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

Koustav Rudra
koustav.rudra@cse.iitkgp.ernet.in

Shruti Rijhwani and **Rafiya Begum** and **Kalika Bali** and **Monojit Choudhury**
{t-shruri, t-rafbeg, kalikab, monojitc}@microsoft.com

Niloy Ganguly
niloy@cse.iitkgp.ernet.in

Abstract

Linguistic research of multilingual societies has indicated that there is usually a preferred language for expression of emotion and sentiment (Dewaele, 2010). The 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 (a) a language identifier for detecting English, Hindi, and Hindi-English Code-switched tweets and (b) a classifier for opinion detection in these languages and classifying opinionated tweets into positive, negative and neutral. Our study indicates that Hindi (i.e. the native language) is preferred over English for expression of negative opinion and swearing. Narrative-evaluative switching and positive reinforcement are the most common pragmatic functions of code-switching on Twitter.

1 Introduction

More than half of the world's population is multilingual (Grosjean, 2010). The pattern of language use in a multilingual society is a complex interplay of socio-linguistic, discourse 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). 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 the three language classes to further categorize them into opinionated and non-opinionated tweets. Sentiment detectors then classify the opinionated tweets as positive, negative or neutral. Our study shows that there is significant 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 non-opinionated statement to an opinion (especially with negative polarity). This is referred to as the *narrative–evaluative* function of switching (Sanchez, 1983). Using the same experimental technique, we also bring out other functions of CS, such as reinforcement and sarcasm/humor.

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 a language detection system for English, Romanized Hindi and Hi-En CS tweets, which shows an absolute 40% gain on tweet-level language detection accuracy over the current state-of-the-art. (b) We develop one of the first opinion and sentiment classifiers for Romanized Hindi and CS Hi-En tweets with 4% higher accuracy than the only known previous such attempt (Sharma et al., 2015b). (c) We present a novel methodology for automatically detecting certain classes of pragmatic functions of code-switching through opinion and sentiment detection.

The rest of the paper is organized as follows: Section 2 introduces language preference, pragmatic and discourse function studies on multilingualism and code-switching, provides a primer to Hindi-English bilingualism as manifested on Twitter and other online social media platforms, and presents the fundamental questions and hypotheses that the current research seeks to answer. Sections 4, 5.1 and 5.2 discuss the dataset creation, language identification, and opinion and sentiment detection techniques respectively. Section 6 summarizes the observations on the tweet corpus, and evaluates the hypotheses in light of these observations. Finally, we conclude in Section 7 by summarizing our observations and highlighting future directions.

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. Once we have established the context, we shall present the central questions of interest and formulate the hypotheses that we will attempt to verify.

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 a little 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 native 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, 2015), 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 $\mathit{\Lambda} = \{h, e, m\}$ be the set of languages: Hindi (h), English (e) and Mixed (m), i.e., code-switched. Let $\mathit{\Sigma} = \{d, r\}$, be the set of scripts:[1] 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 $T = \{t_1, t_2, \dots t_{|T|}\}$ 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 prefered 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 prefered language for expression of opinion on Twitter. Therefore, we expect

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

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

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

And similarly,

$$pr(\otimes|hr;T) < pr(\otimes|er;T) \tag{7}$$

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

$$pr(-|hd;T) \approx pr(-|hr;T) > pr(-|er;T) \tag{8}$$

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

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

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 comes from proficient bilinguals and have both Hi and En fragments, we can reformulate our hypotheses at a tweet level. Let $m^h r(T)$ and $m^e r(T)$ respectively denote the set of Hi and En fragments in $mr(T)$.

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

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

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

$$pr(-|m^h r;T) > pr(-|m^e r;T) \tag{11}$$

$$\frac{pr(-|m^h r;T)}{pr(+|m^h r;T)} > \frac{pr(-|m^e r;T)}{pr(+|m^e r;T)} \tag{12}$$

Likewise, the above hypotheses also apply for the Devanagari script, though for technical reasons discussed later, we will not be able to test them here.

Instead of comparing aggregate statistics on $mr(T)$, it is also interesting to look at the sentiment of $m^h r(t_i)$ and $m^e r(t_i)$ 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 $m^h r(t_i)$ has sentiment $\diamond$ and $m^e r(t_i)$ 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 Function of Code-switching: We say, 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 $\diamond \leftrightarrow \diamond'$ 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

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

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). Twitter does not identify the language of tweets written in Romanized Indic languages correctly. We use a Hi-Enlanguage identification system (section 5.1) to counter this problem. We experiment on different corpora that are subsets of the tweets collected (after filtering).

