SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets

Parth Patwa[1*] **Gustavo Aguilar**[2*] **Sudipta Kar**[2] **Suraj Pandey**[3]
Srinivas PYKL[1] **Björn Gambäck**[4] **Tanmoy Chakraborty**[3]
Thamar Solorio[2] **Amitava Das**[5]
[1]IIIT Sri City, India [2]University of Houston, TX [3]IIIT Delhi, India
[4]NTNU, Norway [5]Wipro AI Labs, India
[1]{parthprasad.p17, srinivas.p}@iiits.in
[2]{gaguilaralas, skar3, tsolorio}@uh.edu
[3]{suraj18025, tanmoy}@iiitd.ac.in
[4]gamback@ntnu.no [5]amitava.das2@wipro.com

Abstract

In this paper, we present the results of the SemEval-2020 Task 9 on Sentiment Analysis of Code-Mixed Tweets (SentiMix 2020).[1] We also release and describe our Hinglish (Hindi-English) and Spanglish (Spanish-English) corpora annotated with word-level language identification and sentence-level sentiment labels. These corpora are comprised of 20K and 19K examples, respectively. The sentiment labels are - Positive, Negative, and Neutral. SentiMix attracted 89 submissions in total including 61 teams that participated in the Hinglish contest and 28 submitted systems to the Spanglish competition. The best performance achieved was 75.0% F1 score for Hinglish and 80.6% F1 for Spanglish. We observe that BERT-like models and ensemble methods are the most common and successful approaches among the participants.

1 Introduction

The evolution of social media texts such as blogs, micro-blogs (e.g., Twitter), and chats (e.g., WhatsApp and Facebook messages) has created many new opportunities for information access and language technologies. However, it has also posed many new challenges making it one of the current prime research areas in Natural Language Processing (NLP).

Current language technologies primarily focus on English (Young, 2020), yet social media platforms demand methods that can also process other languages as they are inherently multilingual environments.[2] Besides, multilingual communities around the world regularly express their thoughts in social media employing and alternating different languages in the same utterance. This mixing of languages, also known as code-mixing or code-switching,[3] is a norm in multilingual societies and is one of the many NLP challenges that social media has facilitated.

1.1 Code-Mixing Challenges

In addition to the writing aspects in social media, such as flexible grammar, permissive spelling, arbitrary punctuation, slang, and informal abbreviations (Baldwin et al., 2015; Eisenstein, 2013), code-mixing has introduced a diverse set of linguistic challenges. For instance, multilingual speakers tend to code-mix using a single alphabet regardless of whether the languages involved belong to different writing systems

[*]Equal contribution.
[1]https://ritual-uh.github.io/sentimix2020/

[2]Statistics show that half of the messages on Twitter are in a language other than English (Schroeder, 2010).
[3]We use code-mixing and code-switching interchangeably.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 774–790
Barcelona, Spain (Online), December 12, 2020.

(i.e., language scripts). This behavior is known as transliteration, and code-mixers rely on the phonetic patterns of their writing (i.e., the actual sound) to convey their thoughts in the foreign language (i.e., the language adapted to a new script) (Sitaram et al., 2019). Another common pattern in code-mixing is the alternation of languages at the word level. This behavior often happens by inflecting words from one language with the rules of another language (Solorio and Liu, 2008). For instance, in the second example below, the word *pushes* is the result of conjugating the English verb *push* according to Spanish grammar rules for the present tense in third person (in this case, the inflection *-es*). The Hinglish example shows that phonetic Latin script typing is a popular practice in India, instead of using Devanagari script to write Hindi words. We capture both transliteration and word-level code-mixing inflections in the Hinglish and Spanglish corpora of this competition, respectively.

Aye_{HI} aur_{HI} $enjoy_{EN}$ $kare_{HI}$
Eng. Trans.: come and enjoy

No_{SP} me_{SP} $\boldsymbol{pushes}_{EN}$ $please_{EN}$
Eng. Trans.: Don't push me, please

Considering the previous challenges, code-mixing demands new research methods where the focus goes beyond simply combining monolingual resources to address this linguistic phenomenon. Code-mixing poses difficulties in a variety of language pairs and on multiple tasks along the NLP stack, such as word-level language identification, part-of-speech tagging, dependency parsing, machine translation, and semantic processing (Sitaram et al., 2019). Conventional NLP systems heavily rely on monolingual resources to address code-mixed text, limiting them when properly handling issues such as phonetic typing and word-level code-mixing.

1.2 Code-Mixing as a Global Linguistic Phenomenon

Naturally, code-mixing is more common in geographical regions with a high percentage of bi- or multilingual speakers, such as in Texas and California in the US, Hong Kong and Macao in China, many European and African countries, and the countries in South-East Asia. Multilingualism and code-mixing are also widespread in India, which has more than 400 languages (Eberhard et al., 2020) with about 30 languages having more than 1 million speakers. Language diversity and dialect changes trigger Indians to frequently change and mix languages, particularly in speech and social media contexts. As of 2020, Hindi and Spanish have over 630 million and over 530 million speakers (Eberhard et al., 2020), respectively, ranking them in 3rd and 4th place based on the number of speakers worldwide, which speaks of the relevancy of using these languages in our code-mixing competition.

1.3 SentiMix Overview

This paper provides an overview of the SemEval-2020 Task 9 competition on sentiment analysis of code-mixed social media text (SentiMix). Specifically, we provide code-mixed text annotated with word-level language identification and sentence-level sentiment labels (negative, neutral, and positive). We release our Hinglish (Hindi-English) and Spanglish (Spanish-English) corpora, which are comprised of 20K and 19K tweets, respectively. We describe general statistics of the corpora as well as the baseline for the competition.

We received 61 final submissions for Hinglish and 28 for Spanglish, adding to a total number of 89 submissions. We received 33 system description papers. We provide an overview of the participants' results and describe their methods at a high level. Notably, the majority of these methods employed BERT-like and ensemble models to reach competitive results, with the best performers reaching 75.0% and 80.6% F1 scores for Hinglish and Spanglish on held-out test data, respectively. We hope that this shared task will continue to catch the NLP community's attention on the linguistic code-mixing phenomenon.

