

MaCmS: Magahi Code-mixed Dataset for Sentiment Analysis

Priya Rani, Gaurav Negi, Theodorus Fransen,
John P McCrae
Data Science Institute, University of Galway

HOST INSTITUTION



PARTNER INSTITUTIONS



Introduction

- Sentiment analysis, the process of determining emotions or opinions expressed in text, presents unique challenges in code-mixed languages
- These languages often intertwine, creating a complex linguistic landscape.
- Despite the challenges, successful sentiment analysis in such contexts can provide valuable insights into the sentiments of multilingual communities on social media.



Contributions

- **MaCmS**: an annotated Magahi (MAG), Hindi (HIN), English (ENG) code-mixed dataset for sentiment analysis. To the best of our knowledge, this is the first Magahi code-mixed dataset for sentiment analysis.
- A linguistic analysis of the structure of code-mixing between two closely related languages, Magahi and Hindi. Which helps understand when and where the code-mixing is happening.
- A statistical analysis of the dataset based on the language preference used by the speakers, which indicates the emotions and attitude of the speakers and the sentiment they show.
- We also provide some baseline models for sentiment analysis at the sentence and language-specific span levels.

HOST INSTITUTION



PARTNER INSTITUTIONS



Dataset Overview



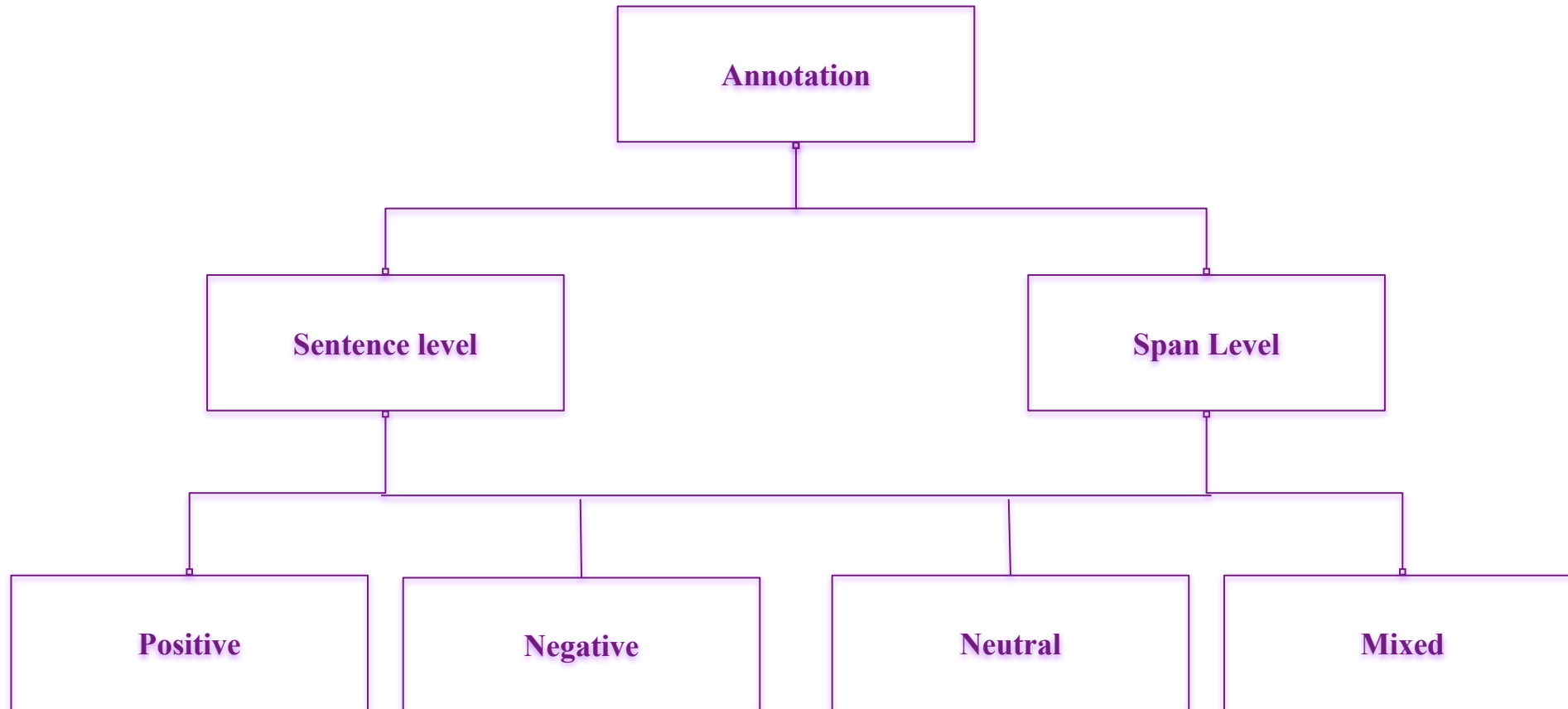
MaCmS	Statistics
Sentences	5000
Span Sentences	750
Total Span	2642

