Corpus Creation and Emotion Prediction for Hindi-English Code-Mixed Social Media Text

Deepanshu Vijay*, **Aditya Bohra***, **Vinay Singh, Syed S. Akhtar, Manish Shrivastava**
International Institute of Information Technology
Hyderabad, Telangana, India
{deepanshu.vijay, aditya.bohra, vinay.singh, syed.akhtar}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Emotion Prediction is a Natural Language Processing (NLP) task dealing with detection and classification of emotions in various monolingual and bilingual texts. While some work has been done on code-mixed social media text and in emotion prediction separately, our work is the first attempt which aims at identifying the emotion associated with Hindi-English code-mixed social media text. In this paper, we analyze the problem of emotion identification in code-mixed content and present a Hindi-English code-mixed corpus extracted from twitter and annotated with the associated emotion. For every tweet in the dataset, we annotate the source language of all the words present, and also the causal language of the expressed emotion. Finally, we propose a supervised classification system which uses various machine learning techniques for detecting the emotion associated with the text using a variety of character level, word level, and lexicon based features.

1 Introduction

Micro-blogging sites like Twitter and Facebook encourage users to express their daily thoughts in real time, which often result in millions of emotional statements being posted online, everyday. Identification and analysis of emotions in social-media texts are of great significance in understanding the trends, reviews, events and human behaviour. Emotion prediction aims to identify fine-grained emotions, i.e., Happy, Anger, Fear, Sadness, Surprise, Disgust, if any present in the text. Previous research related to this task has mainly been focused only on the monolingual text (Chen et al., 2010; Alm et al., 2005) due to the availability of large-scale monolingual resources. However, usage of code mixed language in online posts is very common, especially in multilingual societies like India, for expressing one's emotions and thoughts, particularly when the communication is informal.

Code-Mixing (CM) is a natural phenomenon of embedding linguistic units such as phrases, words or morphemes of one language into an utterance of another (Myers-Scotton, 1993; Muysken, 2000; Duran, 1994; Gysels, 1992). Following are some instances from a Twitter corpus of Hindi-English code-mixed texts also transliterated in English.

T1 : *"I don't want to go to school today, teacher se dar lagta hai mujhe."*
Translation : "I don't want to go to school today, I am afraid of teacher."

T2 : *"Finally India away series jeetne mein successful ho hi gayi :D"*
Translation : "Finally India got success in winning the away series :D"

T3 : *"This is a big surprise that Rahul Gandhi congress ke naye president hain."*
Translation : "This is a big surprise that Rahul Gandhi is the new president of Congress."

The above examples contain both English and Hindi texts. **T1** expresses *fear* through Hindi phrase *"dar lagta hai mujhe"*, *happiness* is expressed in **T2** through a Hindi-English mixed phrase *"jeetne mein successful ho hi gayi"*, while in **T3**, *surprise* is expressed through English phrase *"This is a big surprise"*.

Since very few resources are available for Hindi-English code-mixed text, in this paper we present our initial efforts in constructing the corpus and annotating the code-mixed tweets with associated emotion and the causal language for that emotion. We strongly believe that our initial efforts in constructing the annotated code-mixed emotion corpus will prove to be extremely valuable for researchers working on various natural processing

* These authors contributed equally to this work.

Proceedings of NAACL-HLT 2018: Student Research Workshop, pages 128–135
New Orleans, Louisiana, June 2 - 4, 2018.

tasks on social media.

The structure of the paper is as follows. In Section 2, we review related research in the area of code mixing and emotion prediction. In Section 3, we describe the corpus creation and annotation scheme. In Section 4, we discuss the data statistics. In Section 5, we summarize our classification system which includes the pre-processing steps and construction of feature vector. In Section 6, we present the results of experiments conducted using various character-level, word-level and lexicon features. In the last section, we conclude our paper, followed by future work and the references.

