012; Mohammad et al., 2013; Mittal et al., 2013; Rosenthal et al., 2015) focus on monolingual tweets and sentences. Recently, there has been a couple of studies on sentiment detection of code-switched tweets (Vilares et al., 2015; Sharma et al., 2015b). Sharma et al. (2015b) use Hindi SentiWordNet and normalization techniques to detect sentiment in Hi-En CS tweets.

We propose a two-step classification model. We first identify whether a tweet is opinionated or non-opinionated ($\otimes$). If the tweet is opinionated, we further classify it according to its sentiment ($+$, $-$ or 0). Fig. 2 shows the architecture of the proposed model. Two-step classification was empirically found to be better than a single four-class classifier.

We develop individual classifiers for each language class (er, hr, hd, mr) using an SVM with RBF kernel from Scikit-learn (Pedregosa et al., 2011). We use the **SAC** dataset (Sec. 4) as training data and features as described in Sec. 5.3.

5.3 Classifier Features

For opinion classification (opinion or $\otimes$), we propose a set of event-independent lexical features and Twitter-specific features. (i) **Subjective words:** Expected to be present in opinion tweets. We use lexicons from Volkova et al. (2013) for er and Bakliwal et al. (2012) for hd. We Romanize the hd lexicon for the hr classifiers (ii) **Elongated words:** Words with one character repeated more than two times, e.g. *sooo, naaahhhhi* (iii) **Exclamations:** Presence of contiguous exclamation marks (iv) **Emoticons**[4] (v) **Question marks:** Queries are generally non-opinionated. (vi) **Wh-words:** These are used to form questions (vii) **Modal verbs:** e.g. *should, could, would, cud, shud* (viii) **Excess hashtags:** Presence of more than two hashtags (ix) **Intensifiers:** Generally used to emphasize sentiment, e.g., *we shouldn't get* **too** *comfortable* (x) **Swear words**[5]**:** Prevalent in opinionated tweets, e.g. *that was a* f__ing *no ball!!!! #indvssa* (xi) **Hashtags:** Hashtags might convey user sentiment (Barbosa et al., 2012). We manually identify hashtags in our corpus that represent explicit opinion. (xii) **Domain lexicon:** For hr, & hd category tweets, we construct sentiment lexicons from 1000 manually annotated tweets. Each word or phrase in this lexicon represents $+$, or $-$, or 0 sentiment. (xiii) **Twitter user mentions** (xiv) **Pronouns:** Opinion is often in first person using pronouns like *I* and *we*.

For sentiment classification, we use emoticons, swear words, exclamation marks and elongated words as described above. We also use subjective words from various lexicons (Mohammad and Turney, 2013; Volkova et al., 2013; Bakliwal et al., 2012; Sharma et al., 2015a). Additionally, we use – (i) **Sentiment words:** From Hashtag Sentiment and Sentiment140 lexicons (Mohammad et al., 2013). We also manually annotate hashtags from our dataset that represent sentiment. (ii) **Negation:** A negated context is tweet segment that begins with a negation word and ends with a punctuation mark (Pang et al., 2002). The list of negation words are

[4]The list of emoticons was extracted from Wikipedia

[5]Swear word lexicons from *noswearing.com*, *youswear.com*

Classifier	er	hd	hr	mr
Opinion	72.6	72.0	79.9	73.5
Sentiment	64.4	61.5	62.7	63.4

Table 2: Accuracy of the opinion and sentiment classifiers. All values are in %.

taken from Christopher Potts' sentiment tutorial[6].

The mr opinion classifier uses the output from the er and hr classifiers as features (Fig. 2), along with an additional feature that represents whether the majority of the words in the tweet are Hindi or not. A similar strategy is used for mr sentiment detection.

5.4 Evaluation

We evaluated the language labeling system on the $\boldsymbol{LLC_{Test}}$ corpus, on which the precision (recall) values were $0.93(0.91)$, $0.90(0.85)$ and $0.88(0.92)$ for er, hr and mr classes respectively. The tweet-level classification accuracy was 89.8%.

