2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts

Oyesh Mann Singh
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County (UMBC)
Baltimore, Maryland, USA
osingh1@umbc.edu

Sandesh Timilsina
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County (UMBC)
Baltimore, Maryland, USA
stimilsina@umbc.edu

Bal Krishna Bal
Information and Language Processing Research Lab (ILPRL)
Department of Computer Science and Engineering
Kathmandu University
Kathmandu, Nepal
bal@ku.edu.np

Anupam Joshi
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County (UMBC)
Baltimore, Maryland, USA
joshi@umbc.edu

***Abstract*—With the increase in internet access and the ease of writing comments in the Nepali language, fine-grained sentiment analysis of social media comments is becoming more and more pertinent. There are a number of benchmarked datasets for high-resource languages (English, French, and German) in specific domains like restaurants, hotels or electronic goods but not in low-resource languages like Nepali. In this paper, we present our work to create a dataset for the targeted aspect-based sentiment analysis in the social media domain, set up a dataset benchmark and evaluate using various machine learning models. The dataset comprises of code-mixed and code-switched comments extracted from Nepali YouTube videos. We present convincing baselines using a multilingual BERT model for the Aspect Term Extraction task and BiLSTM model for the Sentiment Classification Task achieving 57.978% and 81.60% F1 score respectively.**

***Index Terms*—aspect based, abusive sentiment analysis, Nepali, natural language processing, YouTube, social media**

I. Introduction

Sentiment Analysis is one of the hot topics in Natural Language Processing field. It is one of the important tool responsible for directly or indirectly impacting lives of people in the online community and can impact the growth/sales of various products and businesses. Early researchers [1], [2] classify the sentiment in a sentence level, be it a comment or a text-piece. However, there is a growing trend of research in Aspect Based Sentiment Analysis (ABSA) for product reviews [3]–[5] to identify the positive and negative sentiment of various aspects in the given text related to the restaurant and electronics product domain. Targeted Aspect Based Sentiment Analysis (TABSA) [6], [7] is a variation of ABSA where the task is to identify the sentiment of various aspects based on its target in a given unit of text. Similarly, there are a notable research in low-resource languages [8]–[10] to identify the aspects and its associated polarity in the texts.

Nepali language is written in Devanagari script, so it is similar in some way to Hindi (also written in Devanagari script) language in aspects like the vocabulary, the grammar, or in spoken form. However, Hindi language differ with Nepali in terms of the morphology, language-specific grammar rules, vocabulary and semantics. Therefore, in this project, we propose our sentiment analysis model trained specifically for the Nepali language.

In this paper, we discuss the targeted aspect based sentiment analysis dataset based on social media texts in the Nepali language. We also discuss about the data collection pipeline, annotation schema and machine learning techniques to identify the aspect terms, targeted terms and the polarity based on the aspect category. Unlike previous works in this field, our dataset contains the comments extracted from popular Nepali YouTube videos under the News & Politics category. We annotated target terms as Named Entities (PER, ORG, LOC and MISC) as users are most likely to comment on political figures or organizations in videos under News & Politics category.

Our goal of this project is to setup a TABSA dataset benchmark and analyze various machine learning techniques to identify the aspect terms/category, its sentiment polarity along with the target term/category for the Nepali language texts in the social media domain.

Our contributions in this paper are as follows:

- We contribute to the development of annotated data resources in the Nepali language by publicly releasing the dataset developed as part of this project.
- We propose a dataset benchmark to identify abusive sentiment terms and classify polarity in the social media domain.
- We provide the dataset baselines based on multilingual BERT and BiLSTM models and also analyze the

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

results.

II. Related Work

Sentiment analysis has become an integral part of our digital lives with its applications touching almost all NLP domains from social media to finance and markets. In the earlier research, sentiment analysis was mainly focused on identifying the polarity of the whole text blob and fine-grained subjectivity was not given much attention. As a result, the deep semantic knowledge within a text remained largely unanalyzed.

To address such issues, there is a growing interest in aspect-based sentiment analysis to identify the polarity specific phrases within texts based on the predefined aspect categories. [11], [12] introduced a technique to capture the contextual polarity of specific phrases or expressions along with its sources and target spans. Furthermore, [13] successfully demonstrated a two-stage technique to annotate target terms and its corresponding sentiment polarity in subjective opinions in Arabic although it is morphologically challenging. [3]–[5] provided an eminent framework to annotate aspect term, aspect category, opinion target expression and sentiment polarity in a given text. The texts consists of customer reviews from laptops and hotels domain. The task amassed various techniques from many participants to extract rich representations for fine-grained sentiment analysis.

To analyze the sentiment of the text based on the targets as well, a new scheme called targeted aspect-based sentiment analysis was introduced. [14]–[17] demonstrated the importance of target entities for a sentiment classification task. Similarly, [7] created a targeted aspect-based sentiment dataset, based on the reviews of neighborhood, to predict the aspect and corresponding sentiment polarity for each location entity in a given sentence.

