"Kanglish alli names!" Named Entity Recognition for Kannada-English Code-Mixed Social Media Data

Sumukh S and **Manish Shrivastava**
Language Technologies Research Centre (LTRC)
International Institute of Information Technology, Hyderabad, India (IIIT-H)
sumukhs@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Code-mixing (CM) is a frequently observed phenomenon on social media platforms in multilingual societies such as India. While the increase in code-mixed content on these platforms provides good amount of data for studying various aspects of code-mixing, the lack of automated text analysis tools makes such studies difficult. To overcome the same, tools such as language identifiers and parts of-speech (POS) taggers for analysing code-mixed data have been developed. One such tool is Named Entity Recognition (NER), an important Natural Language Processing (NLP) task, which is not only a subtask of Information Extraction, but is also needed for downstream NLP tasks such as semantic role labeling. While entity extraction from social media data is generally difficult due to its informal nature, code-mixed data further complicates the problem due to its informal, unstructured and incomplete information. In this work, we present the first ever corpus for Kannada-English code-mixed social media data with the corresponding named entity tags for NER. We provide strong baselines with machine learning classification models such as CRF, Bi-LSTM, and Bi-LSTM-CRF on our corpus with word, character, and lexical features.

1 Introduction

With the rising popularity of social media platforms such as Twitter, Facebook and Reddit, the volume of texts on these platforms has also grown significantly. Twitter alone has over 500 million test posts (tweets) per day[1]. India, a country with over 300 million multilingual speakers, has over 23 million users on Twitter as of January 2022[2], and code-switching can be observed heavily on this social media platform (Rijhwani et al., 2017).

Code-switching or code-mixing[3] occurs when "lexical items and/or grammatical features from two languages appear in one sentence"(Muysken, 2000). Multilingual society speakers often tend to switch back and forth between languages when speaking or writing, mostly in informal settings. It is of great interest to linguists because of its relationship with emotional expression (Rudra et al., 2016) and identity. However, research efforts are often hindered by the lack of automated NLP tools to analyse massive amounts of code-mixed data (Rudra et al., 2016).

Named Entity Recognition (NER) is the foundation for many tasks related to Information Extraction. When exploring text corpora, being able to explore and browse them by the people and places mentioned in those texts becomes an essential feature.

Below is an example of a code-mixed Kannada-English tweet which has also been translated into English. Named entities have been tagged along with the language tags (*Ka*-Kannada, *En*-English, *NE*-Named Entity, *Univ*-Universal).

> **T1:** Saanu/*Person/NE* next/*Other/En* month/*Other/En* Gujarat/*Location/NE* visit/*Other/En* madtale/*Other/Ka* #excited/*Other/En* :D/*Other/Univ*
>
> **Translation:** *Saanu will visit Gujarat next month #excited :D*

Kannada is a Dravidian language spoken majorly in the Indian state of Karnataka with over 56 million native and second-language (L2) speakers worldwide. Kannada is also one of the six languages designated as a classical language of India by the Indian Government. In code-mixed Kannada-English data, the mixing can happen at phrase, word, syntactic and morphological levels

[1] https://www.internetlivestats.com/twitter-statistics/

[2] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[3] The terms "code-mixing" and "code-switching" are used interchangeably by many researchers, and we also use these terms interchangeably

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 154–161

too (Appidi et al., 2020). This adds to the fact that the data from Twitter is already difficult to analyse given its short length, high language variation, grammatical errors, unorthodox capitalisation, and frequent use of emoticons, abbreviations and hashtags.

There are widely known solutions for NER on monolingual data of high-resource languages like English (Jiang et al., 2022) and low-resource languages like Kannada (Pallavi et al., 2018, Amarappa and Sathyanarayana, 2015), but the same is not true for CM data. NER for code-mixed social media data in low-resource languages has been explored only recently (details in section 2).

In this paper, we have tried to address this problem for Kannada-English code-mixed social media data by creating the first ever corpus with named entity tags and providing strong baselines for the task of NER.