2 Background and Related Work

(Bali et al., 2014) performed analysis of data from Facebook posts generated by English-Hindi bilingual users. They created the corpus using posts from Facebook pages in which En-Hin bilinguals are highly active. They also collected the data from BBC Hindi News page. Their final corpus consisted of 6983 posts and 113,578 words. Among the 6983 posts 206 posts were in Devanagari Script, 6544 posts in Roman Script, 246 in Mixed Scripta and 28 in Other Script. After annotating the data with the Named Entities, POS Tags, Word Origin and deleting all such posts which had less than 5 words, they performed analysis on data. Their analysis showed that atleast 4.2% of the data is code-switched. Analysis depicted that significant amount of code-mixing was present in the posts. (Vyas et al., 2014) formalized the problem, created a POS tag annotated Hindi-English code-mixed corpus and reported the challenges and problems in the Hindi-English code-mixed text. They also performed experiments on language identification, transliteration, normalization and POS tagging of the dataset. Their POS tagger accuracy fell by 14% to 65% without using gold language labels and normalization. Thus, language identification and normalization are critical for POS tagging. (Sharma et al., 2016) addressed the problem of shallow parsing of Hindi-English code-mixed social media text and developed a system for Hindi-English code-mixed text that can identify the language of the words, normalize them to their standard forms, assign them their POS tag and segment into chunks. (Barman et al., 2014) addressed the problem of language identification on Bengali-Hindi-English Facebook comments. They annotated a corpus and achieved an accuracy of 95.76% using statistical models with monolingual dictionaries. (Raghavi et al., 2015) developed a Question Classification system for Hindi-English code-mixed language using word level resources such as language identification, transliteration, and lexical translation.

In addition to information, text also contains some emotional content. (Alm et al., 2005) addressed the problem of text-based emotion prediction in the domain of children's fairy tales using supervised machine learning. (Das and Bandyopadhyay, 2010) deals with the extraction of emotional expressions and tagging of English blog sentences with Ekman's six basic emotion tags and any of the three intensities: low, medium and high. (Xu et al., 2010) built a Chinese emotion lexicon for public use. They adopted a graph-based algorithm which rank words according to a few seed emotion words. (Wang et al., 2016) performed emotion analysis on Chinese-English code-mixed texts using a BAN network. (Joshi et al., 2016; Ghosh et al., 2017) performed Sentiment Identification in Hindi-English code-mixed social media text.

3 Corpus Creation and Annotation

We created the Hindi-English code-mixed corpus using tweets posted online in last 8 years. Tweets were scrapped from Twitter using the Twitter Python API[1] which uses the advanced search option of twitter. We have mined the tweets by selecting certain hashtags from politics, social events, and sports, so that the dataset is not limited to a particular domain. The hashtags used can be found in the appendix section. Tweets retrieved are in the json format which consists all the information such as timestamp, URL, text, user, retweets, replies, full name, id and likes. An extensive semi-automated processing was carried out to remove all the noisy tweets. Noisy tweets are the ones which comprise only of hashtags or urls. Also, tweets in which language other than Hindi or English is used were also considered as noisy and hence removed from the corpus. Furthermore, all those tweets which were written either in pure English or pure Hindi language were removed, and thus, keeping only the code-mixed tweets. In the annotation phase, we further removed all those tweets which were not expressing any emotion.

[1] https://pypi.python.org/pypi/twitterscraper/0.2.7

```
<tweet>
<id>954297321843433472</id>
<word lang="other">@sachin_rt</word>
<word lang="hin">sab</word>
<word lang="hin">cheezo</word>
<word lang="hin">ke</word>
<word lang="hin">bare</word>
<word lang="hin">mai</word>
<word lang="eng">tweet</word>
<word lang="hin">kartey</word>
<word lang="hin">ho</word>
<word lang="hin">toh</word>
<word lang="other">#delhiAirpollution</word>
<word lang="hin">kaise</word>
<word lang="hin">bhol</word>
<word lang="hin">gaye</word>
<word lang="hin">jo</word>
<word lang="eng">national</word>
<word lang="eng">emergency</word>
<word lang="hin">hai,</word>
<word lang="eng">play</word>
<word lang="eng">a</word>
<word lang="eng">fair</word>
<word lang="eng">game</word>
<word lang="hin">sirji</word>
</tweet>
<emotion>       <causal_language>
Sadness         Mixed
</emotion>      </causal_language>
```

