Leveraging BERT, MWE, and ML Models to Detect Emotions and Threats in Urdu

Bénédicte Diot-Parvaz-Ahmad, Pierre Magistry, Ilaine Wang and Damien Nouvel

INALCO ERTIM, 2 rue de Lille, 75007 Paris, France

Abstract

Our participation in the EmoThreat 2022 challenge relied on 3 distinct methods. We used algorithms based on BERT, a traditional statistic method and unsupervised Multi Word Expression (MWE) extraction to resolve the two proposed tasks to classify Urdu tweets. Task A consisted in classifying tweets into 6 classes of emotions. Task B was subdivided into 2 sub-tasks: B1) detecting threat and B2) specifying if the threat targets a single person or a group. For task A, the unsupervised MWE extraction method obtained the best results. For task B, the BERT based approach was more efficient to identify threats and the ML traditional method turned out to be the best among challenge participants to detect individual or collective threat. After discussing the types of errors, we suggest some guidelines for further improvement of those models.

Keywords

Urdu, BERT models, MWE, Machine Learning,

1. Introduction

The EmoThreat 2022 challenge consisted in two classification tasks for Urdu tweets corpora. In task A, candidates had to classify tweets into 6 classes of emotions. Task B was sub-divided into 2 sub-tasks: the first one aimed at detecting threats for a given tweet, while the goal of the second one was to distinguish threats to individuals or groups of persons. Our research team ER-TIM, was interested in participating to the EmoThreat 2022 campaign for several reasons. First, members of the team, among which an Urdu speaker is present, conducted a project about text classification with assistance (active learning) for which Urdu was one of the experimented languages. Second, Urdu still lacks resources for NLP in comparison with other South Asian languages. We also wanted to test machine learning (including fine tuning of large language models) for different classification systems: unsupervised extraction of MWE for task A, BERT based models like RoBERTa, MURIL or statistical method for task B. After describing the data and systems used for each task (Sections 3 and 4), we present the results obtained with each method (Section 5), analyse the errors produced (Section 6).

FIRE 2022: Forum for Information Retrieval Evaluation, December 9–13, 2022, India

benedicte.parvazahmad@inalco.fr (B. Diot-Parvaz-Ahmad); pierre.magistry@inalco.fr (P. Magistry); ilaine.wang@inalco.fr (I. Wang); damien.nouvel@inalco.fr (D. Nouvel)

0000-0002-4103-5926 (B. Diot-Parvaz-Ahmad); 0000-0002-9296-8902 (P. Magistry); 0000-0002-0047-9117 (I. Wang); 0000-0001-8866-4028 (D. Nouvel)

2. Literary Review

Several approaches have been experimented on very similar tasks during the last years. In Zameen and Sayeed [1], the authors used an approach based on Markov chains for predicting the sentiment of Urdu tweets (classification along 3 classes), which is very close to task A. In Khan et al. [2], deep learning methods had been implemented for Urdu sentiment analysis and in Khan et al. [3], the BERT model was used for classifying sentiments in Urdu texts with multi-labels. In Muhammad et al. [4], authors used in their ML approach techniques based on n-grams at word and character levels to extract relevant features for detecting abusive language in Urdu social medias (close to task B). In Das et al. [5], the base models used again for detecting abusive language in Indic languages including Urdu were m-BERT and MuRIL. Our work was based on these approaches and an original one using MWE unsupervised extraction.

3. Datasets

Organizers of EmoThreat 2022 provided train and test datasets for each task. For task A, the train set consisted of 7,801 tweets and the test set of 1,951 tweets. For task B, there were 3,564 tweets in the train set and 635 in the test one. As commonly seen in classification tasks, datasets are unbalanced, especially for task B. Further details are given in Amjad et al. [6].

We also used an extra set of tweets from Batra et al. [7], which contained more than one million Urdu tweets. We selected and annotated a small part of these to enrich our train dataset.

