



University of
Zagreb

M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets

Gaurish Thakkar¹, Sherzod Hakimov², Marko Tadić¹

¹Faculty of Humanities and Social Sciences, University of Zagreb

²Computational Linguistics, University of Potsdam



LREC-COLING  2024

Introduction

- In recent years, multimodal natural language processing, aimed at learning from diverse data types, has garnered significant attention.
- However, there needs to be more clarity when it comes to analysing multimodal tasks in multi-lingual contexts.
- The process of annotating supervised datasets for natural language processing (NLP) tasks is a labour-intensive endeavour requiring significant investment of time, financial resources, and effort

Introduction

- A straightforward approach for enhancing pre-existing publicly accessible datasets to conduct multimodal (image & text) sentiment analysis on Twitter called M2SA (Multimodal Multilingual Sentiment Analysis)
- The fundamental hypothesis underlying our utilisation of the unimodal dataset posits that, given its gold annotation, the Twitter dataset can be linked to an image that has not been previously examined or employed in the context of multimodal sentiment classification

M2SA

- Tweets are optionally accompanied with images (or videos).
- These images provide additional context for the text.

Tweets can be examined to determine whether there are any additional modalities present.



Multimodal Multilingual Sentiment Analysis (M2SA)

- Data collection
 - Hugging Face Datasets, European Language Grid, and GitHub etc.
 - specific keywords such as **twitter sentiment analysis dataset**, **social media sentiment analysis dataset**, and **twitter sentiment shared tasks**.
- Hydrating a tweet
 - The compiled list of datasets undergoes the process of querying tweet information using the Twitter API
- Dataset filtering
 - The initially collected datasets are then subjected to manual checking to exclude tasks unrelated to sentiment analysis.