2 Related Work

Linguists (Verma, 1976; Bokamba, 1988; Singh, 1985) studied the phenomena of code-mixing and intra-sentential code-switching and found that processing code-mixed language is much more complicated

than monolingual text. Code-mixing is often found on social media which contains a lot of nonstandard spellings of words and unnecessary capitalization (Das and Gambäck, 2014), making the task more difficult. Naturally, the difficulty will increase as the amount of code-mixing increases. To quantify the level of code-switching between languages in a sentence, Gambäck and Das (2016) introduced a measure called Code Mixing Index (CMI) which considers the number of tokens of each language in a sentence and the number of tokens where the language switches.

Finding the sentiment from code-mixed text has been attempted by some researchers. Mohammad et al. (2013) used SVM-based classifiers to detect sentiment in tweets and text messages using semantic information. Bojanowski et al. (2017) proposed a skip-gram based word representation model that classifies the sentiment of tweets and provides an extensive vocabulary list for language. Giatsoglou et al. (2017) trained lexicon-based document vectors, word embedding, and hybrid systems with the polarity of words to classify the sentiment of a tweet. Sharma et al. (2016) attempted shallow parsing of code-mixed data obtained from online social media, and Chittaranjan et al. (2014) tried word-level identification of code-mixed data to classify the sentiment. Some researchers also tried normalizing the text with lexicon lookup for sentiment analysis of code-mixed data (Sharma et al., 2015).

To advance research in code-mixed language processing, few workshops have also been conducted. Four successful series of Mixed Script Information Retrieval have been organized at the Forum for Information Retrieval Evaluation (FIRE) (SahaRoy et al., 2013; Choudhury et al., 2014; Sequiera et al., 2015; Banerjee et al., 2016). Three workshops on Computational Approaches to Linguistic Code-Switching (CALCS) have been conducted which included shared tasks on language identification and Named Entity Recognition (NER) in code-mixed data (Solorio et al., 2014a; Molina et al., 2016; Aguilar et al., 2018). For our SentiMix Spanglish dataset, we adopt the SentiStrength (Vilares et al., 2015) annotation mechanism and conduct the annotation process over the unified corpus from the three CALCS workshops.

3 Task Description

Although code-mixing has received some attention recently, properly annotated data is still scarce. We run a shared task to perform sentiment analysis of code-mixed tweets crawled from social media. Each tweet is classified into one of the three polarity classes - Positive, Negative, Neutral. Each tweet also has word-level language marking. We release two datasets - Spanglish and Hinglish.

We used CodaLab[4,5] to release the datasets and evaluate submissions. Initially, the participants had access only to train and validation data. They could check their system's performance on the validation set on a public leaderboard. Later, a previously unseen test set was released, and the performance on the test set was used to rank the participants. Only the first three submissions on the test set by each participant were considered, to avoid over-fitting on the test set. The ranking was done based on the best out of the three submissions. There was no distinction between constrained and unconstrained systems, but the participants were asked to report what additional resources they have used for each submitted run.

We release 20k labeled tweets for Hinglish and $\approx$ 19k labeled tweets for Spanglish. In both the datasets,[6] in addition to the tweet level sentiment label, each tweet also has a word-level language label. The detailed distribution is provided in Table 1. Some annotated examples are provided in Table 2. Although this task focuses on sentiment analysis, the data has word-level language marking and can be used for other NLP tasks.

3.1 Evaluation Metric

To evaluate the performance and rank the participants, we use weighted F1 score on the test data, across the positives, negatives, and neutral examples.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where,

[4]Hinglish: `https://competitions.codalab.org/competitions/20654`

[5]Spanglish: `https://competitions.codalab.org/competitions/20789`

[6]Both the datasets are available at `https://zenodo.org/record/3974927#.XyxAZCgzZPZ`

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}} \text{ , Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

The F1 scores are calculated for each class and then their average is weighted by support (number of true instances for each class). We use a weighted F1 score since the number of instances per class is not equal. Other than the F1 score, we also calculate precision and recall for each class to analyze and have a better understanding of *false positives* and *false negatives*.

4 Dataset

The datasets consist of tweets labeled into one of the three classes:

- **Positive (Pos)**: Tweets which express happiness, praise a person, group, country or a product, or applaud something. Hinglish example: *"bholy bhayaa. Uffffff dil jeet liya ap ne. Love you imran bhai. Mind blowing ap ki acting hai."* (bholy bhayaa, you won hearts. love you imran bhai your acting is mind blowing). Spanglish example: *"We all here waiting pa ke juege mex :)"* (We all here waiting for Mexico to play :)).

- **Negative (Neg)**: Tweets which attack a person, group, product or country, express disgust or unhappiness towards something, or criticize something. Hinglish example: *"You efficiency of anchoring a program is continuously deteriorating. Ab to dekhne ki himmat hi nahi"* (Your efficiency of anchoring is continuously deteriorating. Now can't even dare to watch it) Spanglish example: *"Eres una cualkiera yes u are."* (You are a tramp, yes you are.)

- **Neutral (Neu)**: Tweets which state facts, give news or are advertisements. In general those which don't fall into the above 2 categories. Hinglish example: *"Nahi wo is news ko defend kerne ki koshesh ker rhe hain h"* (No, they are trying to defend this news). Spanglish example: *"My phone looks ratchet todo crack"* (My phone looks ratchet all crack).

Language	Split	Total	Positive	Neutral	Negative
Hinglish	Train	14,000	4,634 (33.10%)	5,264 (37.60%)	4102 (29.30%)
	Validation	3,000	982 (32.73%)	1,128 (37.60%)	890 (29.67%)
	Test	3,000	1,000 (33.33%)	1,100 (36.67%)	900 (30%)
	Total	20,000	6,616 (33.08%)	7,492 (37.46%)	5892 (29.46%)
Spanglish	Train	12,002	6,005 (50.03%)	3,974 (33.11%)	2,023 (16.85%)
	Validation	2,998	1,498 (49.96%)	994 (33.15%)	506 (16.87%)
	Test	3,789	3,061 (80.78%)	206 (5.43%)	522 (13.77%)
	Total	18,789	10,564 (56.22%)	5,174 (27.53%)	3,051 (16.23%)

Table 1: Class-wise statistics of the dataset for train, validation, and test set. We put special care to make a balanced class-wise distribution for Hinglish.