$\boldsymbol{T_{All}}$: All tweets after filtering. This corpus contains 0.43M unique tweets.

$\boldsymbol{T_{BL}}$: Tweets from users who are certainly Hi-En bilinguals, which are approximately 55% (0.24M) of the tweets in $\boldsymbol{T_{All}}$. We define a user to certainly be Hi-En bilingual if there is at least one tweet from the user which is mr, or if the user has tweeted at least once in Hindi (hd or hr) and once in English (er).

$\boldsymbol{T_{spo}}, \boldsymbol{T_{mov}}, \boldsymbol{T_{pol}}, \boldsymbol{T_{eve}}$: Corpora containing tweets from different topics – sports, moves, politics, and events respectively (Table 1).

$\boldsymbol{T_{CS}}$: mr tweets that contain inter-sentential CS. We define these as tweets with at least one sequence of 5 contiguous Hindi words and one sequence of

Topic	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

5 contiguous English words. This corpus contains 3,357 tweets.

$\mathbf{SAC}$: 1000 monolingual tweets (er, hr, hd) and 260 mixed (mr) tweets manually annotated with sentiment and opinion labels for training and testing the classifiers. 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). Effectively, the tweets are classified into the four classes in the set $\diamond$. If a tweet contains both opinion and non-opinion parts, 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 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 detection system (section 5.1).

Apart from $\mathbf{SAC}$ and $\mathbf{LLC_{Test}}$, all corpora are subsets of $\boldsymbol{T_{All}}$.

5 Method

Figure 1 diagrammatically summarizes our experimental method. We identify the language used in each tweet before detecting opinion and sentiment.

5.1 Language Detection

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 En or Hi). We design minor modifications to the system to improve its performance on Twitter data, which are omitted here due to paucity of space.

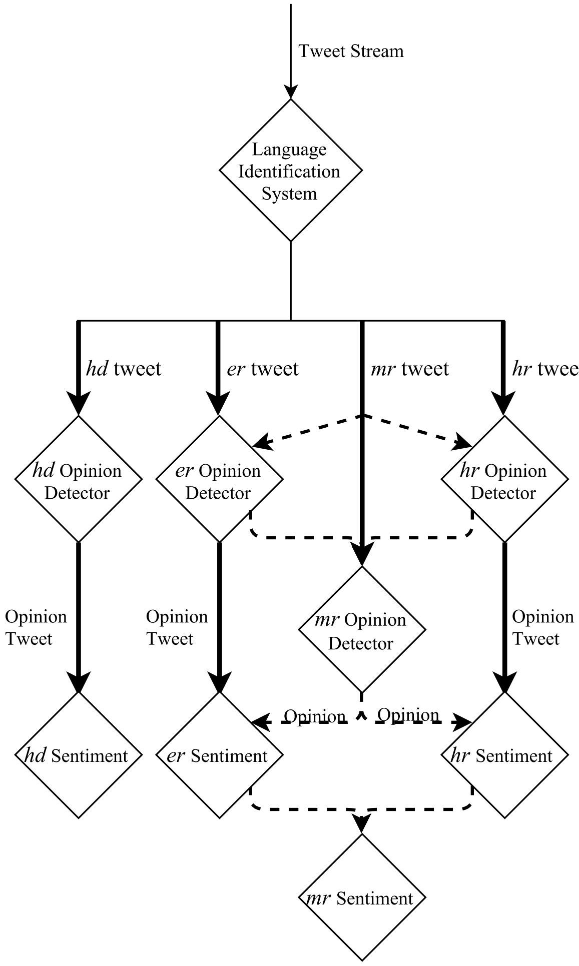


Figure 1: Overview of the experimental method.

5.2 Opinion and Sentiment Detection

Most existing research in opinion detection (Qadir, 2009; Brun, 2012; Rajkumar et al., 2014) and sentiment analysis (Mohammad, 2012; Mohammad et al., 2013; Mittal et al., 2013; Rosenthal et al., 2015) focuses on monolingual tweets and sentences. Recently, there has been interest in 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 two-step classification model. We first identify whether a tweet is opinionated or not ($\otimes$). If the tweet is opinionated, we further classify it according to its sentiment ($+$, $-$ or 0). Fig. 1 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 the Scikit-learn package (Pedregosa et al., 2011). We use the $\mathbf{SAC}$ dataset (Sec. 4) as training data and features as described in Sec. 5.3.