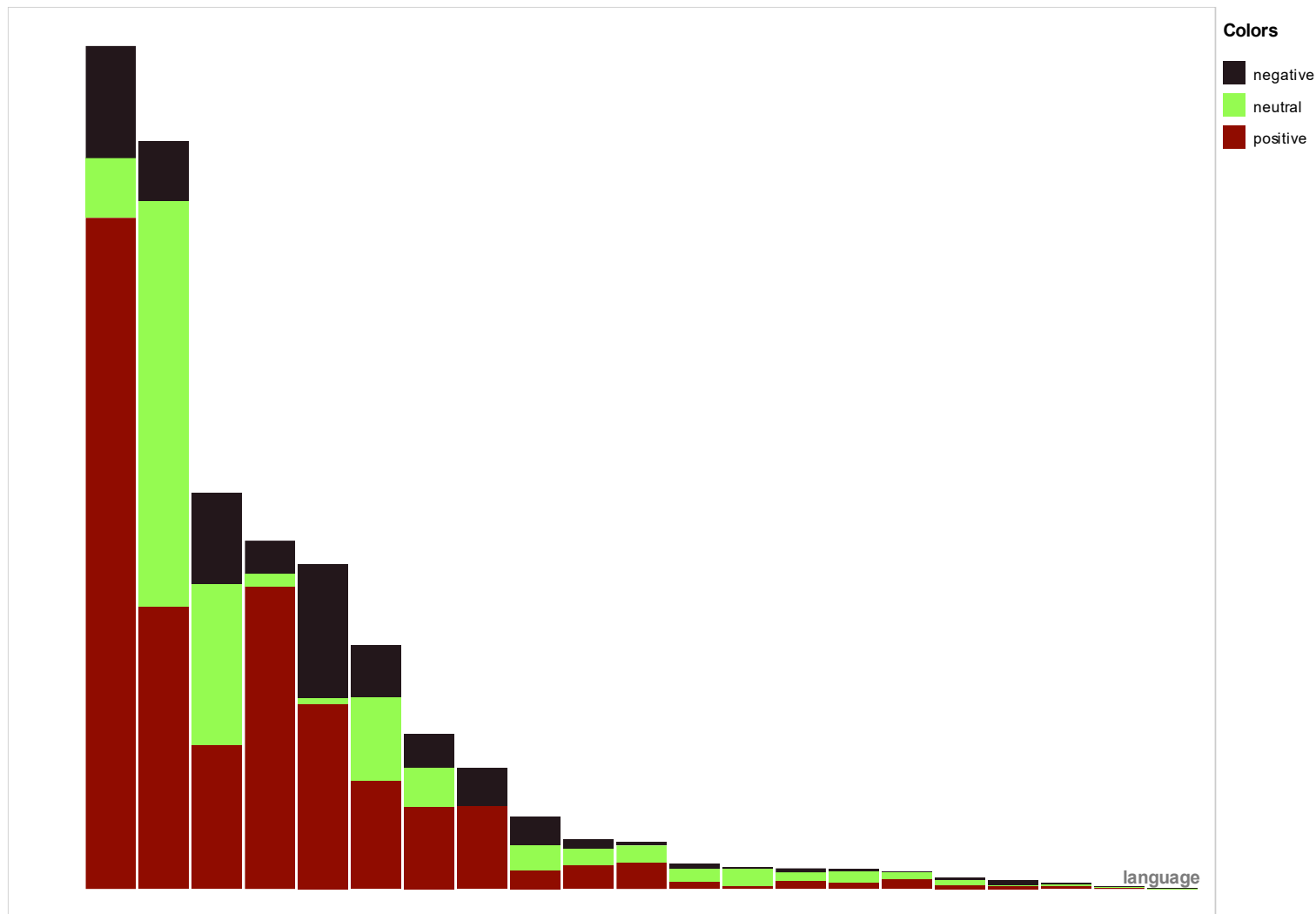
Preprocessing

- Removal of all black and white images.
- Tweet normalisation for USERS, URLs and HASHTAGS
- Filtering of tweets with text content less than five characters, not accounting for USER and URL tags.
- Deduplication is performed using tweet IDs.
- Checking if the same tweet ID has more than one label assigned and employing a majority vote when needed.
- Filtering of tweets with corrupted or no images or with images of less than 200 × 200 pixels size.
- Checking the language tag in the tweet JSON and see if it matches the target language.
- Translation of English tweets for lower-resourced languages using the NLLB5 machine translation (MT) model.

Dataset Fields

- tweetid: unique identifier for the tweet.
- normalised-text: text obtained after applying preprocessing steps.
- language: the language of the text.
- translated-text: text in the target language obtained using the NLLB model.
- image-paths: list of images associated with the tweet.
- label: POSITIVE|NEGATIVE|NEUTRAL

Dataset Distribution



Lang	Dataset name
ar	SemEval-2017
ar	TM-Senti@ar
bg	Twitter-15@Bulgarian
bs	Twitter-15@Bosnian
da	AngryTweets
de	xLiMe@German, Twitter-15@German, TM-Senti@de
en	SemEval-2013-task2, SemEval-2015, SemEval-2016
en	CB COLING2014 vanzo
en	CB IJCOL2015 ENG castellucci
en	RETWEET
es	xLiMe@spanish
es	Copres14
es	mavis@tweets
es	Twitter-15@Spanish
es	JOSA corpus
es	TASS 2018, 2019, 2020
es	TASS 2012, 2013, 2014, 2015
fr	DEFT 2015
hr	InfoCoV-Senti-Cro-CoV-Twitter
hr	Twitter-15@Croatian
hu	Twitter-15@Hungarian
it	CB IJCOL2015 ITA castellucci
it	xLiMe@Italian
it	sentipolc16
it	TM-Senti@it
lv	Latvian tweet corpus
mt	Malta-Budget-2018, 2019, 2020
pl	Twitter-15@Polish
pt	Twitter-15@Portuguese
pt	Brazilian tweet@tweets
ru	Twitter-15@Russian
sq	Twitter-15@Albanian
sr	doiserbian@tweet
sr	Twitter-15@Serbian
sv	Twitter-15@Swedish
tr	BounTi Turkish
zh	TM-Senti@zh-ids