Figure 1: Annotated Instance for tweet "*@sachin_rt sab cheezo ke bare main tweet kartey ho toh #delhiAirpollution kaise bhol gaye jo national emergency hai, play a fair game sirji*"

3.1 Annotation

The annotation step was carried out in following two phases:

Language Annotation : For each word, a tag was assigned to its source language. Three kinds of tags namely, 'eng', 'hin' and 'other' were assigned to the words by bilingual speakers. 'eng' tag was assigned to words which are present in English vocabulary, such as "successful", "series" used in **T2**. 'hin' tag was assigned to words which are present in the Hindi vocabulary such as "naye"(new), "hain"(is) used in **T3**. The tag 'other' was given to symbols, emoticons, punctuations, named entities, acronyms, and URLs.

Emotion and Causal Language Annotation : We annotated the tweets with six standard emotions, namely, Happiness, Sadness, Anger, Fear, Disgust and Surprise (Ekman, 1992, 1993). Hindi and English were annotated as the two causal languages. Since emotion in a statement can be expressed through the two languages separately, and also through mixed phrases like: "mujhe fear hai", it is thus essential to annotate the data with four kinds of causal situations (Lee and Wang, 2015), i.e. Hindi, English, Mixed and Both. Next, we further discuss these situations in detail.

Hindi means the emotion of the given post is solely expressed through Hindi text. In the example, **T4** happiness is expressed through Hindi text.

T4 : *"Bahut badiya, ab sab okay hai surgical strike ke baad."*
Translation : "Very good, now everything is okay after the surgical strike."

English means the emotion of the given post is solely expressed through English text. **T5** is an example that expresses surprise through English text.

T5 : *"He is in complete shock, itni property waste ho gayi uski."*
Translation : "He is in complete shock that so much of his property has been wasted."

Both means the emotion of the given tweet is expressed through both Hindi and English text. Since a user can express a kind of emotion using multiple phrases, it is essential to incorporate the case when same emotion is expressed through both the languages. **T6** is an example where sadness is expressed through both Hindi and English texts.

T6 : *"Demonetisation ko Saal hogaye hai..ab toh chod do..these are the people jo har post ko @narendramodi NoteBandi aur Desh ki Sena se jod dete hai.. grow up man..have a life for Gods sake."*
Translation : "It has been one year of Demonetisation. Please Leave it now. These are the people who relates every post with @narendramodi, NoteBandi and Army of this country. Grow up man. Have a life for Gods sake."

Mixed means the emotion of the given tweet is expressed through one or multiple Hindi-English mixed phrases. **T7** is an example which expresses sadness through the mixed phrase 'dekhke sad lagta hai'.

T7 : *"In this country gareeb logo ki haalat dekhke sad lagta hai."*
Translation : "It is sad to see the condition of poor people in this country."

Annotation of this dataset is performed by two of the co-authors who are native Hindi speakers and have proficiency in both Hindi and English. Figure 1 shows an instance of annotation, where both the emotion and the caused language is annotated. In a given tweet, for each emotion, annotator marked whether it expresses that emotion along with it's caused language. The annotated dataset with the classification system is made available online[2].

3.2 Inter Annotator Agreement

Annotation of the dataset to identify emotion in the tweets was carried out by two human annotators having linguistic background and proficiency in both Hindi and English. In order to validate the quality of annotation, we calculated the inter-annotator agreement (IAA) between the two annotation sets of 2866 code-mixed tweets using Cohen's Kappa coefficient. Table 1 shows the results of agreement analysis. We find that the agreement is significantly high. This indicates that the quality of the annotation and presented schema is productive. Furthermore, the agreement of emotion annotation is lower than that of caused language, which probably is due to the fact that in some tweets, emotions are expressed indirectly.