Both the Hinglish and Spanglish datasets are released using the previous sentiment label scheme. However, each dataset has been annotated separately as the studies were independent before the organization of this competition. We provide the data collection and annotation details in the following subsections.

4.1 Hinglish

Data Collection: First, we make a list of all the Hindi tokens from the dataset provided by (Patra et al., 2018). From that list, we remove those tokens which are common to Hindi and English (example '*the*' can be used in both the languages). Then we use Twitter API [7] to crawl those tweets from twitter which have at least one word from the list. The list has 10786 tokens. Some words from the list are: *kuch, tu, gaya, raha, aaj, apne, tum, gaye, sath* etc.

Language and Sentiment Annotation: For word-level language marking we use an automated tool released by Bhat et al. (2014). The tokens are labeled into HIN - Hindi, ENG - English, or O - other. For tweet level sentiment labels, we took the help of around 60 annotators who were bilingual/multilingual, proficient in Hindi and had Hindi as their first or second language. Each tweet was shown to two annotators, and it was selected if their annotations matched, else it was discarded. They used a simple website designed for this purpose to annotate the data. Each tweet was shown on a page that had a radio button for each label. The annotators first had to enter their unique id, then they could either select a sentiment option for a tweet and send or choose to skip the tweet.

Statistics: Table 1 gives detailed class-wise distribution of the tweets. Although Neutral is the majority class for Hinglish, the dataset is not too imbalanced. The class-wise distribution is similar for all three splits. Table 2 shows some examples of tweets marked with language and sentiment tags. The average CMI for Hinglish train, validation, and test set is 25.32, 25.53, and 25.13 respectively. The inter-annotator agreement is 55%.

4.2 Spanglish

Data Collection: We use the Spanish-English data from the CALCS workshops (Solorio et al., 2014b; Molina et al., 2016; Aguilar et al., 2018). In the first workshop (Solorio et al., 2014b), the data was collected by crawling tweets from specific locations with a strong presence of Spanish and English speakers (e.g., California and Texas). The collection process was conducted using common words from each language through the Twitter API.[7] In the second workshop (Molina et al., 2016), the organizers provided a new test set collected with a more elaborated process. They selected big cities where bilingual speakers are common (e.g., New York and Miami). Then, they localized Spanish radio stations that showed code-mixed tweets. Such radio stations led to users that also practice code-mixing. Similar to the third workshop (Aguilar et al., 2018), we take the CALCS data and extend it for sentiment analysis. It is worth noting that a large number of tweets in the corpora only contain monolingual text (i.e., no code-mixing). Considering that, and after merging the two corpora, we prioritize the tweets that show code-mixed text to build the SentiMix corpus. We ended up incorporating 280 monolingual tweets per language (English, Spanish) in the test set.

Annotation: Since we use the data from the previous CALCS workshops, we did not need to undergo the token-level annotation process for language identification (LID). We adopted the CALCS LID label scheme, which is comprised of the following eight classes: `lang1` (English), `lang2` (Spanish), `mixed` (partially in both languages), `ambiguous` (either one or the other language), `fw` (a language different than `lang1` and `lang2`), `ne` (named entities), `other`, and `unk` (unrecognizable words).

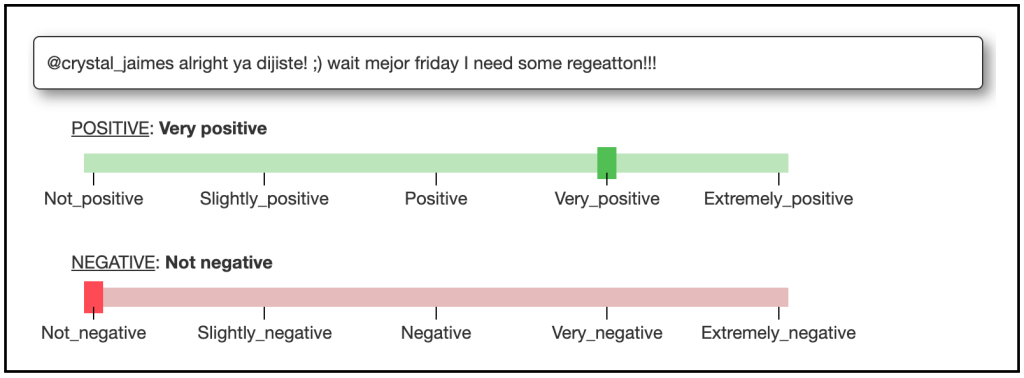

Figure 1: A screenshot of the Spanglish annotation interface.

For the annotations of the sentiment labels, we follow the SentiStrength[8] strategy (Thelwall et al., 2010; Vilares et al., 2015). That is, we provide positive and negative sliders to the annotators. Each

[7] `https://developer.twitter.com/`
[8] `http://sentistrength.wlv.ac.uk/`

slider denotes the strength for the corresponding sentiment, and the annotators can choose the level of the sentiment they perceived from the text (see Figure 1). The range of the sliders is discrete and included strengths from 1 to 5 with 1 being *no strength* (i.e., no positive or negative sentiment) and 5 the strongest level. Using two independent sliders allowed the annotators to process the positive and negative signals without excluding one from the other, letting them provide mixed sentiments for the given text (Berrios et al., 2015). Once the sentiment strengths were specified, we converted them into a 3-way sentiment scale (i.e., `positive`, `negative`, and `neutral`). We simply subtract the negative strength from the positive strength, and mark the text as `positive` if the result was greater than zero, `negative` if less than zero, or `neutral` otherwise.

We annotate each tweet with the help of three annotators from Amazon Mechanical Turk.[9] We regulate the annotations by using quality questions within every assignment[10] of a HIT (Human Intelligence Task). Every assignment has ten tweets, two of them were for quality control (i.e., the annotation was already known) and the other eight tweets were the ones to annotate.[11] The annotators had to have at least one quality control tweet right so that the assignment (i.e., the ten tweets) was not automatically rejected. Since the sentiment analysis task is arguably arbitrary, we provided multiple valid levels of strength for the quality control tweets. If an assignment was rejected, then another annotator was automatically required to complete the HIT until three annotations were accepted. Also, we automatically approved HITs if their 3-way sentiment inter-annotator agreement was over 66%.[12] Otherwise, we evaluated manually the annotations and decide whether to extend the assignments or mark the sentiment labels ourselves for the trivial cases. After merging the annotations, we gave a pass over the data and manually corrected annotations that were unambiguously wrong.