The structure of the paper is as follows. In Section 2, we review the related work. In Section 3, we discuss the annotation methodology and challenges involved. In Section 4, we describe the steps involved in corpus creation and data statistics. In Section 5, we describe our baseline systems. In Section 6, we present the results of the experiments conducted. Finally, in section 7, we conclude the paper and discuss the future prospects.

2 Background and Related Work

A lot of work has been done in Named Entity Recognition (NER) for resource rich language and newswire data such as such as English (Finkel et al., 2005), German (Tjong Kim Sang and De Meulder, 2003), and Spanish (Copara Zea et al., 2016). However, the noisy data from social media platforms like Twitter are different from traditional textual resources due to slacker grammatical structure, spelling variations, abbreviations and more (Ritter et al., 2011). NER for monolingual tweets was explored in Ritter et al. (2011) and Li et al. (2012).

Bali et al. (2014) analysed Facebook posts generated from Hindi-English bilingual users and confirmed the presence of significant code mixing in them. 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, normalise them to their standard forms, assign them their POS tag and segment into chunks. Bhargava et al. (2016) proposed a hybrid model for NER on Hindi-English and Tamil-English CM dataset.

Appidi et al. (2020) reported a work on annotating CM Kannada-English data collected from Twitter and creating POS tags for this corpus. Singh et al. (2018a) presented an automatic NER of Hindi-English CM data while Singh et al. (2018b) and Srirangam et al. (2019) have presented a corpus for NER in Hindi-English and Telugu-English CM data respectively. For Kannada-English CM data, Sowmya Lakshmi and Shambhavi (2017) have proposed an automatic word-level Language Identification (LID) system for sentences from social media posts.

To the best of our knowledge, the corpus created for this paper is the first ever Kannada-English code-mixed social media corpus with Named Entity tags.

3 Annotation Methodology

We label the tags with the present three Named Entity tags 'Person', 'Organisation', 'Location', which using the BIO standard become six NE tags. B-Tag refers to beginning of a named entity and I-Tag refers to the intermediate of the entity, if the name is split into multiple tokens. We use the 'Other' tag for for tokens that don't lie in any of the six NE tags.

'Per' tag refers to the 'Person' entity which is the name of a person, twitter handles and common nick names of people.

The 'Org' tag refers to 'Organisation' entity which is the name of a socio-political organisation like 'Bharatiya Janatha Party', 'BJP', 'JDS'; institutions like 'RBI' and 'Canara bank'; social media companies like 'Youtube', 'Twitter', 'Facebook', 'WhatsApp', 'Google', etc.

'Loc' tag refers to the location named entity which is assigned to the names of places for eg. 'Mysore', 'Shimoga', '#Bengaluru', etc.

The following is an instance of annotation with these tags-

> **T2:** Tomorrow/*Other* ,/*Other* Chandu/*B-Per* Reddy/*I-Per* avru/*Other* Mysore/*B-Loc* alliro/*Other* NVIDIA/*B-Org* Graphics/*I-Org* office/*Other* visit/*Other* madtaare/*Other* !/*Other*
>
> **Translation:** *Tomorrow, Chandu Reddy will visit NVIDIA Graphics office in Mysore!*

The ones which does not lie in any of the mentioned tags are assigned 'Other' tag.

3.1 Challenges

Following are the challenges with annotating Kannada-English code-mixed social media data-

- Word-level/morpheme-level code mixing between Kannada and English makes the problem harder as a CM word is a combination of two words from different languages. This is very common for the mixing of a noun from English language or a named entity and prepositions from Kannada language.

  For example, "*companyge*" is used as a single word in code-mixed Kannada-English sentence which roughly translates (depending on context) to "*to the company*" in English.

  Another common occurrence is the addition of "*-galu*" to indicate plural form of words in Kanglish. For example - "*cargalu*" for "cars", "*companygalu*" for "companies", "*bookgalu*" for "books", etc.