The opinion and sentiment classifiers were evaluated using 10-fold cross validation on the **SAC** dataset. Table 2 details the class-wise accuracy. For comparison, we also reimplemented the dictionary and dependency-based method by Qadir (2009). The accuracy of the opinion classifier on the er tweets was found to be 65.7%, 7% lower than our system. We also compared our mr sentiment classifier with that of Sharma et al. (2015b). As their method performs two class sentiment detection ($+$ and $-$), we select such tweets from **SAC**. Their system achieves an accuracy of 68.2%, which is 4% lower than the accuracy of our system.

An analysis of the errors showed more false negatives (i.e., opinions labeled $\otimes$) than false positives in opinion classification. Sentiment misclassification is uniformly distributed.

Table 3 reports the accuracy of the opinion classifier for feature ablation experiments. For all three language-script pairs, lexicon and non-word (emoticons, elongated words, hashtags, exclamation) features are the most effective, though all features have some positive contribution towards the final accuracy of opinion detection. For hr and hd tweets, domain knowledge is significant, as shown by the 4% accuracy drop with removing the domain lexicon.

[6]http://sentiment.christopherpotts.net/lingstruc.html

Ablated Feature(s)	er	hr	hd
NONE	72.6	79.9	72.0
mention	70.1	79.3	70.8
lexicon	**68.1**	**75.9**	**66.6**
subjective	69.7	79.8	70.3
wh-words	71.0	79.3	70.1
modal verb	71.1	79.3	71.3
intensifier	71.3	76.6	69.6
slang	70.0	79.2	70.6
pronoun	71.6	79.7	70.3
domain lex.	N.A.	77.0	**67.7**
non-word	**67.7**	**75.6**	68.9

Table 3: Feature ablation experiments for the opinion classifiers. NONE represents the case when all features were used. The two smallest values (pertaining to the two most effective features) are shown in bold.

Corpus	T_{BL}	T_{All}	T_{pol}	T_{mov}
$\|er(T)\|/\|T\|$	0.65	0.79	0.76	0.70
$\|hd(T)\|/\|T\|$	0.12	0.08	0.13	0.04
$\|hr(T)\|/\|T\|$	0.08	0.05	0.05	0.09
$\|mr(T)\|/\|T\|$	0.15	0.08	0.06	0.17

Table 4: Distribution across classes in Λ

6 Experiments and Observations

In this section, we report our experiments on 430,000 unique tweets (T_{All}), and its various subsets as defined in Sec 4. First, we run the language detection system on the corpora. Table 4 shows the language-wise distribution. We see that language preference varies by topic, which is not surprising. Due to paucity of space, the correlation between language usage and topic will not be discussed at length here, but we will highlight cases where the differences are striking.

We apply the language-specific opinion and sentiment classifiers to tweets detected as the corresponding language class. In the following subsections, we empirically investigate the hypotheses.

6.1 Status of Hypotheses I and II

Table 5 shows $pr(\otimes|\lambda\sigma;T)$, $pr(-|\lambda\sigma;T)$ and $pr(-|\lambda\sigma;T)/pr(+|\lambda\sigma;T)$ for T_{All}, T_{BL} and two randomly selected topics – Movie and Politics. The statistics are fairly consistent over the corpora, with slight differences but similar trends in T_{mov}.

Statistic	$\lambda\sigma$	T_{BL}	T_{All}	T_{pol}	T_{mov}
$pr(\otimes\|\lambda\sigma;T)$	er	0.34	0.35	0.37	0.29
	hd	0.45	0.47	0.48	0.49
	hr	0.38	0.39	0.37	0.49
$pr(-\|\lambda\sigma;T)$	er	0.16	0.17	0.22	0.07
	hd	0.18	0.17	0.19	0.16
	hr	0.24	0.25	0.27	0.13
$\frac{pr(-\|\lambda\sigma;T)}{pr(+\|\lambda\sigma;T)}$	er	0.35	0.38	0.59	0.11
	hd	3.00	3.27	5.67	1.90
	hr	1.46	1.60	1.96	0.55

Table 5: Sentiment across languages: Statistics concerning hypotheses I and II.