Similarly, [8]–[10] created an aspect-based sentiment analysis dataset in Hindi, Bangla, and Indonesian languages respectively in various domains like electronic products, sports, and restaurant.

However, there is no such work explored on the aspect-based sentiment analysis in the Nepali language. [18] developed a Nepali Sentiment Corpus which is a collection of sentences from the News domain with binary-level annotation whether it is subjective or objective. They also developed Nepali SentiWordNet called Bhavanakos, which is a translated version of English SentiWordNet. They identify a sentence as being subjective or not based on the sentiment words in the text using the Nepali SentiWordNet. Since we cannot perfectly estimate whether a sentence is subjective or objective merely based on words, we need a deeper analysis on phrase-level, because the same word might carry different sentiment based on different contexts.

[19] created a Nepali sentiment corpus from social media reviews on books and movies. They performed experiments on a document-level to classify whether it is positive or negative. On the phrase-level annotation, they annotate opinion expression terms, opinion targets and opinion holders but not much description is provided.

III. NepSA Dataset

Nepali Sentiment Analysis Dataset[1] (NepSA) is a named targeted aspect-based sentiment analysis dataset. We collected the comments from the most popular Nepali YouTube channels having the highest subscribers under the News & Politics category. The dataset consists of 3068 comments extracted from 37 different YouTube videos of 9 different YouTube channels. We used binary sentiment polarity schema and divided the comments into 6 aspect categories *General*, *Profanity*, *Violence*, *Feedback*, *Sarcasm* and *Out-of-scope* to annotate the data. All the targeted annotations are created considering the target entity towards which the sentiment is expressed and not on the general understanding of the sentence. The target entities are divided mainly into *Person*, *Organization*, *Location* and *Miscellaneous*.

A. Dataset Preparation & Preprocessing

Initially, we went through some of the Nepali YouTube videos and its comments manually. We found out many cases where abusive terms along with decent terms were among the top comments on YouTube. Therefore, we decided to create an ABSA dataset for social media in the Nepali language to analyze the comments on the granular level.

We first listed out ten popular Nepali "News & Politics" YouTube channels with the highest number of subscribers. Then we filtered out the top 10 videos which were released in the year 2019 from each channel. The selection was based on the number of views on the videos as of September 2019. We collected up to top 100 comments from each of those videos based on the number of *Likes* on the comments along with the following constraints:

- The comment should have at least one Devanagari character.
- The comment should have at least 5 words and at max 50 words.
- The comments might contain native (Nepali), code-mixed (Romanized) and code-switched (English + Nepali) terms in a single sentence.
- Emojis were removed because BRAT [20] could not handle it effectively during annotation.

After collecting all the comments from YouTube, we lemmatize using [21] which follows a rule-based approach. Since this lemmatizer was developed very early, it was not able to lemmatize every word perfectly. Due to language evolution and gradual change in political scenario, people in Nepal have been coining various terms like झोले, मण्डले . This lemmatizer over-lemmatizes in such cases, separating

[1]The code and dataset is available at https://github.com/oya163/nepali-sentiment-analysis

झो and ले . Therefore, during our annotation process, we join together such words manually. No any special processing is done for code-mixed or code-switched terms. However, all repetitive punctuations were reduced to unit count.

B. Aspect Categories

We annotate any terms in a comment which lies under *General*, *Profanity*, *Violence*, *Feedback*, *Sarcasm* and *Out-of-scope* categories based on the context of the video. We assume our aspect term to be a noun phrase. The definition of each aspect category is explained clearly in the annotation guidelines[2] for the annotators. The examples provided might contain spellings mistakes as they are directly taken from social media comments. The distribution of these aspect terms are presented in table III.

C. Target Categories

We assume that an abusive comment in social media is either targeted or untargeted. We tagged target entities as *Person*, *Organization*, *Location* and *Miscellaneous*. In this dataset, target entities are annotated similar to Nepali NER dataset [22] and based on the annotation guidelines provided in CoNLL 2003 [23]. The statistic for the target categories is presented in table II

D. Annotation Procedure

We used the BRAT [20] annotation tool to annotate the dataset. Annotators first began by reading the annotation guidelines and examples. Each annotator was then required to annotate a small subset of the data. After completion, an inter-annotator agreement was calculated and disagreements were discussed. This procedure was repeated with gradual changes in the guidelines until a reasonable agreement was reached.