In addition to those datasets, we also conducted a small annotation of the test dataset. Our idea was to have an idea of the annotation agreement between our Urdu speaker and EmoThreat organizers and of the required human effort to conduct the considered task. Those annotations were not used to learn models for our run submissions. This resulted in 60 tweets annotated, among which 14 of them did not receive the same class than those provided by organizers. The agreement rate is 0.78 and Cohen's Kappa 0.67, which reveals a strong agreement.

4. Systems

Three approaches have been implemented. For both tasks, we used models provided by the HuggingFace repository. In addition, we used and unsupervised MWE extraction for task A and a machine learning and a linguistic-based approach for task B.

4.1. HuggingFace models

In order to classify tweets in relevant categories, we experimented the famous HuggingFace transformers, by searching available models able to process Urdu (including multilingual ones). Among proposed models, we conducted a preliminary evaluation for `roberta-urdu-small`[1], `muril-large-cased'` [8], `'albert-large-urdu'` (no reference has been found) and `'bert-large-uncased'` [9] (multilingual). All these models are used with the `AutoTokenizer` and

[1]https://github.com/urduhack/urduhack

`AutoModelForSequenceClassification`[2].

Considering task B, we directly implemented the B2 subtask, which allows to submit both B1 and B2 runs. We also extended our dataset with an external corpus of tweets[10], our idea was to manually annotate tweets similar to those that were misclassified in the training dataset in a cross-validation setting. The first step was to retrieve misclassified tweets and order them by scoring function of the HuggingFace model. Then, our Urdu speaker examined the considered tweets and selected 15, that were obviously wrong and which seem not too complex from a NLP standpoint (i.e. we asked the Urdu speaker to imagine those that would be the easier to annotate for a machine). From each of these 15 tweets, we selected 10 similar tweets in our external dataset. The similarity uses the mean of the last layer of tokens computed by the model as the representation of a single tweet and a cosine measure. Subsequently, the Urdu speaker manually annotated 218 of those retrieved tweets, to complement the provided training set.

4.2. Unsupervised MWE extraction

Multiword expressions (MWE) are very likely to carry emotions and to be more reliable clues than single-token words when used as features for text classification. One main issue consists in being able to spot such MWE when they appear in spontaneous text such as tweets. It can be very difficult, costly and time-consuming to keep a lexicon up-to-date with the most recent usage on social media. In a previous work [11], we proposed to rely on unsupervised segmentation based only on raw data to extract salient chunk of text to generate features for a MaxEnt classifier. This solution was proven efficient on Chinese micro-blogging data, but the main algorithm was not specifically restricted to the Chinese script.

We took the opportunity offered by this challenge to apply the algorithm for Urdu language (a very different script) and to evaluate it in a slightly different set-up (3 classes versus 7 classes). Unlike the previous work on Chinese, we strictly limited ourselves to the provided training data and did not extract any emotion lexicon from other sources. The algorithm consists in two steps. Firstly, it extracts autonomous chunks of text based on unsupervised text segmentation [12]. Secondly, it uses these chunks as features for logistic regressions. To meet the multiclass, multioutput design of the task, we use 7 one-versus-others classifiers.

As the Perso-Arabic alphabet differs significantly from the Chinese script, we combined two different segmentation strategies for the first step. One is character-based, it performs segmentation of a sentence into tokens from the sequence of characters (one can think of this procedure as removing spaces from the input and finding them back). The other one processes input as a sequence of tokens and extracts autonomous multitoken expressions. It does not re-segment the input into larger units, but rather yields all MWE candidates.

One benefit from the logistic regression is that it is very straightforward to inspect. It computes weights for the whole vocabulary of features and allows us to check which words and MWE contribute the most to the decision.

[2]The `AutoModelForSequenceClassification` library handles multiclass and task A is multilabel. Due to lack of time, we just selected the latest sentiment in the order of the columns in the dataset to learn models.