Statics of the Dataset

Examples

Number	Sentence	Translation	Sentiment
Sent: 1	Bahut sundar bahut wait kara hi apne vedio ke	Very nice I am eagerly waiting for your video	Positive
Sent: 2	तोरा दोनो के सामने सलमान खान शारुखखान अमीर खान अजय देवगनसब फैल है जी हमको बिहारी होने पे इतना गरब है कि बता नय सकीवो हा जी तोर दोनो के वीडियो देखो हीवो तो मन खुश हो जाहो	Salman Khan, Shahrukh Khan and Ajay Devgan are nothing in front of you both; I am so proud to be a Bihari that I can't express. Moreover, I am delighted whenever I see your videos	Positive
Number	Span and (language)	Translation	Sentiment
Sent: 1 Span1	Bahut sundar bahut (Hindi)	Very nice	Positive
Span2	wait (English)	wait	Neutral
Span3	kara hi apne vedio ke (Magahi)	For your video	Neutral
Sent: 2 Span1	तोरा दोनो के सामने सलमान खान शारुखखान अमीर खान अजय देवगनसब फैल है जी (Magahi)	Salman Khan, Shahrukh Khan and Ajay Devgan are nothing in front of you both	Positive
Span2	हमको बिहारी होने पे इतना गरब है कि (Hindi)	I am so proud to be bihari that	Positive
Span3	बता नय सकीवो हा जी तोर दोनो के वीडियो देखो हीवो तो मन खुश हो जाहो (Magahi)	I can't express. Moreover, I am delighted whenever I see your videos.	Positive

Annotation Overview



HOST INSTITUTION

PARTNER INSTITUTIONS

Annotation Workflow

01	Disagreement between the annotators	<ul style="list-style-type: none">• Re-Annotate
02	One disagreement and three agree	<ul style="list-style-type: none">• Majority wins
03	Two agreement and two different opinion	<ul style="list-style-type: none">• Majority wins
04	Equal distribution for both agreement and disagreement	<ul style="list-style-type: none">• Reannotate

HOST INSTITUTION

PARTNER INSTITUTIONS

Inter-annotator Agreement

- Annotators Details
 - 4 annotators (2 male and 2 female)
 - Mother Tongue - Magahi
 - Aged -18-28
- 3 Phases of Annotation
 - Phase 1 & 2 had 500 sentences each
 - Phase 1 and 2 approx 300 spans each
- Krippendorf's @ score

Phase	Sentence-level	Span-Level
Phase-1	0.67	0.69
Phase-2	0.72	0.76
Final	0.78	0.76