After a reasonable agreement was reached, each annotator started to annotate the remaining dataset equally. The comments were annotated at a sentence level. To find the underlying entity-aspect relationship, the annotators were asked to determine the aspect terms and entity terms if present and identify the relationship between them. Phrase-level annotation was performed only for *General*, *Profanity*, *Violence* and *Feedback* categories. Each aspect term was also given a polarity value to determine if it is positive/neutral or negative. Therefore, the dataset can be divided into two schemas, fine-grain and coarse-grain, its example is presented in figures 1 and 2. As seen on the figure 2, only the end of the sentence is tagged with aspect category which is equivalent to tagging the whole sentence.

Translation: *I loved the talk by the one wearing Red Coat but I think another stupid guy is a traitor.*

[2]https://github.com/oya163/nepali-sentiment-analysis/tree/master/guidelines

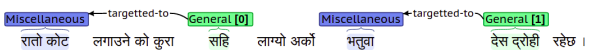

Figure 1. Annotation sample under fine-grained schema

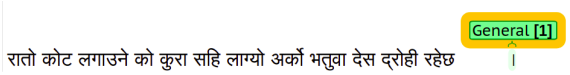

Figure 2. Annotation sample under coarse-grained schema

E. Inter-annotator agreement

15% of the total dataset was annotated by two annotators whose native language is Nepali and the remaining were annotated individually on 50/50 proportion. We used the F1 score to measure the pairwise agreement between the two annotators for this task. The overall pair-wise inter-annotator agreement was found to be 0.703 when considering all aspect categories and target entities for instance based matching[3]. The agreements on different aspect categories varied, with some having higher agreement score, which is presented in table I. The F1 on *Organization* is 0 because there was no instance of it in the pair-wise annotated set. In the table I instance-level column represent exact word matching while token-level column represent subset word matching. The total mean F1 for Polarity identification is 0.602, calculated based on its associated aspect category. The total mean F1 score for relationship classification is 0.605, calculated based on whether aspect category is targeted or untargeted, if targeted we take its target entity into account as well.

IV. Corpus Analysis

From this dataset, we see that there is a diversity in terms of the use of language in social media platforms. We see that there is a growing use of code-mixed, code-switched and transliterated comments in the Nepali language based on their convenience. In our dataset, the ratio

[3]https://github.com/kldtz/bratiaa

Labels	Instance	Token	Polarity	Relation
Feedback	0.154	0.486	0.308	-
General	0.33	0.442	0.609	-
Location	0.444	0.462	-	-
Misc	0.4	0.432	-	-
Org	0	0	-	-
Outofscope	0.732	0.732	-	-
Person	0.987	0.993	-	-
Profanity	0.705	0.635	0.636	-
Sarcasm	0.4	0.211	-	-
Violence	0.55	0.689	0.6	-
Targeted	-	-	-	0.652
Untargeted	-	-	-	0.486
Total	**0.71**	**0.725**	**0.602**	**0.605**

Table I
Pair-wise F1 measurement for aspect term extraction based on instance & token matching; sentiment polarity identification and relationship identification

of English words to Nepali words = 0.0185. We also see the maximum usage of new words such as: झोले *(sycophant)*, मण्डले *(bootlicker)*. These words do not exist in the Nepali dictionary but have evolved with time and change in the political scenario. The sentences are unstructured and there are a lot of spelling mistakes which, at times, makes it difficult to interpret.

The dataset statistics show that *General* negative sentiment is more dominant among all the aspect categories. Similarly, mild *Profane* and positive *Feedback* terms are also widely used. It is interesting to see that the least number of comments were related to *Violent* sentiment. The length of the aspect terms on average is 2 words with most of the aspect words tagged are less than 5 words. 41% of the aspect terms are unigram whereas 36% and 15% are bigrams and trigrams respectively. Most of the unigrams are in *General* and *Profanity* categories whereas bigrams and trigrams are more in *Violence* and *Feedback* categories. Some of the most frequent words in comments are as follows :

- General : सलाम , झोले, चोर, दलाल, भ्रष्ट
  Translation: salute, sycophant, thief, broker, corrupt
- Profanity : चोर, साला, खाते, कुकुर
  Translation: thief, moron, homeless, dog,
- Violence : हत्यारा, यातना, गोलि हानी, मार्ने, कुटपिट
  Translation: killer, torture, shoot, kill, beat,
- Feedback : गर्नु पर्छ, बनाउनु पर्छ, जेल हाल्नु पर्छ, बोल्न देउ, हुन देउ
  Translation: should do, should make, should be jailed, let him/her speak, let it be done.

Similarly, in terms of target, most of the comments are directed towards an individual rather than a group or organization. These individuals are mostly political figures, journalists and panelists shown in the video. In most of the cases, the sentiment expressed towards these individuals was negative with mild profane terms whereas some TV journalists (Ravi Lamichhane, Sushil Pandey) and figures (Rabindra Mishra, Kulman Ghising) had positive comments. Overall, a large number of comments expressed negative sentiment towards political leaders. The dataset consists of 2136 single-targeted, 705 multi-targeted and 2491 untargeted instances of aspect terms. Single-targeted means each aspect term is associated with only one target entity, multi-targeted means each aspect term is associated with multiple target entity and untargeted means an aspect term does not have any association to the target entity.