4.3. Linguistics-based approach

The linguistic method addresses challenges such as EmoThreat 2022 by thinking about linguistic cues that may be used to solve the issue. For emotion classification, linguistic cues are mainly lexical and affect lexicons, for instance thesauri such as SentiWordNet [13] have been thoroughly used to automatically tag texts with a polarity ('positive', 'negative' and sometimes 'neutral') and/or an emotion. However, to the best of our knowledge, there is no readily available affect lexicon for Urdu yet[3]. Added to the inherent complexity of emotion classification, increased by the fact that task A requires multi-label classification on very short texts, we decided not to use this approach for this task. On the other hand, the linguistics-based approach is more appropriate for Task B, as threats could be defined with lexical cues (i.e., lexical words like گستاخ *insolence* or کٹوانا *to cause to be cut* as well as grammatical words ورنہ *otherwise* and syntactic cues (i.e., typical injunctive forms like verbs in the subjunctive mood or future tense).

The methodology behind this take on Task B consists in using linguistic features as machine learning features. To do so, we first annotated both train and test datasets using Stanza's Universal Dependencies (UD) model[4] for Urdu [15]. Stanza tokenises each tweet and provides properties for each token, namely the lemma, the syntactic head and their relation, inflectional features if any, as well as two parts-of-speech (POS) tags: `upos` from UD, and `xpos` for a more specific POS tagging. All of these features were tested individually to see whether or not they help our models. We then decided to keep only lemmas and xpos as they were the only features that seemed to have a positive impact on our results.

Each word is thus represented as follows: `token`, `token_lemma`, `token_POS`. This method allows us to disambiguate homographs using their lemma or their POS. For instance, مریں گے ضرور مریں گے *They will die, for sure* is presented to our model as a set of 15 elements: (1) مریں, (2) گے, (3) ضرور, (4) مریں, (5) گے, (6) مر_مریں, (7) گے_گے, (8) ضرور_ضرور, (9) مر_مریں, (10) گے_گے, (11) مریں_VM, (12) گے_VAUX, (13) ضرور_JJ, (14) مریں_VM, (15) گے_VAUX.[5]

It is important to note that as any automatic system, Stanza is prone to making errors. Those errors definitely restrain the positive impact our annotation could make, but they are not necessarily a waste. The first reason is that the model only uses the feature it finds useful, and the second reason is linked to the systematic nature of those errors. As long as an error is the same both in the train and the test datasets, it can be used rightfully by the model.

5. Results

HuggingFace models We report performances of HuggingFace models and number of epochs in Figures 1 (task A) and 2 (task B2). Comparing models clearly shows that `roberta`, `muril` obtain better scores than `albert` and `bert`, while comparison between the former models does not show significant differences. The number of epochs doesn't impact much results. We also

[3]It is noteworthy that SentiWordNets for other South Asian languages, namely Hindi, Bengali and Telugu, have been developed [14]

[4]These models are named after the set of treebanks they were trained on, freely available online here: https://universaldependencies.org

[5]Where `VM` stands for main verb, `VAUX` for auxiliary verb and `JJ` for adjective.