- Users tend to use colloquial words/slang on social media and have their own preference of native words. For example, *baralilla* is a Kannada word and it can be written as *brlilla*, *barlilla*, etc.

- Misspelled words are very common on social media. For example, a word like *tonight* could be written as *tonight, tonite, tonihgt, ton8, etc.,* which posed a significant challenge while building spelling agnostic models.

4 Corpus and statistics

4.1 Data collection

Data collection is a vital step while dealing with any problem with any neural-network based approaches (Roh et al., 2021). As there are only a few sources for code-mixed low-resource language data, this would be challenging as it is difficult to build supervised models.

The corpus that we created from Twitter[4] for Kannada-English code-mixed tweets contains tweets from December 2020 to August 2022. We used hashtags related to city names where Kannada is widely spoken, politics, movies, events, and trending hashtags in collecting the corpus. We also manually identified some of the Twitter account that posted often with code mixing between Kannada and English languages.

Using the twitter API, we retrieved around 222,124 tweets. The following types of tweets were identified and removed-

- Tweets having only English or only Kannada.
- Tweets having only URLs, emojis or hashtags.
- Tweets with less than 5 tokens.

After manually filtering the data with the steps mentioned above, we were left with 3,759 code-mixed Kannada-English tweets. We tokenized these sentences and removed URLs from the same in an effort to reduce the noise.

Label	Count of tokens
Kannada	20,380
English	19,701
Named Entities	8,096
Universal	5,208
Total number of tokens	53,385
Avg. tweet length	14.2
Total tweets	3,759

Table 1: Corpus statistics

Tag	Count of tokens
B-Per	3,729
I-Per	787
B-Org	1,338
I-Org	750
B-Loc	1,137
I-Loc	355

Table 2: NER tag statistics

4.2 Data statistics

The corpus has a total of 53,385 tokens which were tagged for the 7 tags mentioned in the Section 3. The corpus statistics and the tag statistics can be seen in Table 1 and Table 2 respectively.

The corpus will be made available online for public use at the earliest.

4.3 Inter Annotator Agreement

Annotation of the dataset for NE tags in the tweets was carried out by 2 human annotators having linguistic background and proficiency in both Kannada and English based on the methodology in Section 3. In order to validate the quality of annotation,

[4]http://twitter.com/

Tag	Cohen Kappa score
B-Per	0.97
I-Per	0.96
B-Org	0.97
I-Org	0.91
B-Loc	0.96
I-Loc	0.94

Table 3: Inter Annotator Agreement

we calculated the inter annotator agreement (IAA) between the 2 annotation sets of 3,759 code-mixed tweets having 53,385 tokens using Cohen's Kappa (Cohen, 1960). Table 3 shows the results of agreement analysis. We find that the agreement is significantly high. Furthermore, the agreement of 'I-Loc' and 'I-Org' annotation are relatively lower than that of 'I-Per', and this is because of the presence of uncommon/confusing words in these entities.

Disagreements about the tags were resolved through discussions between the annotators to reach a mutual agreement.

5 Experiments

In this section, we present the experiments using different combinations of features and systems. In order to determine the effect of each feature and parameters of the model we performed several experiments using some set of features at once and all at a time simultaneously changing the parameters of the model, like criterion ('Information gain', 'gini') and maximum depth of the tree for decision tree model, regularization parameters and algorithms of optimization like 'L2 regularization', 'Avg. Perceptron' and 'Passive Aggressive' for CRF. Optimization algorithms and loss functions in LSTM. We used 5 fold cross validation in order to validate our classification models. We used 'scikit-learn' and 'keras' libraries in Python for the implementation of the above algorithms.

The training, validation, and testing for all our experiments were 60%, 10%, and 30% of the total data, respectively.