Table 1: Languages and their corresponding dataset names

Experiments

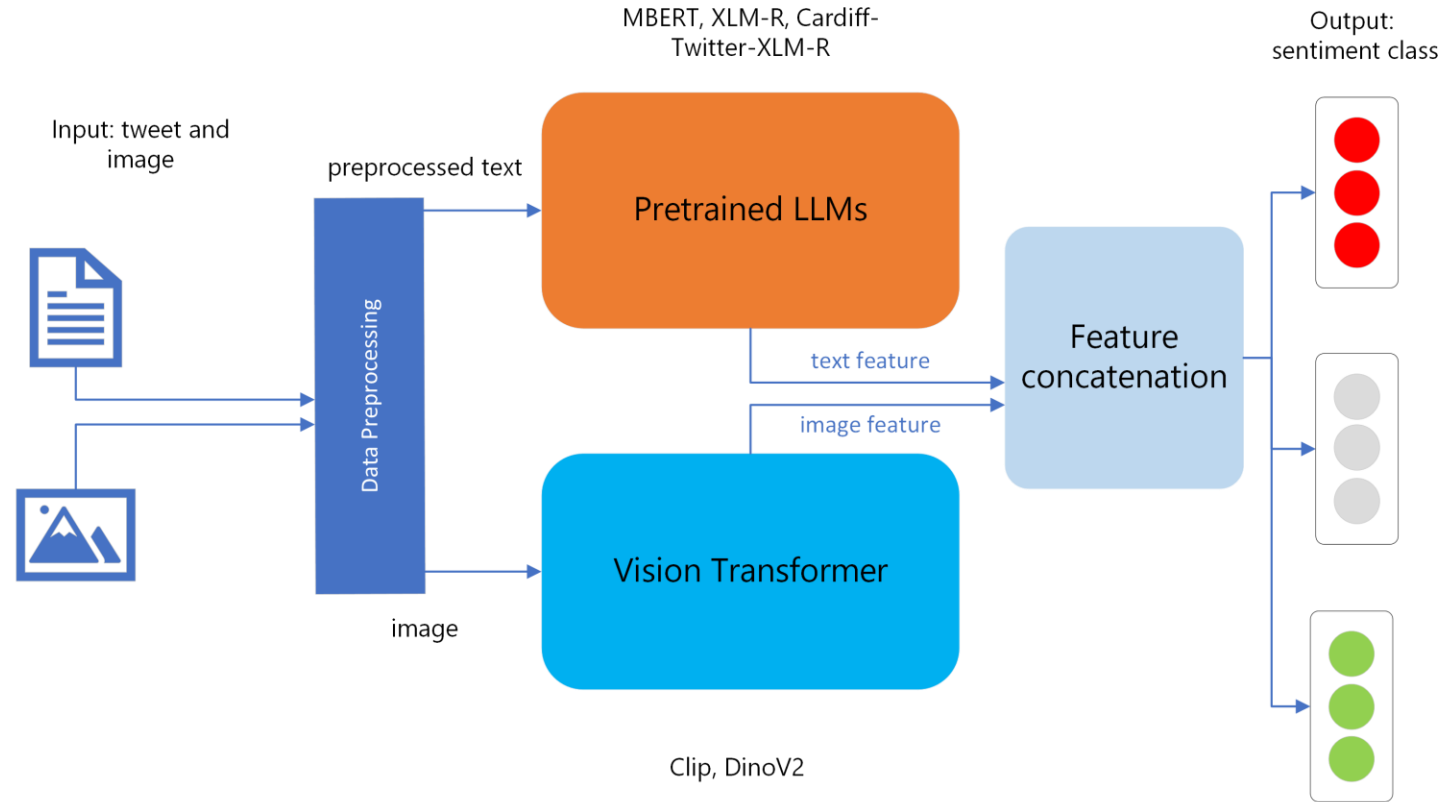
- Model Configurations
 - Unimodal vs. Multimodal
 - Original data vs. Inclusion of translation
 - Monolingual vs. Multilingual