Agreement Scores

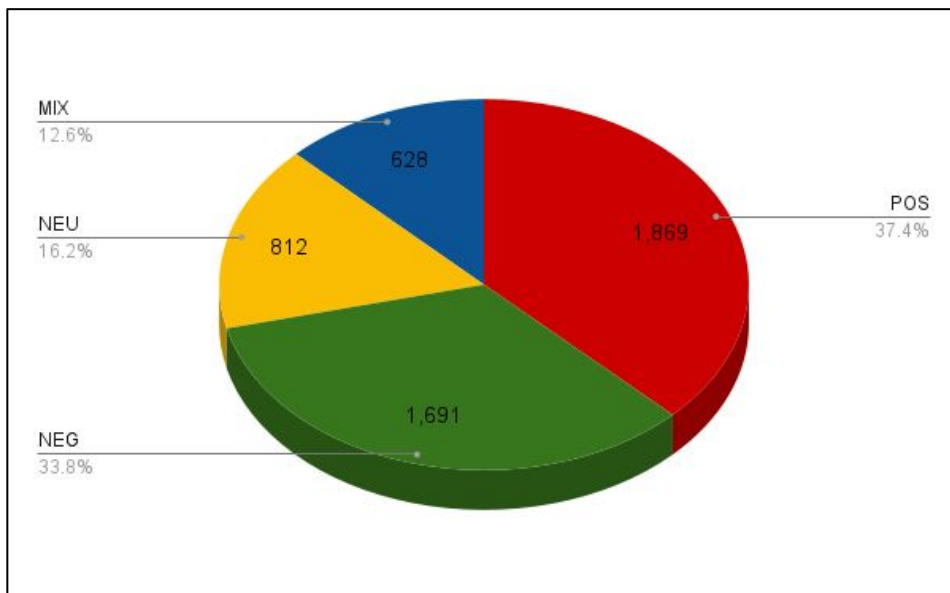
HOST INSTITUTION



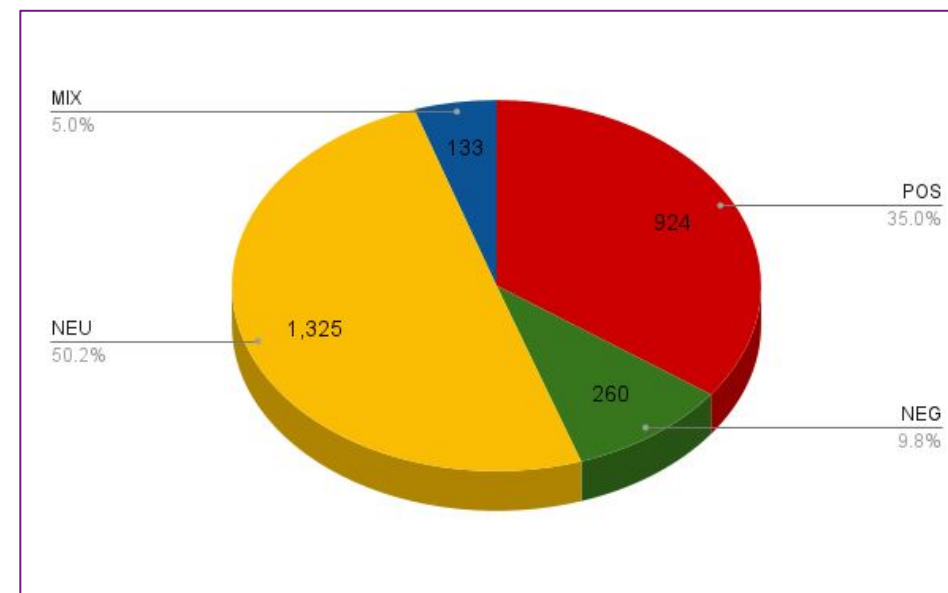
PARTNER INSTITUTIONS



Statistical Analysis



Sentence-level data distribution after annotation.