	Cohen Kappa
Emotion	0.902
Caused Language	0.945

Table 1: Inter Annotator Agreement.

4 Data Statistics

We retrieved 3,55,448 tweets from Twitter. After manually filtering the tweets as described in Section 3, we found that only 5546 tweets were code-mixed tweets. Table 2 shows the distribution of data across different emotion categories. Out of 5546 code-mixed tweets, only 2866 tweets were expressing any emotion. The remaining tweets were removed from our dataset, thus keeping only those code-mixed tweets which were expressing any of the six emotions. Also, it is vital to note that some of the tweets contained multiple phrases depicting different emotions. These emotions could be caused by any of the four causal languages. As a result, total number of causal language annotations is more than the number of tweets in the dataset. Usually, a user while posting a tweet feels only one kind of emotion. Hence all such tweets are neglected to avoid any conflict between the literal depiction and the implicit conveyance of emotions in the tweets. This resulted in 2698 emotional code-mixed tweets. Table 3 shows the count of sentences in which emotion was expressed in English, Hindi, Both and Mixed. It clearly shows that in most of the sentences emotion is expressed through a mixed Hindi-English phrase.

Emotion	**Sentences**
Happiness	595
Sadness	878
Anger	667
Fear	85
Disgust	291
Surprise	182
Multiple Emotions	168
Total sentences	2866

Table 2: Data Distribution.

Caused Language	**Sentences**
English	113 (3.7%)
Hindi	1301 (43%)
Mixed	1483 (49%)
Both	127 (4.1%)

Table 3: Caused Language Distribution.

5 System Architecture

After developing the annotated corpus, we try to detect emotion in the code-mixed tweets. We break down the process of emotion detection into three sub-processes: pre-processing of raw tweets, feature identification and extraction and finally, the classification of emotion as happiness, sadness, and anger. It is important to note that classification is carried out only for three classes i.e., 'happiness', 'sadness' and 'anger', as number of tweets which express 'fear', 'disgust' and 'surprise' are extremely limited. The steps have been discussed in sequential order.

[2]https://github.com/deepanshu1995/Emotion-Prediction

5.1 Pre-processing of the code-mixed tweets

Following are the steps which were performed in order to pre-process the data prior to feature extraction.

1. **Removal of URLs:** All the links and URLs in the tweets are stored and replaced with "URL", as these do not contribute towards emotion of the text.

2. **Replacing User Names:** Tweets often contain mentions which are directed towards certain users. We replaced all such mentions with "USER."

3. **Replacing Emoticons :** All the emoticons used in the tweets are replaced with "Emoticon". Before replacing, the emoticons along with their respective counts are stored since we use them as one of the features for classification.

4. **Removal of Punctuations:** All the punctuation marks in a tweet are removed. However, before removing them we store the count of each punctuation mark since we use them as one of the features in classification.

5.2 Feature Identification and Extraction :

In our work, we have used the following feature vectors to train our supervised machine learning model.

1. **Character N-Grams (C):** Character N-Grams are language independent and have proven to be very efficient for classifying text. These are also useful in situations when the text suffers from errors such as misspellings (Cavnar et al., 1994; Huffman, 1995; Lodhi et al., 2002). Groups of characters can help in capturing semantic meaning, especially in the code-mixed language where there is an informal use of words, which vary significantly from the standard Hindi and English words. We use character n-grams as one of the features, where n varies from 1 to 3.

2. **Word N-Grams (W) :** Bag of word features have been widely used to capture emotion in a text (Purver and Battersby, 2012) and in detecting hate speech (Warner and Hirschberg, 2012). Thus we use word n-grams, where n varies from 1 to 3 as a feature to train our classification models.