Statistics: The Spanglish class-level distribution of the partitions appear in Table 1. Notably, the data is highly imbalanced towards the positive class covering about 56% in the entire Spanglish corpus, while the negative and neutral classes account for around 16% and 27%, respectively. The reason for this imbalance distribution is that we did not collect the data following a sentiment-oriented crawling strategy (e.g., searching by sentiment-related keywords). Instead, we just extended the original corpus, which happens to be mostly positive. The intention to proceed in this way is to enrich the original corpus annotations with sentiment-level labels. Moreover, the splits do not share the same distribution (i.e., development and test are more skewed than training) because we were annotating data on-demand rather than having available the entire corpus at any stage of the competition. Some annotated examples are provided in Table 2. The average CMI for the train, validation, and test sets are 21.84, 20.52, and 17.23, respectively.

5 Baseline

We develop our baseline system using the pre-trained *multilingual* BERT (M-BERT; Devlin et al. (2019)). M-BERT was trained on 104 languages' entire Wikipedia dump and the WordPiece (Wu et al., 2016) vocabulary of this model contains 110K sub-word tokens from these 104 languages. To balance the risk of low-resource languages being under-represented or over-fitted due to small training resources during pre-training, exponentially smoothed weighting was performed on the data during pre-training data creation and vocabulary creation. Although M-BERT was trained on monolingual data from different languages, it is capable of multilingual generalization in code-switching scenarios (Pires et al., 2019).

We use the Transformers (Wolf et al., 2019) library to implement our framework and we fine-tune the pre-trained `BERT-Base, Multilingual Cased` model separately for each of the two languages. Based on our observation on the training split for each dataset, we set the highest sequence length to 40 and 56 tokens for Spanglish and Hinglish, respectively. Then, we fine-tune the model for three epochs using AdamW (Loshchilov and Hutter, 2019) optimizer ($\eta = 2e^{-5}$).

[9] `https://requester.mturk.com/`

[10] An assignment is done by a single annotator.

[11] We use the assignment review policy `ScoreMyKnownAnswers/2011-09-01`.

[12] We use the HIT review policy `SimplePlurality/2011-09-01`.

Language	Tweet	Class
Spanglish	@username$_{other}$ ha$_{lang2}$ pos$_{ambiguous}$ have$_{lang1}$ fun$_{lang1}$ its$_{lang1}$ pretty$_{lang1}$ te$_{lang2}$ subes$_{lang2}$ al$_{lang2}$ horse$_{lang1}$ its$_{lang1}$ cute$_{lang1}$ lol$_{lang1}$ (@username ah, then have fun, it's pretty, you ride the horse, it's cute lol)	Positive
Spanglish	Cuando$_{lang2}$ Mis$_{lang2}$ parents$_{lang1}$ me$_{lang2}$ dejan$_{lang2}$ ir$_{lang2}$ el$_{lang2}$ date$_{lang1}$ me$_{ambiguous}$ Keda$_{lang2}$ Mal$_{lang2}$ /$_{other}$.$_{other}$ -$_{other}$ No$_{lang2}$ MAMEN$_{lang2}$ (When my parents let me go, my date is cancelled / . - You're kidding me!)	Negative
Spanglish	Tengo$_{lang2}$ hungry$_{lang1}$ mhm$_{unk}$ (I'm hungry mhm)	Neutral
Hinglish	Congratulations$_{ENG}$ Sir$_{ENG}$ we$_{ENG}$ proud$_{ENG}$ of$_{ENG}$ you$_{ENG}$ ..$_{O}$ Aap$_{HIN}$ pr$_{HIN}$ pura$_{HIN}$ jakeen$_{HIN}$ hai$_{HIN}$ ..$_{O}$ aap$_{HIN}$ bohat$_{HIN}$ achaa$_{HIN}$n home$_{HIN}$ minister$_{ENG}$ Honga$_{HIN}$ ..$_{O}$)$_{O}$ (Congratulations sir we are proud of you.. We believe in you.. You will be a very good home minister..)	Positive
Hinglsih	Hostelite$_{ENG}$ k$_{ENG}$ naam$_{HIN}$ pe$_{HIN}$ dhabba$_{HIN}$ ho$_{HIN}$ tum$_{HIN}$ (you are a blot on the name of a hostelite)	Negative
Hinglish	Warm$_{ENG}$ up$_{ENG}$ match$_{ENG}$ to$_{ENG}$ theek$_{HIN}$ thaak$_{HIN}$ chal$_{HIN}$ ra$_{HIN}$ hai$_{HIN}$ (Warm up match is going fine)	Neutral

Table 2: Examples of labeled tweets. Code-mixing often refers to the juxtaposition of linguistic units from two or more languages in a single conversation or sometimes even a single utterance. These examples emphasize on the fact that people don't do only phrase, or tag-mixing as it was a belief in the linguistic forum until now.

6 Participation and Top Performing Systems

We received an overwhelming response for both Hinglish and Spanglish. 61 teams submitted their systems for Hinglish and 28 teams submitted their systems for Spanglish. 16 teams submitted to both Hinglish and Spanglish. We received 33 system description papers in total. The embeddings and techniques used by the participants are tabulated in Table 5. The team names, Codalab names, and their corresponding description papers are provided in Appendix (Table 6). We provide a summary of the top teams below (Codalab usernames are mentioned in parentheses) :

Top Hinglish Systems @ SentiMix

- **KK2018 (kk2018)** used pre-trained XLM-R(Conneau et al., 2019) which was trained with 100 languages. They trained it with adversarial (intentionally designed to make model cause a mistake) examples. To create adversarial examples, they used the formula proposed by (Miyato et al., 2016) where the perturbation is created using the gradient of the loss function.
- **MSR India (genius1237)** used embeddings from XLM-R as inputs to a classification layer. They also do so with multiligual BERT.
- **Reed (gopalvinay)** Finetuned BERT and claimed that pre-training of BERT is not of much use. They also tried bag-of-words based feedforward networks.
- **BAKSA (ayushk)** used XLM-R(Conneau et al., 2019) multilingual embeddings (a transformer-based masked language model trained on 100 languages) followed by ensemble model of CNN and self attention architecture.