5.3 Classifier Features

We propose a set of event-independent lexical features and Twitter-specific features for opinion classification (opinion or $\otimes$). (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 to use 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**[2] (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**[3]**:** 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) **Twitter user mentions** (xiii) **Pronouns:** Opinion is often in first person using pronouns like *I* and *we*.

For sentiment classification features, 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 taken from Christopher Potts' sentiment tutorial[4].

Classifier	er	hd	hr	mr
Opinion	73.6	72.0	75.6	71.3
Sentiment	64.4	60.2	62.9	62.6

Table 2: Accuracy of the Opinion and Sentiment Classifiers. All values are in %.

As depicted in Fig. 1, the mr opinion classifier uses the output from the er and hr opinion classifiers as features, along with an additional feature that represents whether majority of the words in the tweet are Hindi or not. We use a similar strategy to create the mr sentiment detector.

5.4 Evaluation

We evaluated the language detection 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 word-level labeling accuracy was 89.8%. Importantly, the misclassification was largely homogeneous.

The opinion and sentiment classifiers were evaluated using 10-fold cross validation on the $\boldsymbol{SAC}$ dataset. Table 2 details the class-wise accuracy. For a 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%, 8% less 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 $\boldsymbol{SAC}$. Their system achieves an accuracy of 68.2%, which is 4% lower than the accuracy of our system.

The opinion classification shows more false negatives (i.e., opinions labeled $\otimes$) than false positives.

[2] The list of emoticons was extracted from Wikipedia

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

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

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 3: Distribution across classes in Λ

The sentiment misclassification is uniformly distributed.

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 3 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 and topic will not be discussed at length here. However, the general trend for the class distributions over the sets Λ and $\diamond$ are similar for all the topic restricted corpora, T_{All} and T_{BL}. Therefore, topic specific statistics will be omitted in the following discussions.

We apply the language-specific opinion and sentiment classifiers to tweets detected as the corresponding language class. In the following sections, we discuss our observations and attempt to validate our hypotheses (Sec. 3).

6.1 Testing Hypotheses I and II

Table 4 shows the statistics $pr(\otimes|\lambda\sigma;T)$, $pr(-|\lambda\sigma;T)$ and $pr(-|\lambda\sigma;T)/pr(+|lambda\sigma;T)$ for T_{All} T_{BL} and two randomly selected topics - Movie and Politics. Recall that we need the first statistic to investigate **Hypothesis I** (Eq 6 and 7), and t

Contrary

to Hypothesis I, we observe hr and hd having slightly higher fractions of non-opinionated tweets ($\otimes$) than er, in all the represented corpora. However, the difference is not large enough for statistical inference and we do not see any strong language preference for opinion expression. Therefore, we do not claim to either verify or invalidate Hypothesis I.

Hypothesis II states that Hindi is the preferred language for expressing negative sentiment and is represented by equations (8) and (9). Equation (8) inspects the fraction of negative tweets for er, hd and hr. As Table 4 shows, $pr(-|hd;T)$ and $pr(-|hr;T)$ are greater than $pr(-|er;T)$ in almost all instances, with varying degrees of difference between the fractions. This observation is supported by the validity of equation (9) across all corpora. $pr(-|\lambda\sigma;T)/pr(+|\lambda\sigma;T)$ is distinctly lower for er than hd or hr, indicating the preference for in Hindi while expressing negative sentiment, as compared to positive sentiment. These observations provide very strong evidence for Hypothesis II. The preference for English when tweeting positive sentiment is also apparent.

$\lambda\sigma$	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	0.45	0.47
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	0.15
hr	0.24	0.25	0.27	0.13
$pr(-\|\lambda\sigma;T)/pr(+\|\lambda\sigma;T)$				
er	0.35	0.38	0.59	0.11
hd	1.78	1.71	2.22	1.07
hr	1.46	1.60	1.96	0.55

Table 4: Sentiment across languages: Statistics for testing Hypothesis I and II.

	$m^h r$	$m^e r$
$pr(\otimes\|L;T_{CS})$	0.34	0.46
$pr(-\|L;T_{CS})$	0.38	0.09
$pr(-\|L;T_{CS})/pr(+\|L;T_{CS})$	5.18	0.23
$\otimes' \leftrightarrow \otimes$	0.99	
$- \leftrightarrow +$	6.35	

Table 5: Statistics on T_{CS}

6.2 Testing Hypotheses Ia and IIa

Hypothesis Ia states that tweets with both Hindi and English fragments tend to express opinion in the Hindi fragment ($m^h r$). It is valid if equation

(10) holds on the T_{CS} corpus. In absolute terms, equation (10) is true as the fraction of non-opinion, $pr(\otimes|\lambda\sigma; T_{CS})$, is greater for English fragments (Table 5). However, as with Hypothesis I, the difference is likely statistically insignificant.