5.1 Conditional Random Field (CRF)

Conditional Random Fields (CRFs) are a class of statistical modelling methods applied in machine learning that takes neighboring sample context into account for tasks like classification. In NER using the BIO standard annotation, I-Org cannot follow I-Per(Tjong Kim Sang and Veenstra, 1999). Since here we are focusing on sentence level and not individual positions, CRFs are suitable and produce better performance measures for NER task.

5.2 Random Forests

Random Forest is a classifier that fits a number of decision trees on various subsets of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (Pedregosa et al., 2011).

On our corpus, a random forest with a max depth of 32, with Gini index as the criterion yielded the best results.

5.3 BiLSTM

Long Short Term Memory (LSTM) is a special kind of RNN architecture that is well suited for classification and making predictions based on time series data. LSTMs are capable of capturing only past information. In order to overcome this limitation Bidirectional LSTMs are proposed where two LSTM networks run in forward and backward directions capturing the context in either directions.

The best result that we came through on our corpus was with a BiLSTM using 'softmax' as activation function, 'adam' as optimizer and 'sparse categorical cross-entropy' for our loss function along with random initialisations of embedding vectors.

5.4 BiLSTM-CRF

The BiLSTM-CRF is a combination of bidirectional LSTM and CRF (Huang et al., 2015;Lample et al., 2016). The BiLSTM model can be combined with CRF to enhance recognition accuracy. This combined model of BiLSTM-CRF inherits the ability to learn past and future context features from the BiLSTM model and use sentence-level tags to predict possible tags using the CRF layer. BiLSTM-CRF has been proved to be a powerful model for sequence labeling tasks like NER (Panchendrarajan and Amaresan, 2018).

After hyperparameter tuning, we found that '*softmax*' as activation function, '*rmsprop*' for optimiser, '*categorical cross-entropy*' as loss function and random initialisations of embedding vectors yielded the best results on our corpus.

5.5 Features

The features to our machine learning models consist of lexical, word-level and character features such as char N-Grams of size 2 and 3 in order to capture the information from emojis, mentions, suffixes in social media like '#', '@', numbers in

the string, numbers, punctuation. Features from adjacent tokens are used as contextual features.

1. **Capitalization:** In social media, people tend to use capital letters to refer to the names of persons, organizations and persons; at times, they write the entire name in capitals(von Däniken and Cieliebak, 2017)to give particular importance or to denote aggression. This gives rise to a couple of binary features. One feature is to indicate if the beginning letter of a word is capitalized, and the other is to indicate if the entire word is capitalized.

2. **Mentions and Hashtags:** People use '@' mentions to refer to persons or organizations, they use '#' hashtags in order to make something notable or to make a topic trending. Thus the presence of these two gives a reasonable probability for the word being a named entity which counts under proper nouns.

   Take the following sentence for example - "*@rakshit nim movies andre tumba ishta, namma #Sandalwood industry improve maadi!"*.

   The token "*@rakshit*" is referring to a person (B-Per tag) and "*#Sandalwood*" is the name of the Kannada film industry (B-Org tag). They are identified by the symbols @ and #. It is important to note that not all hashtags will be a named entity, so we need to understand the word context to correctly classify.

3. **Word N-Grams:** Bag of words has been the standard for languages other than English (Jahangir et al., 2012) in tasks like NER. Thus, we use adjacent words as a feature vector to train our model as our word N-Grams. These are also called contextual features. We used trigrams in the paper.

4. **Character N-Grams:** Character N-Grams are proven to be efficient in the task of classification of text and are language-independent (Majumder et al., 2002). They are helpful when there are misspellings in the text (Cavnar and Trenkle, 1994;Huffman, 1995;Lodhi et al.). Group of chars can help in capturing the semantic information. Character N-Grams are especially helpful in cases like code mixed language where there is free use of words, which vary significantly from the standard Kannada-English words.