Span-level data distribution after annotation

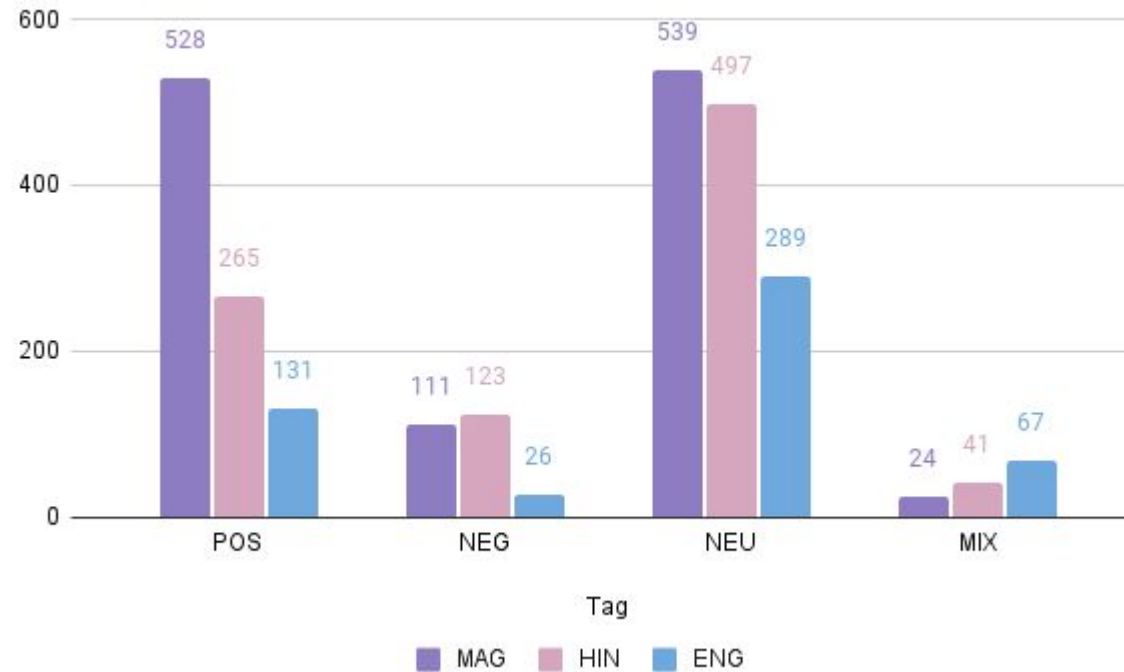
HOST INSTITUTION



PARTNER INSTITUTIONS



Statistical Analysis



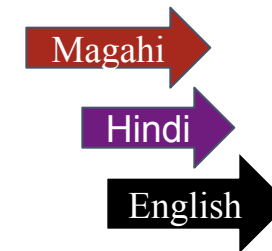
Language-specific span-level data distribution

HOST INSTITUTION

PARTNER INSTITUTIONS

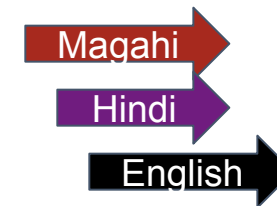
Linguistic Analysis

- Like most closely related languages, **Hindi** and **Magahi** share a lexicon, possibly due to various reasons like phonetic similarity in the spelling across languages. For example:
 - `u' : -- `that' in **Magahi**, `You' in **English**
- When the speakers try to express their strong emotions or gain attention, they code-mix a substantial number of functional words, including Wh-words, adjectives, pronouns, determiners, etc.
 - **E Match me aapka dubbing matching ho gya Sirji**
this match in your-3SG.HON Dubbing match finish-PST AUX Sirji
Translation: Your dubbing got matched in this match Sirji