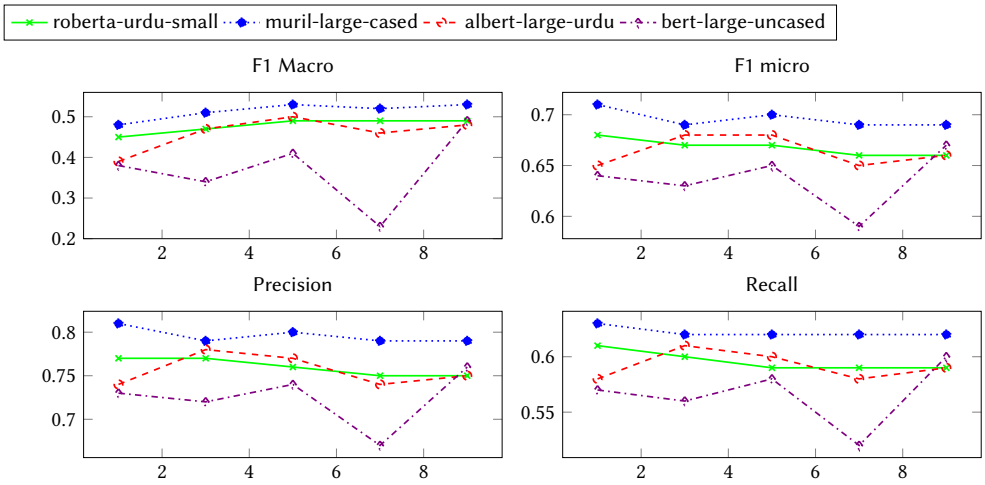

Figure 1: F1 macro (F1M), F1 micro (F1m), Precision and Recall of Models for Task A at different Epochs

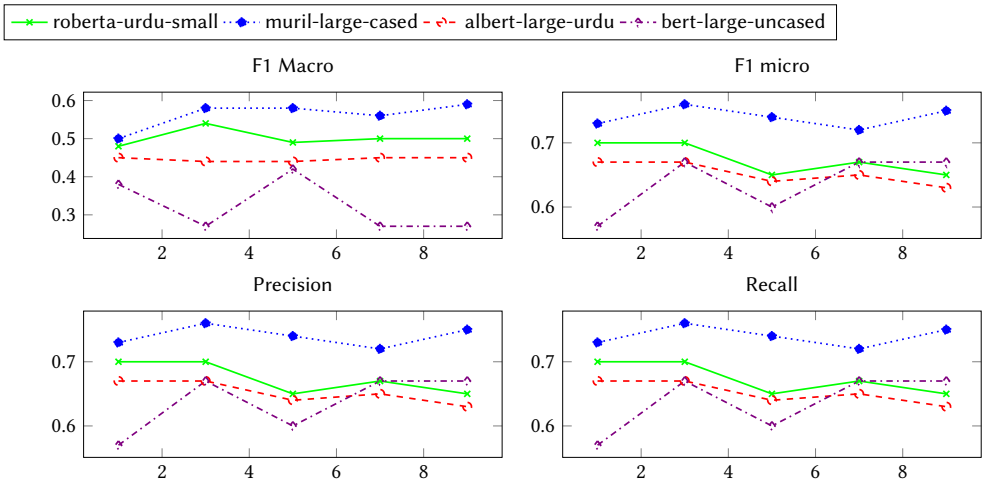

Figure 2: F1 macro (F1M), F1 micro (F1m), Precision and Recall of Models for Task B at different Epochs

note that precision is better than recall on task A, while they are balanced for task B. Our final runs use the `roberta` model with 3 epochs[6].

Results obtained on the train set with cross validation (5 folds) and train / test split are reported in Table 1 and 2. Unsurprisingly, there is a difference between cross-validation and train/test split, which is more important for task B. We also see a great difference between macro (F1M) and micro (F1m) evaluations, which is due to the unbalancing of dataset in terms of categories for both tasks. The official scoring scheme relied on macro scores. For task A, HuggingFace models have a precision much higher than recall, this is most probably due to the fact that we naively convert multilabel dataset to a multiclass one.

[6]Other parameters are left standard with learning rate 2e-5 and weight decay 0.01

Model	5 Folds Train Cross Validation				Train / Test			
	F1M	F1m	P	R	F1M	F1m	P	R
unsup MWE+logit	0.57±0.02	0.70±0.01	0.82±0.01	0.62±0.01	0.60	0.72	0.81	0.64
roberta	0.49±0.01	0.68±0.01	0.78±0.01	0.61±0.01	0.47	0.67	0.77	0.60
muril	0.52±0.01	0.70±0.01	0.80±0.01	0.63±0.01	0.51	0.69	0.79	0.62

Table 1
Detailed Scores of Models for task A

Results for task A comparing models in the train set with cross validation (5 folds) and train / test split 1 reveal a significant difference between F1 macro (F1M) and F1 micro (F1m). We note that `roberta` model outperforms `muril` on F1M scores. Interestingly, the traditional approach based on a logistic regression outperforms transformer-based models. This may be due to our method being trained specifically and only on the task data, creating a model better suited to this dataset. Our team ranked 5th for task A with our MWE method.