Tag	RF	CRF	BiLSTM	BiL-CRF
B-Per	0.32	0.82	0.81	0.84
B-Org	0.70	0.63	0.65	0.63
B-Loc	0.37	0.70	0.82	0.81
I-Per	0.35	0.55	0.57	0.62
I-Org	0.23	0.52	0.46	0.55
I-Loc	0.30	0.46	0.41	0.45
Other	0.95	0.97	0.96	0.97
Wtd avg	0.89	0.93	0.92	0.94

Table 4: F1-scores for CRF, BiLSTM and BiLSTM-CRF respectively with the weighted average at the end.

Feature removed	Precision	Recall	F1
Capitalisation	0.74	0.53	0.61
Mentions, hashtags	0.72	0.57	0.63
Char n-gram	0.65	0.41	0.50
Word n-Gram	0.62	0.44	0.51
Common symbols	0.75	0.48	0.58
Numbers in String	0.78	0.56	0.65

Table 5: Weighted average scores when a specific feature is removed for the BiLSTM-CRF model.

5. **Common Symbols:** It is observed that currency symbols as well as brackets like '(', '[', etc. symbols in general are followed by numbers or some mention not of importance. Hence, these are a good indicator for the words following or before to not being an NE.

6. **Numbers in String:** In social media content, users often express legitimate vocabulary words in alphanumeric form for saving typing effort, to shorten message length, or to express their style. Examples include words like 'n8' ('night'), 'b4' ('before'), etc. We observed by analyzing the corpus that alphanumeric words generally are not NEs, therefore, serves as a good indicator for negative examples.

6 Results and Discussion

Table 4 captures performance of all models for our dataset. Our best model is the BiLSTM-CRF which achieved a weighted average F1-score of 0.94 with '*softmax*' activation function, '*rmsprop*' optimiser, '*categorical cross-entropy*' loss function and random initialisations of embedding vectors. As BiLSTM-CRF can efficiently use both past and future input features from BiLSTM and sentence level tags from CRF, we see that the accuracy is enhanced.

Word	Truth	Predicted
Banashankari	B-Loc	B-Loc
alliro	Other	Other
BESCOM	B-Org	B-Org
kacheeri	Other	Other
alli	Other	Other
work	Other	Other
siktu	Other	Other
Bharat	B-Per	B-Loc
annavrige	Other	Other

Table 6: BiLSTM-CRF example (T1) prediction

Word	Truth	Predicted
Javalli	B-Loc	B-Loc
village	Other	Other
alli	Other	Other
Jnanadeepa	B-Org	B-Org
School	I-Org	I-Org
sersudvi	Other	Other
nan	Other	Other
maga	Other	Other
Suhas	B-Per	B-Per
puttanige	Other	I-Per

Table 7: BiLSTM-CRF example (T2) prediction

Table 5 shows results of our abalation study after removing each particular feature. We can see that the N-grams features have the most impact on our F1-scores, and this is understandable as char n-grams are helpful when there are misspellings and capturing semantic information when there is free use of words which vary significantly from standard word of Kannada and English words.

On analysing some of the results from the model, we see that the intermediate tags of location and organisation is lower than that of a name. This can be explained with the fact that there are uncommon/confusing words in the oraganisation and location names. For example, the word "*Bhaarath*", one of the names for the country India, is "B-Loc" while the words "*Bharat*" and "*Bhaarti*" are common first names in India which are tagged as "B-Per". Furthermore, there are confusing words like "*Bali*" which is a city in Indonesia, but in Kannada, it means "*near*". This can be seen in the example provided in Table 6 where the word "*Bharat*" is referring to a person with that name while our model is predicting that the word is a location, referring to the country India.

We tested a random tweet with the BiLSTM-CRF model that we trained, and here is the model predicted tags along with the ground truth tags in the Table 7. We noticed that the I-Per is predicted incorrectly for the Kannada word *puttanige* (an endearment word for kids) as this word is very similar to some of the common last names in southern part of India such as *Puttanna* and *Puttagere*. The low scores for intermediate tags (*I-per, I-Org* and *I-Loc*) can be attributed to these reasons along with the "noisiness" of the social media data which tends to have misspelled words and colloquial forms of words. This gets more difficult with Kannada-English code-mixed data as mixing happens at word-level, mostly for Kannada language prepositions and named entities or English language nouns (Section 3.1).