3. **Emoticons (E) :** We also use emoticons as a feature for emotion classification since they often represent textual portrayals of a writer's emotion in the form of symbols. For example, ':o('and ':(' express sadness, ':)'and ';)' express happiness. We use a list of Western Emoticons from Wikipedia.[3]

4. **Punctuations (P):** Punctuation marks can also be useful for emotion classification. Users often use exclamation marks when they want to express strong feelings. Multiple question marks in the text can denote surprise, excitement, and anger. Usage of an exclamation mark in conjunction with the question mark indicates astonishment and annoyed feeling. We count the occurrence of each punctuation mark in a sentence and use them as a feature.

5. **Repetitive Characters (R) :** Users on social media often repeat some characters in a word to stress upon particular emotion. For example, 'lol' (abbreviated form of laughing out loud) can be written as 'loool', 'looool'. 'Happy' can be written as 'happppyyyy,' 'haaappyy'. We stored the count of all such words in a tweet in which a particular character is repeated more than two times in a row and use them as one of the features.

6. **Uppercase Words (U) :** Users often write some words in a text in capital letters to represent shouting and anger (Dadvar et al., 2013). Hence for every tweet, we count all such words which are completely written in capital letters and contain more than 4 letters and use it as a feature.

7. **Intensifiers (I):** Users often tend to use intensifiers for laying emphasis on sentiment and emotion. For example in the following code-mixed text,
"Wo kisi se baat nahi karega because he is too sad",
Translation : *"He will not talk to anyone because he is too sad"*.
"too" is used to emphasize on the sadness of the boy. A list of English intensifiers was taken from wikipedia[4]. For creating the list of Hindi intensifiers, English intensifiers were

[3] https://en.wikipedia.org/wiki/List_of_emoticons
[4] https://en.wikipedia.org/wiki/Intensifier

Class	Weight
Happiness	4
Sadness	2
Anger	1

Table 4: Weights assigned to classes

transliterated to Hindi. Also Hindi words found in the corpus which are usually used as intensifiers were incorporated in the list. We count the number of intensifiers in a tweet and use the count as a feature.

8. **Negation Words (N) :** We select negation words to address variance from the desired emotion caused by negated phrases like "not sad" or "not happy". For example the tweet "It's diwali today and subah jaldi uthna padega!! Not happy" should be classified as a sad tweet, even though it has a happy unigram. To tackle this problem we define negation as a separate feature. A list of English negation words was taken from Christopher Pott's sentiment tutorial[5]. Hindi negation words were manually selected from the corpus. We count the number of negations in a tweet and use the count as a feature.

9. **Lexicon (L) :** It has been demonstrated in (Mohammad, 2012) that emotion lexicon features provide a significant gain in classification accuracy when combined with corpus-based features, if training and testing sets are drawn from the same domain. We used the (Mohammad and Turney, 2010, 2013) emotion lexicon containing 14182 unigrams both of English and Hindi. The words in Hindi emotion lexicon were written in the Devanagri[6] script and had to be transliterated into Roman Script by the authors. Each word in the lexicon is given a association score of 1 if it is related to a emotion otherwise the association score is 0. A weight was given to each word in a lexicon. The exact weight values are mentioned in the Table 4. This assignment of weight ensured that if a word is related to more than one emotion then we don't lose any information.

[5]http://sentiment.christopherpotts.net/lingstruc.html
[6]https://en.wikipedia.org/wiki/Devanagari

Feature Eliminated	Accuracy
None	58.2
Emoticons	58.1
Char N-Grams	42.9
Word N-Grams	57.6
Repetitive Characters	58.2
Punctuation Marks	57.4
Upper Case Words	58.2
Intensifiers	58.2
Negation Words	58.2
Lexicon	57.9

Table 5: Impact of each feature on the classification accuracy of emotion in the text calculated by eliminating one feature at a time.

6 Results and Discussions

This section presents the results for various feature experimentation.