Hypothesis IIa states that Hindi is preferred for negative sentiment within CS tweets. This is strongly supported by $pr(-|m^h r; T_{CS})$, the relatively large fraction of Hindi fragments that have negative sentiment (Table 5). Further, $pr(-|\lambda\sigma; T_{CS})/pr(+|\lambda\sigma; T_{CS})$ is 20 times greater for $m^h r$ than for $m^e r$. Both equations (11) and (12) are satisfied, giving strong evidence for the hypothesis that Hindi is the language of choice for negative sentiment in CS tweets.

We identify **Pragmatic functions of code-switching** that are indicated by a *change in sentiment* between the $m^h r$ and $m^e r$ fragments. Using equation (13) (Sec. 3), we evaluate the preference for switching to a particular language while changing the sentiment.

The *Narrative-Evaluative* function occurs when one fragment is opinionated ($\otimes'$) and the other is non-opinionated ($\otimes$). This function appears in 45.5% of the tweets in $\boldsymbol{T_{CS}}$. $\otimes' \leftrightarrow \otimes$ in Table 5 represents the ratio

$$\frac{pr(h\otimes' \leftrightarrow e\otimes; T_{CS})}{pr(h\otimes \leftrightarrow e\otimes'; T_{CS})}$$

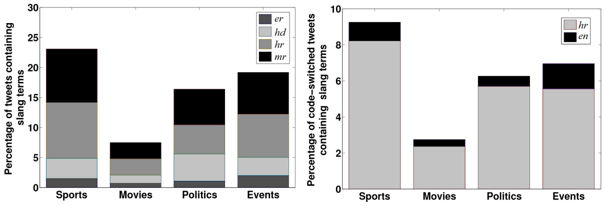
The value is almost 1.0, indicating that there is no preference for switching to Hindi (or English) while switching between opinion and non-opinion.

9.4% of the tweets in $\boldsymbol{T_{CS}}$ show the *Polarity Switch* function, i.e., when one fragment is positive and the other is negative. $- \leftrightarrow +$ in Table 5 represents the ratio

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

The extremely high value for this ratio is evidence for a strong preference of code-switching to Hindi while changing sentiment from positive to negative (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 we cannot evaluate language preference using equation (13). In $\boldsymbol{T_{CS}}$, 18.3% of the tweets show *Positive Reinforcement*, where both fragments are of positive sentiment. *Negative Reinforcement* is defined similarly and is seen in 6.6% of the tweets. The remaining 20.2% tweets in $\boldsymbol{T_{CS}}$ likely have pragmatic functions that cannot be identified based on sentiment.



(a) % Abusive tweets **(b) Swearing pref. in T_{CS}**

Figure 2: Distribution of Abusive Tweets

6.3 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. 2a shows the distribution for T_{spo}, T_{mov}, T_{pol} and T_{eve}. The languages hr and mr have a much higher fraction of abusive tweets than er and hd. Fig. 2b shows the distribution of abusive $m^h r$ and $m^e r$ fragments for tweets in the T_{CS} corpus. Interestingly, over 90% of the swear word occurrences are in $m^h r$. Both distributions strongly suggest a preference for swearing in Hindi (hr and $m^h r$).

7 Conclusion

In this paper, through a purely data-driven approach, we tried to answer a fundamental question regarding multilingualism, namely, is there a preferred language for expression of emotion. We also looked at the various pragmatic functions indicated by code-switching. The entire study has been conducted for Hi-En bilingual users on Twitter. The results indicate a strong preference for using Hindi, which can be safely assumed as L1 for this population, for expressing negative sentiment. This is indicated by the overall opinion distribution, swear word distributions and micro-level analysis of inter-sentential CS tweets.

Previous linguistic studies (Dewaele, 2010; Dewaele, 2004) have already shown a preference for L1 for expressing emotion and swearing. Interestingly, we find that, for expressing positive emotion, English 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? Do such preferences vary across users and other multilingual communities? How representative of the society is this kind of social media study? We would like to systematically explore some of these questions in the future.

Our work 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.

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.

Francois Grosjean. 2010. *Bilingual: Life and Reality*. Harvard University Press.

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, DO 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 WASSA 2015*, page 2.

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)*.

S Vosoughi and D Roy. 2015. Tweet acts: A speech act classifier for twitter. *Submitted to PLoS ONE*.

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.