7 Conclusion and future work

The following are our contributions in this paper.

1. An annotated code-mixed Kannada-English corpus for named entity recognition, which to the best of our knowledge, is the first corpus. The corpus will be made available online soon along with the models.

2. Introducing and addressing Named Entity Recognition (NER) of Kannada-English code-mixed data as a research problem.

3. We have experimented with the machine learning models Random Forest, CRF, BiLSTM and BiLSTM-CRF on our corpus and achieved an F1-score of 0.89, 0.93, 0.93 and 0.94 respectively, which looks good considering the complexity of the task and the amount of research done in this new domain for low resource languages.

As part of future work, we plan to explore downstream tasks like semantic labelling and entity-specific sentiment analysis which makes use of NER for code-mixed data. The size of the corpus can be increased to include more data from varied topics.

8 Acknowledgements

We would like to thank our annotators for their hard work and dedication. We would also like to thank the anonymous reviewers, Prashant Kodali, and Mohsin Mustafa for their valuable feedback.

References

S. Amarappa and S. V. Sathyanarayana. 2015. Kannada named entity recognition and classification (nerc) based on multinomial naïve bayes (mnb) classifier. *CoRR*, abs/1509.04385.

Abhinav Reddy Appidi, Vamshi Krishna Srirangam, Darsi Suhas, and Manish Shrivastava. 2020. Creation of corpus and analysis in code-mixed Kannada-English social media data for POS tagging. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 101–107, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

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, Doha, Qatar. Association for Computational Linguistics.

Rupal Bhargava, Bapiraju Vamsi Tadikonda, and Yashvardhan Sharma. 2016. Named entity recognition for code mixing in indian languages using hybrid approach. In *FIRE*.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Jenny Linet Copara Zea, Jose Eduardo Ochoa Luna, Camilo Thorne, and Goran Glavaš. 2016. Spanish NER with word representations and conditional random fields. In *Proceedings of the Sixth Named Entity Workshop*, pages 34–40, Berlin, Germany. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Stephen Huffman. 1995. Acquaintance: Language-independent document categorization by n-grams. In *TREC*.

Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa, and Xuan Wang. 2012. N-gram and gazetteer list based named entity recognition for Urdu: A scarce resourced language. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 95–104, Mumbai, India. The COLING 2012 Organizing Committee.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building NLP models for social media analysis. *CoRR*, abs/2201.07281.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 721–730, New York, NY, USA. Association for Computing Machinery.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels.

P Majumder, M Mitra, and B. B. Chaudhuri. 2002. N-gram: a language independent approach to ir and nlp.

Pieter Muysken. 2000. *Bilingual speech*.

K. P. Pallavi, L. Sobha, and M. M. Ramya. 2018. Named entity recognition for kannada using gazetteers list with conditional random fields. *Journal of Computer Science*, 14(5):645–653.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional lstm-crf for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

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. *the Journal of machine Learning research*, 12:2825–2830.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018a. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018b. Named entity recognition for Hindi-English code-mixed social media text. In *Proceedings of the Seventh Named Entities Workshop*, pages 27–35, Melbourne, Australia. Association for Computational Linguistics.

B S Sowmya Lakshmi and B R Shambhavi. 2017. An automatic language identification system for code-mixed english-kannada social media text. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pages 1–5.

Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, and Manish Shrivastava. 2019. Corpus creation and analysis for named entity recognition in Telugu-English code-mixed social media data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 183–189, Florence, Italy. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 173–179, Bergen, Norway. Association for Computational Linguistics.

Pius von Däniken and Mark Cieliebak. 2017. Transfer learning and sentence level features for named entity recognition on tweets. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 166–171, Copenhagen, Denmark. Association for Computational Linguistics.