Model	5 Folds Train Cross Validation				Train / Test			
	F1M	F1m	P	R	F1M	F1m	P	R
muril	0.64±0.02	0.70±0.02	0.70±0.02	0.70±0.02	0.58	0.76	0.76	0.76
muril+ext	0.61±0.06	0.70±0.03	0.70±0.03	0.70±0.03	0.55	0.75	0.75	0.75
muril+test	0.61±0.05	0.68±0.04	0.68±0.04	0.68±0.04	0.58	0.76	0.76	0.76
roberta	0.60±0.02	0.66±0.02	0.66±0.02	0.66±0.02	0.54	0.70	0.70	0.70
roberta+ext	0.60±0.01	0.68±0.01	0.68±0.01	0.68±0.01	0.41	0.65	0.65	0.65
roberta+test	0.60±0.02	0.66±0.01	0.66±0.01	0.66±0.01	0.54	0.69	0.69	0.69
LogReg	0.54±0.01	0.62±0.01	0.62±0.01	0.62±0.01	0.47	0.66	0.66	0.66
LogReg+ling	0.55±0.00	0.62±0.00	0.62±0.00	0.62±0.00	0.47	0.64	0.64	0.64
LinearSVC	0.54±0.01	0.62±0.01	0.62±0.01	0.62±0.01	0.48	0.69	0.69	0.69
LinearSVC+ling	0.54±0.01	0.64±0.00	0.64±0.00	0.64±0.00	0.48	0.68	0.68	0.68
HingeSGD	0.53±0.01	0.60±0.01	0.60±0.01	0.60±0.01	0.48	0.65	0.65	0.65
HingeSGD+ling	0.53±0.01	0.60±0.01	0.60±0.01	0.60±0.01	0.44	0.62	0.62	0.62

Table 2
Detailed Scores of Models for Task B2

We conducted deeper analysis for task B depicted in Table 2, by comparing models and augmented training datasets, with external data (+ext) and with our manually annotated test set (+test). Globally, the complementary annotations do not provide clear improvements, at least for F1 macro (F1M), where it even lowers scores in some cases. We will examine those results in details to understand how those complementary data are used by models.

Linguistic Approach This approach described in 4.3 was combined with traditional machine learning algorithms using Scikit-Learn 1.1.2. We first used the benchmarking of classifiers proposed by scikit-learn[7]. This allowed us to narrow down the testing algorithms to three: Logistic Regression, LinearSVC and Stochastic Gradient Descent, which results are reported in

[7]https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html.

Table 2. We observe that none of them do better than the BERT models from HuggingFace and using linguistic cues does not increase results. Nonetheless, for the sake of the challenge, we decided to submit a model based on linguistics as the third run for Task B. Our team ranked 3rd for task B1 using the HuggingFace model and 1st for task B2 with the linguistic approach.

6. Error Analyses

Results for both tasks have been reviewed by an Urdu speaker in order to identify error types and try to explain them.

6.1. Task A

In order to review this task's errors, we have analysed the trends of a table displaying the most significant tokens for each of the 6 classes (anger, disgust, fear, sadness, surprise and happiness). Results having a score above 0.5 only have been considered as relevant. Examples are given by descending order of relevance.

anger: Unsurprisingly, we find here occurrences belonging to the lexical field of "hatred" (نفرتانگیز: *hateful*, نفرت : *hatred*, قابلنفرت: *which deserves hate*), adjectives or nouns pejoratively connoted (بےغیرت: *dishonour*, لعنت: *curse*, شدید: *serious*, گھٹیا: *mediocre*, جاهل: *ignorant*, چور: *thief*, کرپٹ : *corrupt*, قتلعام: massacre), then comes the field of justice (انصاف: justice, عدالت: court), followed by names of political persons or movements (عمران: *Imran*, آرایسایس: *RSS*, نوازشریف : *Nawaz Sharif*), country names and professional or religious categories (پٹواری: *notary*, صحافت: *journalism*, قادیانی: *Ahmadi religious community*, مسلمان: *Muslim*). These named entities are subject of criticism in a restrained political and time context, they cannot be used to build a strong model that could be used for other corpora.