Top Spanglish Systems @ SentiMix

- **XLP (LiangZhao)** augmented the data using machine translation. Then they used pre-trained embeddings made by Facebook Research (XLMs)(Lample and Conneau, 2019) followed by CNN

classifier of linear classifier (fully connected layer). They optimized a weighted loss function based on the complexity of code-mixing.

- **Voice@SRIB (asking28)** applied multiple pre-processing steps and used Ensemble model by combining CNN, self-attention and LSTM based model.
- **Palomino-Ochoa (dpalominop)** combined a transfer learning scheme based on ULMFit (Howard and Ruder, 2018) with the-state-of-the-art language model BERT.
- **HPCC-YNU (kongjun)** used word and character embeddings as input to BiLSTM with attention.

7 Results and Analysis

Table 3 and Table 4 show top 15 participants [13] for Hinglish and Spanglish respectively. For Hinglish the top 15 participants lie between 75% and 68.6% F1 score. The participants in the middle of the table are quite close to each other. 44 participants beat the baseline whereas 17 could not. For Spanglish, the top 15 F1 scores lie between 80.6% and 71.0% and most are in mid 70s. 22 teams were able to beat the baseline whereas 6 could not. The results are much better for positive than for other two classes due to the data imbalance.

Rank	System	Positive			Neutral			Negative			Avg.
		P	R	F1	P	R	F1	P	R	F1	F1
1	KK2018	**84.3**	76.0	**79.9**	65.2	73.1	**68.9**	**78.5**	75.4	**76.9**	**75.0**
2	Genius1237	81.0	**77.8**	79.3	**65.7**	64.3	65.0	72.0	**77.0**	74.4	72.6
3	olenet	78.2	74.4	76.2	62.8	65.3	64.0	75.2	75.7	75.5	71.5
4	gopalanvinay	80.7	74.6	77.5	61.4	67.5	64.3	74.5	71.6	73.0	71.3
5	ayushk	78.8	73.8	76.2	60.9	67.5	64.0	75.3	70.6	72.9	70.7
6	Taha	78.6	72.8	75.6	60.6	70.1	65.0	76.2	67.9	71.8	70.6
7	Miriam	78.0	77.3	77.6	62.6	60.1	61.3	70.7	74.9	72.7	70.2
8	HugoLerogeron	79.2	74.7	76.9	60.3	63.9	62.1	70.6	70.0	70.3	69.5
9	somban	78.6	72.9	75.6	59.4	65.0	62.1	71.8	69.3	70.5	69.1
10	aditya_malte	80.3	69.0	74.2	57.0	**73.5**	64.2	77.3	62.2	69.0	69.0
11	MeisterMorxrc	79.9	70.1	74.7	59.5	65.0	62.1	70.2	71.9	71.0	69.0
12	nirantk	78.9	70.8	74.6	58.3	67.4	62.5	73.2	67.6	70.2	68.9
13	apurva19	78.8	75.8	77.3	61.2	60.8	61.0	67.4	70.8	69.1	68.8
14	c1pher	79.7	69.7	74.4	56.5	**73.5**	63.9	78.3	60.7	68.4	68.7
15	will_go	77.2	70.5	73.7	57.8	70.2	63.4	75.9	63.4	69.1	68.6
45	**Baseline**	72.8	68.8	70.7	56.2	60.2	58.1	69.1	67.4	68.3	65.4

Table 3: Top 15 Results for the **Hinglish** dataset. The systems are ordered by the *Weighted Average F1* (Avg.) scores of the *Postive*, *Neutral*, and *Negative* classes. We report Precision (P), Recall (R), and F1 score for each class separately. In each column, the boldfaced scores are the highest score in that column.

In the previous section, we briefly described the top systems. Here, we group and summarize various techniques used by the systems (Codalab usernames are mentioned in parentheses) :

- **Word Embedding:** Three popular word embedding ways explored by participants. Word2Vec, Glove, FastText. Some participants used character-embedding. Additional resources were also used by participants to train their own embeddings.
- **Classical ML methods:** Classical ML techniques like - logistic regression, Naive Bayes, Perceptron, and SVM have been tested by several researchers. Naive Bayes and its multinomial kernel

[13] Results for all the participants are available at `https://ritual-uh.github.io/sentimix2020/`

Rank	System	Positive			Neutral			Negative			Avg.
		P	R	F1	P	R	F1	P	R	F1	F1
1	LiangZhao	88.3	92.6	**90.4**	18.1	20.9	19.4	59.9	39.5	47.6	**80.6**
2	rachel	89.0	87.7	88.3	16.0	45.1	23.7	**65.3**	24.5	35.7	77.6
3	asking28	84.5	90.1	87.2	6.1	4.9	5.4	43.5	29.9	35.4	75.6
4	dpalominop	91.6	77.2	83.8	12.7	30.6	17.9	42.9	**58.6**	49.5	75.5
5	kongjun	87.1	84.6	85.9	11.1	30.1	16.2	56.1	27.2	36.6	75.3
6	HaoYu	92.9	74.0	82.4	11.9	48.1	19.1	55.2	55.0	55.1	75.2
7	Taha	84.7	89.5	87.0	**51.9**	20.5	**29.4**	10.4	17.5	13.0	75.1
8	meiyim	93.0	73.3	82.0	12.1	**55.8**	19.9	57.7	47.1	51.9	74.5
9	Lavinia_Ap	82.0	97.9	89.2	13.8	3.9	6.1	56.0	8.0	14.1	74.4
10	jupitter	**93.6**	71.8	81.3	11.0	53.9	18.2	58.1	47.9	**52.5**	73.9
11	tangmen	91.8	72.5	81.0	11.3	55.3	18.8	59.8	41.6	49.0	73.0
12	hermosillo748	85.4	81.3	83.3	7.3	21.8	10.9	54.1	26.6	35.7	72.8
13	harsh_6	87.7	77.9	82.5	9.5	23.3	13.5	36.1	39.1	37.5	72.5
14	francesita	80.9	**99.5**	89.2	8.7	1.0	1.7	0.0	0.0	0.0	72.2
15	ajason08	90.1	71.0	79.4	8.2	40.3	13.6	54.7	37.5	44.5	71.0
23	**Baseline**	89.5	63.0	74.0	7.9	49.5	13.6	47.0	31.0	37.4	65.6