We need the first statistic in order to investigate **Hypothesis I** (Eqs. 6 and 7), and the two latter ones for verifying **Hypothesis II** (Eqs. 8 and 9).

Contrary to Eqs. 6 and 7, for all corpora except T_{mov}, we observe the following trend:

$$pr(\otimes|hd;T) > pr(\otimes|hr;T) \geq pr(\otimes|er;T)$$

In other words, hd is more commonly preferred for expressing non-opinions than hr and er. Hypothesis I is clearly untrue for these corpora, though due to the small differences between hr and er, we cannot claim that English is the preferred language for expressing opinions. A closer scrutiny of the corpora revealed that hd tweets mostly come from official sources (news channels, political parties, production houses) and celebrities, which are mostly factual. hr tweets are from general users and show similar trends as English. Thus, in general, there seems to be no preferred language for expressing opinion by the Hi-En bilinguals on Twitter.

In the context of **Hypothesis II**, we see the general pattern (with some topic specific variations):

$$pr(-|hr;T) > pr(-|hd;T) \geq pr(-|er;T)$$

The pattern emerges even more strongly, when we look at $pr(-|\lambda\sigma;T)/pr(+|\lambda\sigma;T)$. The odds of expressing a negative opinion over positive opinion in Hindi is between 1.5 and 6 (T_{mov} exhibits a slightly different pattern but similar preference, T_{pol} shows a stronger preference towards Hindi for negative sentiment), whereas the same for English is between 0.1 and 0.6. In other words, English is more preferred for expressing positive opinion, and Hindi for negative opinion. These observations provide very strong evidence in favor of Hypothesis II.

Statistic	m^hr	m^er
$pr(\otimes\|\lambda\sigma;T_{CS})$	0.39	0.45
$pr(-\|\lambda\sigma;T_{CS})$	0.22	0.14
$pr(-\|\lambda\sigma;T_{CS})/pr(+\|\lambda\sigma;T_{CS})$	2.2	0.34

Table 6: T_{CS} statistics for testing hypotheses Ia and IIa

6.2 Status of Hypotheses Ia and IIa

Recall that **Hypothesis Ia** and **Hypothesis IIa** are essentially same as Hypotheses I and II, but applied on m^hr and m^er fragments from the T_{CS} corpus.

Table 6 reports the three statistics necessary for testing these hypotheses. $pr(\otimes|m^er;T_{CS})$ is slightly greater than $pr(\otimes|m^hr;T_{CS})$, which is what we would expect if Hypothesis Ia was true. However, since the difference is small, we view it as a trend rather than a proof of Hypothesis Ia.

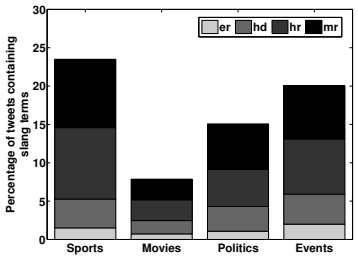
The statistics clearly show that Hypothesis IIa holds true for T_{CS}. The fraction of negative sentiment in m^hr is over 1.5 times higher than that of m^er. Further, the odds of expressing a negative sentiment in Hindi over positive sentiment in Hindi in a code-switched tweet is 6.5 times higher than the same odds for English.

6.3 Switching Functions

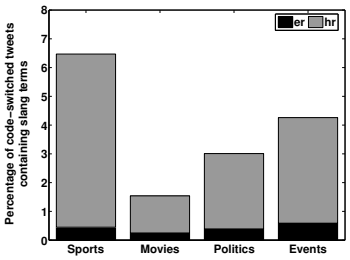
Recall that using Eq. 13 (Sec. 3), we can estimate the preference, if any, for switching to a particular language while changing the sentiment. In particular, research in socio-linguistics has shown that users often switch between languages when they switch from non-opinion ($\otimes$) to opinion ($\{+,-,0\}$). This is called the *Narrative-Evaluative* function of CS (Sanchez, 1983). This function appears in 46.1% of the tweets in $\boldsymbol{T_{CS}}$. We find that

$$\frac{pr(h\{+,-,0\} \leftrightarrow e\otimes;T_{CS})}{pr(h\otimes \leftrightarrow e\{+,-,0\};T_{CS})} = 0.86$$

which indicates that there is no preference for switching to Hindi (or English) while switching between opinion and non-opinion. This is also confirmed above in the context of hypotheses I and Ia. While switching between opinion and non-opinion in a tweet, users do switch language. However, we