A handful of words denoting threat like برداشت *bear*, *support* or باری *turn* are exclusively found in this category. Regarding utterance, the non-polite second personal pronoun combined with ergative case تمنے is very specific of this category.

disgust: a great part of the lexicon stressed in "anger" is also found here: نفرت *hatred*, گھٹیا *mediocre*, قادیانی *Ahmadi*, etc. Some entities and communities are overrepresented here: *army* (فوج), *military* (فوجی), *whole world* (پوریدنیا), *Muslims* (مسلمانوں), *politicians* (سیاستدان) or country names like هندوستان *India*.

fear: We find here occurrences of terms related to "fear", mixed with religion (خداکاخوف: *fear of God*, کےڈرسے : *in fear of*,ڈرھے : *fear is*, خوف: *fright*, اللہتعالی: *God Almighty*. Other themes are: COVID (کروناسے: *from Corona*), Islam (مسلمانوں: *Muslims*), children (بچوںکو: *to the children*), countries (India, China). These occurrences are once again related to specific contexts. Regarding phraseology, several forms of the verb "look like" are recurrent (لگنےلگاھے: *began to look like*, لگاھے/لگرھاھے: *it looks like*). The postposition سے, which indicates the origin and is required to build arguments related to fear, is also very present.

sadness : Related to the lexical field of "sadness" (اداسی: *sadness*, اداس: *sad*, افسوس: *sorrow*,غم : affliction,مجبور : *obliged*,:مایوس *disappointed*), "scorned love" (ٹوٹا: *broken*, محبتسے: *by love*,

broken heart emoji, اکیلے: *lonely*,انجام : *consequence*), and "hatred" (نفرت) Some neutral words are typically linked to this category: شاید(maybe), تعلیم (education), children (بچوں). And more surprisingly: هنسی (laughter), واقعات (event).

surprise : Terms related to surprise (حیرت: *surprise*, بےوجہ: *without any reason*), sometimes denoting some sarcasm (کسخوشیمیں: *in which honor*, واہ : interjection of admiration, تنقید: *criticism*) On the phraseology level, the forms of "know" are overrepresented, especially for the plural present verb (جانتے ہیں), question word کیوں *why*, On the level of utterance, impersonal subjects like آدمی *man* are quite numerous in order to depersonalize the phrase.

happiness : The most salient tokens here are not shared with any other category and make this label quite apart. Emojis are ranked at the top. Then comes the field of "joy" (خوشی: *happiness*, خوشگوار: *happy*, ہاہاہا: *ha ha ha*,) ; some religious formulas and expressions of gratitude are found only here (الحمدللہ : *God be praised*,شکریہ : *thanks*, اللہ: *God*, آمین: *amen*, انشاءاللہ: *if God wants*) ; positive values and adjectives (پیار: *love*, خوبصورت: *beautiful*), Surprisingly, words such as : موت: *death*, orناراض : *angry* are also linked to this category.

Regarding the potential ambiguities, "anger" and "disgust" categories are the closest. Most of significant tokens present in "anger" are also included in "disgust", which is understandable from the context of utterance: tweeters express there grievances against political persons, foreign powers or communities.

Regarding ambiguities in the other categories, the word خوشی *happy* is present in 4 out of 6 classes. Adversely, it is most significant in the "anger" category. "sadness" shares quite a number of tokens with "fear" and to a lesser degree with "surprise". "Surprise" has more shared words with "sadness" and "happiness" appears to be the most singular category as it has the least number of tokens shared with any other category.