Table 4: Top 15 results for the **Spanglish** dataset. The systems are ordered by the *Weighted Average F1* (Avg.) scores of the *Postive*, *Neutral*, and *Negative* classes. We report Precision (P), Recall (R), and F1 score for each class separately. In each column, the boldfaced score is the highest score in that column.

was tried by Zyy1510 (zyy1510). Teams like TueMix (guzimanis), WESSA (ahmed0sultan), C1 (lakshadvani) reported their experiments with Logistic Regression, whereas yet another popular choice Random Forest has been used by IRLab_DAIICT (apurva19), C1 (lakshadvani). SVM was tried by quite a few teams - IUST (Taha), JUNLP (sainik.mahata), WESSA (ahmed0sultan), C1 (lakshadvani), IIITG-ADBU (abaruah).

- **RNN:** RNN, GRU, LSTM, along with their bi-directional varients were explored by several teams. Some of them are gundapusunil (gundapusinil), Team_Swift (aditya_malte), CS-Embed (francesita), GULD@NUIG (koustava), IIT Gandhinagar (vivek_IITGN) etc.

- **CNN for text:** Although RNN is the more popular choices for NLP tasks, quite a few teams also used CNN for text. Some of them are IUST (Taha), FII-UAIC (Lavinia_AP), NLP-CIC (ajason08), NITS-Hinglish-SentiMix (rns2020), Zyy1510 (zyy1510), HCMS (the0ne, talent404).

- **Transformer, BERT and related language models:** The recent trend in NLP is to use highly capable language models like BERT and XLNet. The popular choice, BERT, was tried by MeisterMorxrc (MeisterMorxrc), HinglishNLP (nirantk), IRLab_DAIICT (apurva19), WESSA (ahmed0sultan), C1 (lakshadvani), IIITG-ADBU (abaruah). Some researchers like Deep Learning Brasil - NLP (verissimo.manoel) experimented with XLNet. XLmR was used by Will_go (will_go) , kk2018 (kk2018), FiSSA (jupitter) etc. These type of models gave the best results.

- **Ensembles:** Some teams like Voice@SRIB (asking28), UPB (eduardgzaharia, clementincercel) etc. used ensemble methods. in all cases, ensembles performed better than the their individual models.

- **Special Mentions:** Apart from common practices and architectures quite a few participants explored interesting dimensions and added significant value to this endeavor. We strongly believe these dimensions need to be explored and discussed further.:

  XLP (LiangZhao) used Cross-lingual embeddings which could an interesting way for code-mixed language processing where we have scarcity of annotated data.

	Embeddings				ML Models									LMs			
	Word2Vec	Glove	Character level	fastText	NB	LR	RF	SVM	MLP	CNN	Ensemble	RNN, LSTM and GRU	BERT	XLmR	ALBERT	XLNet
BAKSA										✓		✓	✓	✓		
C1		✓				✓	✓	✓					✓			
CS-Embed	✓											✓				
Deep Learning Brasil - NLP											✓		✓	✓	✓	✓
FII-UAIC										✓						
FiSSA												✓	✓	✓		
gundapusunil	✓	✓	✓	✓		✓	✓	✓	✓							
HCMS			✓	✓				✓		✓		✓	✓			
HPCC-YNU			✓									✓				
HinglishNLP					✓			✓					✓			
IIITG-ADBU								✓					✓			
IIT Gandhinagar		✓										✓				
IRLab_DAIICT	✓					✓	✓				✓		✓			
IUST	✓	✓		✓		✓		✓		✓	✓	✓	✓			
JUNLP								✓								
KK2018													✓	✓		
LIMSI_UPV				✓												
LT3				✓				✓		✓		✓				
MSR India	✓											✓	✓	✓		
MeisterMorxrc													✓			
NITS-Hinglish-SentiMix										✓	✓	✓				
NLP-CIC	✓									✓		✓				
Palomino-Ochoa													✓			
Reed	✓												✓			
Team_Swift				✓								✓		✓		
TueMix						✓				✓		✓				
ULD@NUIG								✓		✓		✓				
UPB								✓		✓		✓	✓	✓		
Voice@SRIB										✓	✓	✓				
WESSA	✓				✓	✓		✓						✓		
Will_go	✓		✓			✓					✓		✓			
XLP										✓			✓	✓		
Zyy1510			✓		✓			✓			✓	✓				

Table 5: Overview of the techniques used by the participants. This is not an exhaustive list. Teams are sorted alphabetically.

UPB (eduardgzaharia, clementincercel) used capsule network with biGRU and showed promising results. The use of capsule networks in NLP tasks need further exploration.

ULD@NUIG (koustava) explored an interesting way to phoneme based Generative Morphemes learning approach. Sub-word based embedding is an interesting new way in the NLP community, but what is the best sub-word unit to choose is still unresolved. Morpheme based approach could be a good alternative, especially for highly spelling variant code-mixed data.

IIT Gandhinagar (vivek_IITGN) tried a new direction by generating sentences using language modeling. Language modeling for code-mixed data is still an under-researched problem.

HPCC-YNU (kongjun) used a Bilingual Vector Gating Mechanism. Vector gating technique got certain success in document classification kinds of applications, but its applications in other NLP dimension demands further exploration.

Will_go (will_go) used Bert and Pseudo labeling. Pseudo Labeling can be a useful strategy for code-mixed languages especially when annotated data is scarce. .

kk2018 (kk2018) reported unique ways to apply adversarial network and its usage in code-mixing. They got very good results.

LIMSI_UPV (somban) gave a way to merge RNN and CNN architecture together for the betterment of sentiment analysis. This could be an interesting way to explore in the future.