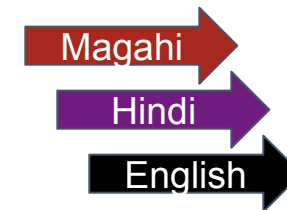
Linguistic Analysis

- बहुते सुंदर अपने मगही मे गीत गइली मन गदगद होगेल।
very beautiful you-2SG.HON Magahi in song sing-PST.2SG.HON mind happy be-PST
Translation: You sang a very beautiful song in Magahi. It was mind-blowing.
- Insertion of the marker **-wa** and the numeral classifier **`go`** with the noun and act like a classifier language rather than a noun class like Hindi.
 - Ek du **go** aur likhiye
one two **go-CLF** more write-PRS.2SG.HON
Translation: Write one or two more.
 - WOW haste haste mera **pet-wa** fat jayega
WOW laugh-PRS.PROG.1SG my stomach-POSS hurt-PRS.PROG.1SG
Translation: My stomach hurts from laughing



Linguistic Analysis

- When the speakers try to quote somebody or some famous expression, they code-mix.
 - Once a legend said ``Laura k sarkaar
Once a legend said ``evil government"
- While talking about culture or regional traditions, the speakers often code-mix a lot as they are comfortable talking about their culture and tradition in their mother tongue.
 - Murna ke khasi khaiye ke hai Murna kab hai ☺
Murna- GEN mutton eat-PRS.1SG be PRS when be-PRS it
Translation: We have to get Murna's party, When is the Murna?



HOST INSTITUTION



PARTNER INSTITUTIONS



Linguistic Analysis

- When the speakers try to express their strong emotion of surprise or any other emotion, they are prone to code-mix between Magahi English as they insert interjections or exclamation in the text. Even when the use the complementizer they code mix.
 - **Bhaiya hum kehrhe hain ki E sarkar ke sabhe kala cita khol da**
Brother I tell-1SG that-COMP this government evil doing bring out
Translation: Brother I am telling you to bring out all the evil doing of this government.



Baseline of the Sentiment Analysis

Models	Precision	Recall	F1 Score
MBert (Sentence)	0.65	0.71	0.68
XLM-R (Sentence)	0.76	0.75	0.75
GeMa (Sentence)	0.68	0.69	0.68
MBert (Span)	0.55	0.52	0.53
XLM-R (Span)	0.50	0.48	0.49
GeMa (Span)	0.58	0.43	0.51

HOST INSTITUTION



PARTNER INSTITUTIONS



Conclusion

- Study the speaker's language preference for a specific context with the help of a newly developed dataset in the case of a closely related code-mixed scenario.
- The analysis concluded that Magahi is used for expressing positive sentiments compared to negative sentiments.
- These results do not agree with the previous studies (Rudra et al., 2016), (Dogruoz et al., 2021), (Agarwal et al., 2017), which state that the speakers prefer the first language to express negative sentiments or while swearing.
- The analysis also concludes a significant association between sentiment and language proportion.
- This association allows us to explore further and experiment to improve the language Identification models for closely related code-mixed scenarios.

HOST INSTITUTION



PARTNER INSTITUTIONS



References

- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.
- Ada Wan. 2016. Leveraging data-driven methods in word-level language identification for a multilingual alpine heritage corpus. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 45–54.
- Goswami, K., Sarkar, R., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. [2020, December]. Unsupervised deep language and dialect identification for short texts. In *Proceedings of the 28th International Conference on Computational Linguistics* [pp. 1606-1617].
- Diab, M., Hirschberg, J., Fung, P., & Solorio, T. [2014, October]. Proceedings of the First Workshop on Computational Approaches to Code Switching. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.
- Barman, U., Das, A., Wagner, J., & Foster, J. [2014, October]. 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* [pp. 13-23].
- Mave, D., Maharjan, S., & Solorio, T. [2018, July]. Language identification and analysis of code-switched social media text. In *Proceedings of the third workshop on computational approaches to linguistic code-switching* [pp. 51-61].
- Gambäck, B., & Das, A. [2014, December]. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India* [pp. 1-7].

HOST INSTITUTION



PARTNER INSTITUTIONS



ACKNOWLEDGMENT

Centre for
Research
Training



This work has been done with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under grant number 18/CRT/6223.

Thank You!

HOST INSTITUTION



PARTNER INSTITUTIONS