6.2. Task B

6.2.1. Overall error reports

Among the 3 runs, **Run-1** (HuggingFace baselines approach) made less errors. Most errors are about the existence or not of a threat: in 124 tweets, a threat was detected whereas there was none. On the opposite, 106 individual threats were not detected.

The confusion between "no threat" and "collective threat" is less (32 tweets) and there are 25 errors of number.

In **Run-2** (HuggingFace baselines approach), there are 27% more errors than for Run-1. The most frequent errors are between "collective" or "individual" threat (167), followed by the confusion between "no threat" and "individual" threat (140). However results are far better here for detecting threat from "individual threat" (only 24 errors)

Run-3 (Linguistic based approach) produced exactly as many errors as Run-2 (HuggingFace baselines approach). It outperformed the former regarding errors between "individual threat" and "no threat" but tagged more often tweets without threat as "individual threats". It was better at distinguishing between "collective threat" and "no threat".

nb errors	true label	tagged label
124	2	0
106	0	2
23	1	0
20	1	2
12	2	1
2	0	1
264		

Table 3
Errors for Task B Run-1

nb errors	true label	tagged label
187	0	2
80	2	0
41	1	2
19	2	1
7	1	0
1	0	1
335		

Table 4
Errors for Task B Run-2

nb errors	true label	tagged label
140	2	0
78	0	2
61	2	1
25	1	0
24	1	2
7	0	1
335		

Table 5
Errors for Task B Run-3

6.2.2. Typology of errors

For this task, errors have been identified on 3 levels: lexical, phrasal and deictic. Some elements were at first considered as core features to identify threatening tweets, but turned out not to be useful as they were present in both classes. Regarding the verbal system, we noticed that second person address, subjunctive mode for jussive phrases and future tense to express consequence were more employed in threat tweets. On a lexical level, moderate abusive language and terms related to threat (نشانہ بنانا *to target*, سزا دینا *to punish*, یاد رکھو *remember!*) are present in both classes, even if they are more numerous in the "threat" one. Besides, no MWE nor phrase expression seems to characterize the "non threat" class.

Conversely, the errors review revealed some elements belonging exclusively to the "threat" category: 1) multiple clause sentences in which a first clause exposes an order and a second a future consequence, both clauses being linked by ورنہ *otherwise*; 2) MWE using برداشت *bear* co-occurring with a negative element (برداش نہیں *don't bear*, ناقابلِ برداشت *unbearable*); 3) other significant MWE as الٹی گنتی *countdown*, وقت آیا ہے کہ *time has come to*, سر کاٹے *cut heads*.

A comparison of the 3 train sub-corpora ("no threat", "individual threat" and "collective threat") with the textometry software TXM [16] corroborates this analysis. First, we searched out the specific terms for each sub-corpus, and found out that morphemes of future (گا, گے) were very specific (respective scores: 37.3 and 16.3) of the **"individual threat"** sub-corpus, as well as the injunctive form of the verb "to die" (مریں *may they/you die*) and first and second person pronouns. Other semantic terms detected in the test set scored very poorly (برداشت *bear*, واجب *obligatory*, وقت *time*), however the study of their cooccurrents leads to interesting results for classification. For the "collective threat" class, the first most specific terms have far smaller scores: بکواس *bullshit* (8.3), بند *closed* (7.9), خبیث *mean* (7.2), کتے *dog* (6.8), سر *head* (5.1). Abusive words are not surprisingly common to both categories. The differences with the test data can be explained by the lack of some terms in the train set.