8 Conclusion and Future Work

SentiMix, sentiment analysis of code-mixed tweets at SemEval 2020 received an overwhelming response for both Hinglish and Spanglish. 61 teams submitted their systems for Hinglish and 28 teams submitted their systems for Spanglish. The best performance achieved was 75.0 % F1 score for Hinglish and 80.6% for Spanglish. We received a total of 33 system description papers. BERT-like models were the most successful among participants. Although the SentiMix task mainly focused on sentiment analysis, the data will serve the NLP community or whoever is interested in the code-mixing problem for these particular languages and in general.

Properly annotated code-mixed data is still scarce. The success of SentiMix motivates us to go further and organize similar events in the future. We plan to add more languages, especially from regions that have a high percentage of bi- or multilingual speakers. We also plan to enrich our datasets with annotations for other tasks (NER, emotion recognition, translation etc). We strongly believe that code-mixing is a new horizon of interest in the NLP community and needs to be further explored in the future.

References

Laksh Advani, Clement Lu, and Suraj Maharjan. 2020. C1 at SemEval-2020 task 9: Sentimix: Sentiment analysis for code-mixed social media text using feature engineering. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia, July. Association for Computational Linguistics.

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. NLP-CIC at SemEval-2020 Task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Lavinia Aparaschivei, Andrei Palihovici, and Daniela Gîfu. 2020. FII-UAIC at SemEval-2020 Task 9: Sentiment analysis for code- mixed social media text using cnn. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China, July. Association for Computational Linguistics.

Somnath Banerjee, Kunal Chakma, Sudip Kumar Naskar, Amitava Das, Paolo Rosso, Sivaji Bandyopadhyay, and Monojit Choudhury. 2016. Overview of the mixed script information retrieval (MSIR). In *Proceedings of FIRE 2016*. FIRE, December.

Somnath Banerjee, Sahar Ghannay, Sophie Rosset, Anne Vilnat, and Paolo Rosso. 2020. LIMSI_UPV at SemEval-2020 Task 9: Recurrent convolutional neural network for code-mixed sentiment analysis. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Wei Bao, Weilong Chen, Wei Bai, Yan Zhuang, Mingyuan Cheng, and Xiangyu Ma. 2020. Will_go at SemEval-2020 Task 9: An accurate approach for sentiment analysis on hindi-english tweets based on bert and pseudo label strategy. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Subhra Jyoti Baroi, Nivedita Singh, Ringki Das, and Thoudam Doren Singh. 2020. NITS-Hinglish-SentiMix at SemEval-2020 Task 9: Sentiment analysis for code-mixed social media text using an ensemble model. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Arup Baruah, Kaushik Amar Das, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2020. IIITG-ADBU at SemEval-2020 task 9: Svm for sentiment analysis of english-hindi code-mixed text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Elizabeth Bear, Diana Constantina Höfels, and Mihai Manolescu. 2020. TueMix at SemEval-2020 Task 9: Logistic regression with linguistic feature set for sentiment analysis of code-mixed social media text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Raul Berrios, Peter Totterdell, and Stephen Kellett. 2015. Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. *Frontiers in Psychology*, 6:428.

Meghana Bhange and Nirant Kasliwal. 2020. HinglishNLP at SemEval-2020 Task 9: Fine-tuned language models for hinglish sentiment detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, page 48–53, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Eyamba G. Bokamba. 1988. Code-mixing, language variation, and linguistic theory:: Evidence from bantu languages. *Lingua*, 76(1):21 – 62.

Bertelt Braaksma, Richard Scholtens, Stan van Suijlekom, Remy Wang, and Ahmet Üstün. 2020. FiSSA at SemEval-2020 Task 9: Fine-tuned for feelings. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using CRF: Code-switching shared task report of MSR India system. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 73–79, Doha, Qatar, October. Association for Computational Linguistics.

Monojit Choudhury, Gokul Chittaranjan, Parth Gupta, and Amitava Das. 2014. Overview of FIRE 2014 track on transliterated search. In *Proceedings of FIRE 2014*. FIRE, December.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India, December. NLP Association of India.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Manoel Verissimo dos Santos Neto, Ayrton Denner da Silva Amaral, Nádia F F da Silva, and Anderson da Silva Soares. 2020. Deep Learning Brasil - NLP at SemEval-2020 Task 9: Sentiment analysis of code-mixed tweets using ensemble of language models. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, editors. 2020. *Ethnologue: Languages of the World.* SIL International, Dallas, TX, USA, twentythird edition.

Jacob Eisenstein. 2013. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, Georgia, June. Association for Computational Linguistics.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Avishek Garain, Sainik Kumar Mahata, and Dipankar Das. 2020. JUNLP at SemEval-2020 Task 9: Sentiment analysis of hindi-english code mixed data using grid search cross validation. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214 – 224.

Vinay Gopalan and Mark Hopkins. 2020. Reed at SemEval-2020 Task 9: Fine-tuning and bag-of-words approaches to code-mixed sentiment analysis. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Koustava Goswami, Priya Rani, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P McCrae. 2020. ULD@NUIG at SemEval-2020 Task 9: Generative morphemes with an attention model for sentiment analysis in code-mixed text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Sunil Gundapu and Radhika Mamidi. 2020. gundapusunil at SemEval-2020 Task 9: Syntactic semantic lstm architecture for sentiment analysis of code-mixed data. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Soroush Javdan, Taha Shangipour ataei, and Behrouz Minaei-Bidgoli. 2020. IUST at SemEval-2020 Task 9: Sentiment analysis for code-mixed social media text using deep neural networks and linear baselines. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Jun Kong, Jin Wang, and Xuejie Zhang. 2020. HPCC-YNU at SemEval-2020 Task 9: A bilingual vector gating mechanism for sentiment analysis of code-mixed text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Ayush Kumar, Harsh Agarwal, Keshav Bansal, and Ashutosh Modi. 2020. BAKSA at SemEval-2020 Task 9: Bolstering cnn with self-attention for sentiment analysis of code mixed text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Frances A. Laureano De Leon, Florimond Guéniat, and Harish Tayyar Madabushi. 2020. CS-Embed at SemEval-2020 Task 9: The effectiveness of code-switched word embeddings for sentiment analysis. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. kk2018 at SemEval-2020 Task 9: Adversarial training for code-mixing sentiment classification. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yili Ma, Liang Zhao, and Jie Hao. 2020. XLP at SemEval-2020 Task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Aditya Malte, Pratik Bhavsar, and Sushant Rathi. 2020. Team_Swift at SemEval-2020 Task 9: Tiny data specialists through domain-specific pre-training on code-mixed data. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2016. Virtual adversarial training for semi-supervised text classification. *ArXiv*, abs/1605.07725.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas, November. Association for Computational Linguistics.