(a) Abusive tweets



(b) Swearing pref. in T_{CS}

Figure 3: Distribution of swear words by language

observe no particular preference for the languages chosen for each part.

We also report two other pragmatic functions:

$$\frac{pr(h- \leftrightarrow e\{+,0,\otimes\}; T_{CS})}{pr(h\{+,0,\otimes\} \leftrightarrow e-; T_{CS})} = 1.98$$

$$\frac{pr(h- \leftrightarrow e+; T_{CS})}{pr(h+ \leftrightarrow e-; T_{CS})} = 10.27$$

The latter function is called *polarity switch*. The extremely high value for these ratios is an evidence for a strong preference towards switching language from English to Hindi while switching to negative sentiment (and switching to English when sentiment changes from negative to positive).

We also observe cases where there is a language switch, but no sentiment switch and hence, we cannot evaluate language preference using Eq. 13 (because $\diamond = \diamond'$). In $\boldsymbol{T_{CS}}$, 15.3% of the tweets show *Positive Reinforcement*, where both fragments are of positive sentiment. *Negative Reinforcement* is defined similarly and is seen in 8.7% of the tweets. Other tweets in $\boldsymbol{T_{CS}}$ likely have pragmatic functions that cannot be identified based on sentiment.

6.4 Language Preference for Swearing

Since there is evidence that the native language (Hindi, in this case) is preferred for swearing (Dewaele, 2004), we computed the fraction of tweets that contain swear words in each language class. Fig. 3a shows the distribution across topics. The languages hr and mr have a much higher fraction of abusive tweets than er and hd. Fig. 3b shows the distribution of abusive $m^h r$ and $m^e r$ fragments for tweets in T_{CS}. Interestingly, over 90% of the swear words occur in $m^h r$. Both distributions strongly suggest a preference for swearing in Hindi.

7 Conclusion

In this paper, through a large scale empirical study of nearly half a million tweets, we tried to answer a fundamental question regarding multilingualism, namely, is there a preferred language for expression of sentiment. We also looked at some of the pragmatic functions of code-switching. Our results indicate a strong preference for using Hindi, L1 for the users from whom these tweets come, for expressing negative sentiment, including swearing. However, we do not observe any particular preference towards Hindi for expressing opinions.

Previous linguistic studies (Dewaele, 2004; Dewaele, 2010) have already shown a preference for L1 for expressing emotion and swearing. However, we observe that for expressing positive emotion, English (which would be L2) is the language of preference. This raises some intriguing socio-linguistic questions. Is it the case that English being the language of aspiration in India, it is preferred for positive expression? Or is it because Hindi is specifically preferred for swearing and therefore, is the language of preference for negative emotion? How do such preferences vary across topics, users and other multilingual communities? How representative of the society is this kind of social media study? We plan to explore some of these questions in the future.

Our study also indicates that inferences drawn on multilingual societies by analyzing data in just one language (usually English), which has been the norm so far, are likely to be incorrect.

Acknowledgement

Koustav Rudra was supported by a fellowship from Tata Consultancy Services.

References

Peter Auer. 1995. The pragmatics of code-switching: a sequential approach. In Lesley Milroy and Pieter Muysken, editors, *One speaker, two languages*, pages 115–135. Cambridge University Press.

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon : A lexical resource for hindi polarity classification. In *Proc. LREC*, Austin, Texas, USA, May.