7. Conclusion

Our participation in the EmoThreat 2022 challenge about sentiment and threats in Urdu tweets was the opportunity to test three different approaches to address classification tasks: Hugging-

Face models based on BERT, unsupervised MWE extraction and linguistics-based approach. It turned out that unsupervised MWE extraction models obtain better scores for task A (6 classes of emotions), with much higher precision than recall. Still, HuggingFace models dedicated to Urdu language (`roberta`, `muril`) obtained satisfying results. For task B, the HuggingFace obtain better results for sub-task B1 (binary detection of threats) while the linguistics-based approached appeared to be the best one in the sub-task B2 (detection of individual/collective threat). Considering results for both tasks leads to the conclusion that HuggingFace models do not outperform other methods. Our error analysis identified numerous lines of research that we intend to explore in the future.

References

[1] N. Zameen, G. Sayeed, Sentiment analysis on urdu tweets using markov chains, SN Computer Science (2020).

[2] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE Access 9 (2021) 97803–97812.

[3] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, Multi-class sentiment analysis of urdu text using multilingual bert, Scientific Reports 12 (2022) 1–17.

[4] P. A. Muhammad, J. Zheng, I. R. Naqvi, A. Mohammed, T. S. Muhammad, Automatic detection of offensive language for urdu and roman urdu, IEEE Access XX (2017).

[5] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: HT '22: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, 2022. URL: https://doi.org/10.1145/3511095.3531277.

[6] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detection and target identification in urdu tweets, IEEE Access 9 (2021) 128302–128313.

[7] R. Batra, Z. Kastrati, A. S. Imran, S. M. Daudpota, A. Ghafoor, A large scale tweet dataset for urdu text sentiment analysis, 2020. doi:10.17632/rz3xg97rm5.1.

[8] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, MuRIL: Multilingual Representations for Indian Languages, 2021. arXiv:2103.10730.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] M. Y. Khan, M. S. Nizami, Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis., in: 2020 IEEE 2nd International Conference On Information Science & Communication Technology (ICISCT), IEEE, 2020.

[11] P. Magistry, S.-K. Hsieh, Y.-Y. Chang, Sentiment detection in micro-blogs using unsupervised chunk extraction, Lingua Sinica 2 (2016) 1–10.

[12] P. Magistry, Unsupervised word segmentation and wordhood assessment: the case for Mandarin Chinese, Ph.D. thesis, Paris 7, 2013.

[13] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource

for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.
[14] A. Das, S. Bandyopadhyay, SentiWordNet for Indian Languages, in: Proceedings of the eighth workshop on Asian language resources, 2010, pp. 56–63.
[15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.
[16] S. Heiden, J.-P. Magué, B. Pincemin, TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement, in: JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data, 2010. URL: http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf.
[17] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, UrduThreat@ FIRE2021: Shared track on abusive threat identification in Urdu, in: Forum for Information Retrieval Evaluation, 2021, pp. 9–11.
[18] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, Overview of the shared task on threatening and abusive detection in Urdu at FIRE 2021, in: FIRE (Working Notes), CEUR Workshop Proceedings, 2021.
[19] N. Ashraf, A. Rafiq, S. Butt, H. M. F. Shehzad, G. Sidorov, A. Gelbukh, Youtube based religious hate speech and extremism detection dataset with machine learning baselines, Journal of Intelligent & Fuzzy Systems (2022) 1–9.
[20] N. Ashraf, R. Mustafa, G. Sidorov, A. Gelbukh, Individual vs. group violent threats classification in online discussions, in: Companion Proceedings of the Web Conference 2020, 2020, pp. 629–633.
[21] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022, in: CEUR Workshop Proceedings, 2022.
[22] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, EmoThreat@FIRE2022: Shared Track on Emotions and Threat Detection in Urdu, in: Forum for Information Retrieval Evaluation, FIRE 2022, Association for Computing Machinery, New York, NY, USA, 2022.
[23] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label emotion classification of urdu tweets, PeerJ Computer Science 8 (2022) e896.
[24] I. Ameer, N. Ashraf, G. Sidorov, H. Gómez Adorno, Multi-label emotion classification using content-based features in twitter, Computación y Sistemas 24 (2020) 1159–1164.