Data
Original
Machine Translated

Text Model
M-BERT
XLNet-Roberta
XLNet-RoBERTa-Sentiment-Multilingual

Vision Model
CLIP
DINOv2

Model Architecture

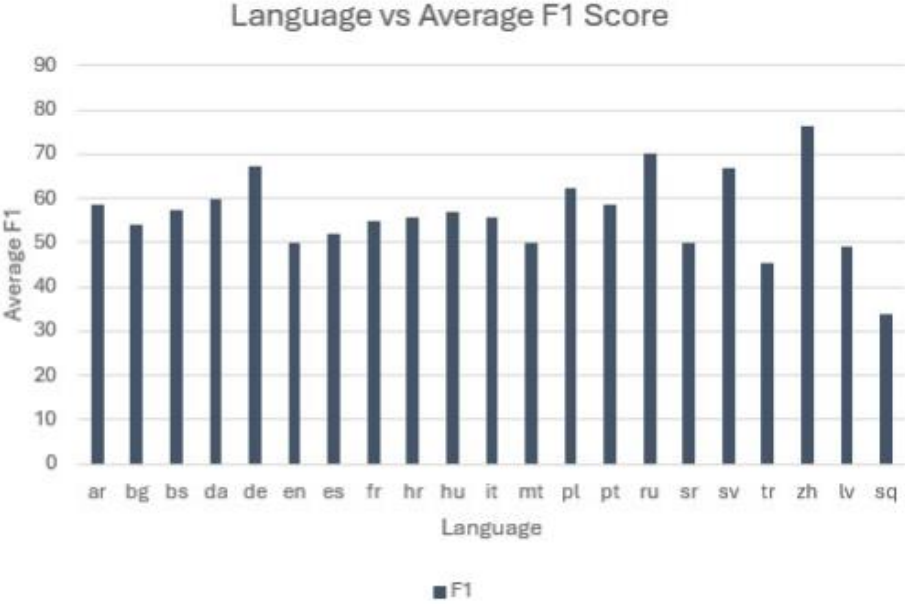
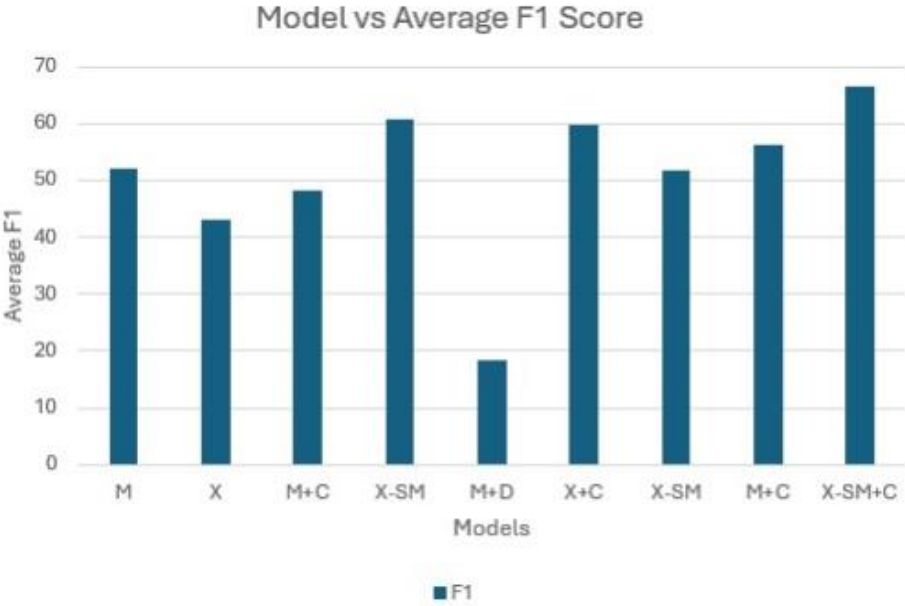