Kalika Bali, Yogarshi Vyas, Jatin Sharma, and Monojit Choudhury. 2014. "i am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. In *Proc. First Workshop on Computational Approaches to Code Switching, EMNLP*.

Glivia A. R. Barbosa, Wagner Meira Jr, Ismael S. Silva, Raquel O. Prates, Mohammed J. Zaki, and Adriano Veloso. 2012. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In *Proc. ACM CHI*, Austin, Texas, USA, May.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *The 1st Workshop on Computational Approaches to Code Switching, EMNLP 2014*.

Inma Muñoa Barredo. 1997. Pragmatic functions of code-switching among Basque-Spanish bilinguals. *Retrieved on October*, 26:528–541.

Caroline Brun. 2012. Learning opinionated patterns for contextual opinion detection. In *COLING (Posters)*, pages 165–174. Citeseer.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *Proceedings of ACL*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Jean-Marc Dewaele. 2004. Blistering barnacles! What language do multilinguals swear in?! *Estudios de Sociolinguistica*, 5:83–105.

Jean-Marc Dewaele. 2010. *Emotions in multiple languages*. Palgrave Macmillan, Basingstoke, UK.

J. A. Fishman. 1971. *Sociolinguistics*. Rowley, Newbury, MA.

Spandana Gella, Jatin Sharma, and Kalika Bali. 2013. Query word labeling and back transliteration for indian languages: Shared task system description.

Yael Maschler. 1991. The language games bilinguals play: language alternation at language boundaries. *Language and communication*, 11(2):263–289.

Yael Maschler. 1994. Appreciation ha'araxa 'o ha'arasta? [valuing or admiration]. *Negotiating contrast in bilingual disagreement talk*, 14(2):207–238.

Lesley Milroy and Pieter Muysken, editors. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. 2013. Sentiment analysis of hindi review based on negation and discourse relation. In *proceedings of International Joint Conference on Natural Language Processing*, pages 45–50.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. 29(3):436–465.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Miwa Nishimura. 1995. A functional analysis of Japanese/English code-switching. *Journal of Pragmatics*, 23(2):157–181.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proc. EMNLP*, pages 79–86.

Rana D. Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is India speaking? Exploring the "Hinglish" invasion. *Physica A*, 449:375–389.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *ACL (2)*, pages 674–679.

Ashequl Qadir. 2009. Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 38–43. Association for Computational Linguistics.

Pujari Rajkumar, Swara Desai, Niloy Ganguly, and Pawan Goyal. 2014. A novel two-stage framework for extracting opinionated sentences from news articles. *TextGraphs-9*, page 25.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of SemEval-2015*.

Rosaura Sanchez. 1983. *Chicano discourse*. Rowley, Newbury House.

Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, and Kunal Chakma. 2015. Overview of fire-2015 shared task on mixed script information retrieval. In *Working Notes of FIRE*, pages 21–27.

Raksha Sharma, Pushpak Bhattacharyya, Ultimate Goal, and Hindi Senti Lexicon Statistics. 2015a. A sentiment analyzer for hindi using hindi senti lexicon.

Shashank Sharma, Pykl Srinivas, and Rakesh Chandra Balabantaray. 2015b. Text normalization of code mix and sentiment analysis. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, pages 1468–1473. IEEE.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2010. Learning to Predict Code-Switching Points. In *Proc. EMNLP*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. *Proceedings of The First Workshop on Computational Approaches to Code Switching, EMNLP*, pages 62–72.

Simo Tchokni, D.O. Séaghdha, and Daniele Quercia. 2014. Emoticons and phrases: Status symbols in social media. In *Eighth International AAAI Conference on Weblogs and Social Media*.

2015a. GET help/languages — Twitter Developers, 8.

2015b. GET search/tweets — Twitter Developers, 8.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodrıguez. 2015. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In *6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. In *Proc. ACL (Vol2: Short Papers)*.

Soroush Vosoughi and Deb Roy. 2016. Tweet acts: A speech act classifier for twitter. In *Tenth International AAAI Conference on Web and Social Media*.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In *Proc. EMNLP*, pages 974–979.