6.1 Feature Experiments

In order to determine the effect of each feature on classification, we performed several experiments by elimination one feature at a time. In all the experiments, we carried out 10-fold cross-validation. We performed experiments using SVM classifier with radial basis function. The results of the experiments performed after eliminating one feature at a time (i.e., Ablation test to test interaction of feature sets) and using the above-mentioned classifier are mentioned in Table 5. Since the size of feature vectors formed are very large, we applied chi-square feature selection algorithm which reduces the size of our feature vector to 1600[7]. In our system, we have used SVM with RBF kernel as they perform efficiently in case of high dimensional feature vectors. For training our system classifier, we have used Scikit-learn(Pedregosa et al., 2011). The results from Table 5 shows that Character N-Grams, Punctuation Marks, Word N-Grams, Emoticons and Upper Case Words are the features which affect the accuracy most. We were able to achieve the best accuracy of 58.2% using the Character N-Grams, Word N-grams, Punctuation Marks and Emoticons as features trained with SVM classifier.

[7]The size of feature vector was decided after empirical fine tuning

7 Conclusion and Future Work

In this paper, we present a freely available corpus of Hindi-English code-mixed text, consisting of tweet ids and the corresponding annotations. We also present the supervised system used for classifying the emotion of the tweets. The corpus consists of 2866 code-mixed tweets annotated with 6 emotions namely happiness, sadness, anger, surprise and sadness and with the caused language, i.e., English, Hindi, Mixed and Both. The words in the tweets are also annotated with the source language of the words. Experiments clearly show that usage of punctuation marks and emoticons result in better accuracy. Char N-Grams feature vector is also important for classification. As it is clear from the results, in the absence of char n-grams, the classification accuracy drops nearly by 16%. This paper describes the initial efforts in emotion prediction in Hindi-English code-mixed social media texts.

As a part of future work, the corpus can be annotated with part-of-speech tags at word level which may yield better results. Moreover, the dataset contains very limited tweets expressing fear, disgust, and surprise as emotion. Thus it can be extended to include more tweets having these emotions. The annotations and experiments described in this paper can also be carried out for code-mixed texts containing more than two languages from multilingual societies, in future.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. " i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

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.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 179–187. Association for Computational Linguistics.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Dipankar Das and Sivaji Bandyopadhyay. 2010. Identifying emotional expressions, intensities and sentence level emotion tags using a supervised framework. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*.

Luisa Duran. 1994. Toward a better understanding of code switching and interlanguage in bilinguality: Implications for bilingual instruction. *The Journal of Educational Issues of Language Minority Students*, 14(2):69–88.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Souvick Ghosh, Satanu Ghosh, and Dipankar Das. 2017. Sentiment identification in code-mixed social media text. *arXiv preprint arXiv:1707.01184*.

Marjolein Gysels. 1992. French in urban lubumbashi swahili: Codeswitching, borrowing, or both? *Journal of Multilingual & Multicultural Development*, 13(1-2):41–55.

Stephen Huffman. 1995. Acquaintance: Language-independent document categorization by n-grams. Technical report, DEPARTMENT OF DEFENSE FORT GEORGE G MEADE MD.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Sophia Lee and Zhongqing Wang. 2015. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 91–99.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.

Saif Mohammad. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Carol Myers-Scotton. 1993. Dueling languages: Grammatical structure in code-switching. claredon.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. Answer ka type kya he?: Learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*, pages 853–858. ACM.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Srivastava, Radhika Mamidi, and Dipti M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Zhongqing Wang, Yue Zhang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1634.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Ge Xu, Xinfan Meng, and Houfeng Wang. 2010. Build chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1209–1217. Association for Computational Linguistics.

A Appendix

A.1 HashTags

Category	Hash Tags
Politics	#budget, #Trump, #swachhbharat, #makeinindia, #SupremeCourt, #RightToPrivacy, #RahulGandhi, #MannKiBaat, #ManmohanSingh, #MakeInIndia, #SurgicalStrike
Sports	#CWCU19, #U19CWC, #icc, #srt, #pvsindhu, #IndvsSA, #kohli, #dhoni
Social Events	#Festivals, #Holi, #Diwali
Others	#bitcoin, #Jio, #Fraud, #PNBScam, #MoneyLaundering, #Scam

Table 6: List of Hashtags used for mining the tweets.