Daniel Palomino and José Ochoa-Luna. 2020. Palomino-Ochoa at SemEval-2020 Task 9: Robust system based on transformer for code-mixed sentiment classification. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Apurva Parikh, Abhimanyu Singh Bisht, and Prasenjit Majumder. 2020. IRlab_DAIICT at SemEval-2020 Task 9: Machine learning and deep learning methods for sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task @icon-2017. *CoRR*, abs/1803.06745.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Rishiraj SahaRoy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview and datasets of FIRE 2013 track on transliterated search. In *Proceedings of FIRE 2013*. FIRE, December.

Stan Schroeder. 2010. Half of messages on Twitter aren't in English [STATS], February. `http://mashable.com/2010/02/24/half-messages-twitter-english/`.

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 *Proceedings of FIRE 2015*. FIRE, December.

S. Sharma, P. Srinivas, and R. C. Balabantaray. 2015. Text normalization of code mix and sentiment analysis. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1468–1473.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti M. Sharma. 2016. Shallow parsing pipeline - Hindi-English code-mixed social media text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345, San Diego, California, June. Association for Computational Linguistics.

Pranaydeep Singh and Els Lefever. 2020. LT3 at SemEval-2020 Task 9: Cross-lingual embeddings for sentiment analysis of hinglish social media text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Abhishek Singh and Surya Pratap Singh Parmar. 2020. Voice@SRIB at SemEval-2020 Task 9 and 12: Stacked ensembling method for sentiment and offensiveness detection in social media. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Rajendra Singh. 1985. Grammatical constraints on code-mixing: Evidence from Hindi-English. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 30(1):33–45.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Empirical Methods on Natural Language Processing, EMNLP-2008*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014a. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October. Association for Computational Linguistics.

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

Anirudh Srinivasan. 2020. MSR India at SemEval-2020 Task 9: Multilingual models can do code-mixing too. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2020. IIT Gandhinagar at SemEval-2020 Task 9: Code-mixed sentiment classification using candidate sentence generation and selection. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Aditya Srivastava and V.Harsha Vardhan. 2020. HCMS at SemEval-2020 Task 9: A neural approach to sentiment analysis for code-mixed texts. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Ahmed Sultan, Mahmoud Salim, Amina Gaber, and Islam El Hosary. 2020. WESSA at SemEval-2020 Task 9: Code-mixed sentiment analysis using transformers. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558.

S.K. Verma. 1976. Code-switching: Hindi-english. *Lingua*, 38(2):153 – 165.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8, Lisboa, Portugal, September. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Qi Wu, Peng Wang, and Chenghao Huang. 2020. MeisterMorxrc at SemEval-2020 Task 9: Fine-tune bert and multitask learning for sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Holly Young. 2020. The digital language divide. *The Guardian*. Retrieved July 28, 2020. `http://labs.theguardian.com/digital-language-divide/`.

George-Eduard Zaharia, George-Alexandru Vlad, Dumitru-Clementin Cercel, Traian Rebedea, and Costin-Gabriel Chiru. 2020. UPB at SemEval-2020 Task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Yueying Zhu, Xiaobing Zhou, Hongling Li, and Kunjie Dong. 2020. Zyy1510 team at SemEval-2020 Task 9: Sentiment analysis for code-mixed social media text with sub-word level representations. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

A Participants

Team	Codalab Usernames	System Description Paper
BAKSA	ayushk, harsh_6	Kumar et al. (2020)
C1	lakshadvani	Advani et al. (2020)
CS-Embed	francesita	Leon et al. (2020)
Deep Learning Brasil - NLP	verissimo.manoel	dos Santos Neto et al. (2020)
FII-UAIC	Lavinia_Ap	Aparaschivei et al. (2020)
FiSSA	jupitter	Braaksma et al. (2020)
gundapusunil	gundapusunil	Gundapu and Mamidi (2020)
HCMS	the0ne, talent404	Srivastava and Vardhan (2020)
HPCC-YNU	kongjun	Kong et al. (2020)
HinglishNLP	Nirantk	Bhange and Kasliwal (2020)
IIITG-ADBU	abaruah	Baruah et al. (2020)
IIT Gandhinagar	vivek_IITGN	Srivastava and Singh (2020)
IRLab_DAIICT	apurva19	Parikh et al. (2020)
IUST	Taha	Javdan et al. (2020)
JUNLP	sainik.mahata	Garain et al. (2020)
KK2018	kk2018	Liu et al. (2020)
LIMSI_UPV	somban	Banerjee et al. (2020)
LT3	c1pher	Singh and Lefever (2020)
MSR India	genius1237	Srinivasan (2020)
MeisterMorxrc	MeisterMorxrc	Wu et al. (2020)
NITS-Hinglish-SentiMix	rns2020	Baroi et al. (2020)
NLP-CIC	ajason08	Angel et al. (2020)
Palomino-Ochoa	dpalominop	Palomino and Ochoa-Luna (2020)
Reed	gopalvinay	Gopalan and Hopkins (2020)
Team_Swift	aditya_malte	Malte et al. (2020)
TueMix	guzimanis	Bear et al. (2020)
ULD@NUIG	koustava	Goswami et al. (2020)
UPB	eduardgzaharia, clementincercel	Zaharia et al. (2020)
Voice@SRIB	asking28	Singh and Parmar (2020)
WESSA	ahmed0sultan	Sultan et al. (2020)
Will_go	will_go	Bao et al. (2020)
XLP	LiangZhao	Ma et al. (2020)
Zyy1510	zyy1510	Zhu et al. (2020)

Table 6: The teams that participated in Sentimix-2020 and submitted system description papers with the corresponding reference thereof. Teams are sorted alphabetically.