Results

Lang	M	X	M+C	X-SM	M+D	X-SM+C	X-SM	M+C	X-SM+C
	monolingual						multilingual		
ar	57.3	64.6	53.6	66.5	25.1	69.1	41.0	61.3	72.7
bg	51.9	38.0	53.7	63.1	11.1	60.5	53.5	57.8	60.8
bs	62.4	57.0	60.5	64.4	35.4	66.5	40.3	63.1	67.9
da	48.8	34.5	46.9	66.9	21.9	59.1	55.1	57.8	75.2
de	68.7	89.1	69.4	90.1	10.7	89.6	56.3	75.3	92.9
en	34.1	18.8	30.4	36.2	6.6	33.0	64.2	52.2	53.7
es	46.5	22.6	36.9	51.6	8.0	46.4	61.4	49.4	59.6
fr	51.1	40.2	50.9	64.5	18.5	64.9	65.8	41.0	51.5
hr	58.5	28.7	56.4	64.6	25.7	55.9	40.5	57.7	63.4
hu	50.9	43.1	50.5	62.5	17.8	66.3	47.3	56.1	63.7
it	40.3	29.8	24.0	55.8	4.4	60.2	54.4	56.6	63.1
mt	60.3	60.3	60.0	68.3	11.9	62.0	35.9	44.0	56.8
pl	67.8	45.3	46.2	68.7	12.7	69.5	51.2	63.8	72.3
pt	67.2	48.1	51.8	64.3	29.5	74.6	48.3	52.8	61.8
ru	65.5	43.9	70.6	73.1	27.1	75.3	64.9	65.7	82.3
sr	42.6	23.4	38.1	49.7	21.6	43.8	48.7	49.9	65.3
sv	68.2	43.0	59.2	73.1	28.7	73.3	54.5	66.0	80.2
tr	45.9	32.1	44.4	49.6	11.6	49.4	47.9	41.3	47.8
zh	57.6	98.9	64.9	99.0	26.3	98.4	43.9	68.7	98.4
lv	22.6	19.0	24.8	22.0	21.5	18.1	76.8	52.4	61.6
sq	20.7	20.7	20.5	20.5	7.8	20.5	33.7	43.5	45.4
bg_mt	26.1	23.5	25.8	23.5	9.1	29.4			
bs_mt	17.3	19.0	15.6	18.5	9.1	20.6			
da_mt	20.7	20.7	20.7	24.5	15.0	24.7			
fr_mt	23.1	23.1	23.1	25.8	13.6	23.4			
hr_mt	34.0	25.4	28.9	34.9	16.5	46.9			
hu_mt	28.7	21.2	22.8	28.6	10.3	28			
mt_mt	30.1	18.6	20.9	43.8	12.0	26.3			
pt_mt	16.4	8.7	10.5	22.9	23.4	21.9			
ru_mt	41.3	17.8	28.8	46.9	23.7	45.6			
sr_mt	18.8	18.8	18.6	25.5	17.8	23.0			
sv_mt	31.7	17.3	24.3	54.6	19.8	34.7			
tr_mt	33.7	30.8	32.5	31.5	13.8	30.8			
zh_mt	38.2	66.7	38.0	78.1	25.3	85.4			

Table 2: F1 comparison of models using visual and textual features. M: M-BERT, C: CLIP, X: XLM-Roberta, X-SM: XLM-RoBERTa-Sentiment-Multilingual, D: DINOv2. {lang}_mt: it refers to the model that uses data from original tweets and their translations for that specific lower-resourced language. The value included within a cell containing model headers signifies the model's performance on the test set for the specific language indicated by the lang column. Monolingual training involves the use of data from a single language, whereas multilingual training involves the incorporation of training data from multiple languages. The best result for each language is highlighted in bold.

Average Scores



Conclusion

- This paper presents the model architecture trained on the dataset extracted from various sources for multimodal sentiment classification in a multilingual context
- We employed a straightforward methodology to enhance an existing unimodal dataset from Twitter, transforming it into a multimodal one
- Numerous models have been trained utilising textual data and a combination of textual and visual modalities
- Training a single model for all languages multilingual and multimodal data yielded the best performance across many languages.

Thank you