V. Dataset Statistics

NepSA dataset is categorized into two major schema, fine-grained and coarse-grained schema. Under the fine-grained schema, as shown in figure 1, we tag specific words/phrases which expresses sentiment falling under a category defined in subsection III-B. There are many cases where multiple aspects are present in a sentence targeted to multiple target entities.

Target entities	Count
Person	2327
Location	432
Organization	300
Miscellaneous	1235

Table II
Target entities statistics under fine-grained schema

Polarity	General	Profanity	Violence	Feedback
0	1203	344	122	447
1	1971	120	190	84
Total	3174	464	312	531

Table III
Aspect terms statistics under fine-grained schema

Under fine-grained schema, we tag only four major categories *General*, *Profanity*, *Violence* and *Feedback* whereas under coarse-grained *Sarcasm* and *Out-of-scope* is also tagged. Fine-grained data schema is used for sequence labelling task to extract aspect terms and target terms whereas coarse-grained data is used for classification task to classify sentiment polarity and relationship between target and aspect term.

Table III shows there are more negative *General* terms. We can infer that people are expressing criticism with *General* negative terms in News & Politics videos showing that they are unsatisfied mostly with political leaders. Moreover, the dataset has more mild *Profane* terms, more extremely *Violent* terms and more positive *Feedback* terms.

Table IV presents the total count of positive and negative sentences categorized under various aspects. Although, 5135 total sentences were tagged, we only used *General*, *Profanity*, *Violence* and *Feedback* category for our experiments bringing down the total count of sentences to 4035.

Aspects	0	1	Total
General	1052	1783	2835
Profanity	302	105	407
Violence	114	171	285
Feedback	426	82	508
Sarcasm	-	-	166
Out-of-scope	-	-	934
Total	1894	2141	5135

Table IV
Total count of sentences under coarse-grained schema

VI. Experiments

We start our experiment by training BiLSTM+CRF [24], [25] for aspect term extraction task and BiLSTM [26] & CNN for sentiment polarity classification task. We split the total dataset into 80%, 10%, 10% for train, test and dev set and perform 5-fold cross-validation for both of the tasks. For the sentiment classification task, the dataset is split grouped by aspect. The training is stopped if the validation loss does not decrease after 5 epochs. We train a 300-dimension skip-gram fasttext word embeddings using gensim [27]. We train two different types of embeddings,

Monolingual and Multilingual. Monolingual embedding is trained on Nepali texts collected only from Nepali National Corpus (NNC) [28]. Multilingual embedding is trained on text collected from NNC + Nepali OSCAR [29] + English texts extracted[4] from latest 169899 articles from Wikipedia dump[5]. This is particularly helpful since our dataset contains code-switched comments.

Embeds	**Multilingual**			**Monolingual**		
Dataset	**P**	**R**	**F1**	**P**	**R**	**F1**
Overall	0.607	0.539	**0.571**	0.640	0.509	0.567
Target	0.790	0.800	**0.794**	0.848	0.747	**0.794**
Aspect	0.450	0.394	**0.419**	0.473	0.358	0.407

Table V
PRECISION, RECALL AND F1 SCORE FOR ASPECT TERM EXTRACTION TASK BASED ON MONOLINGUAL AND MULTILINGUAL EMBEDDINGS USING BILSTM+CRF MODEL. THIS TABLE SHOWS THE SIGNIFICANCE OF USING MULTILINGUAL EMBEDDING FOR SOCIAL MEDIA DATA.

We use pre-trained multilingual BERT [30] to fine-tune on our dataset running the experiment for 4 epochs. We use huggingface [31] framework for both of the tasks. This BERT architecture has 12 attention heads, 12 hidden layers, and 768 hidden size. We use Adam algorithm with weight decay fix for optimization with learning rate at 5e-5 and epsilon at 1e-8.

Model	**P**	**R**	**F1**
BiLSTM+CRF	60.70	53.95	57.07
BERT	58.14	57.88	**57.98**

Table VI
MODEL WISE COMPARISON FOR ASPECT TERM EXTRACTION TASK

A. Aspect Term Extraction

This task resembles the sequence labelling task where we tag each token of a given sentence with predefined aspect category or named entities. We experiment with four major categories *General*, *Profanity*, *Violence* and *Feedback* under Aspect Category and *Person*, *Organization*, *Location* and *Miscellaneous* under Target Entities. Table V represents the scores from aspect term experiment using BiLSTM+CRF model with various approaches in dataset, *Overall* represents the dataset consisting of both Aspect Category and Target Entity terms, *Target* represents dataset containing only target entities without the labels from Aspect Category and *Aspect* represents dataset containing only aspect categories without the labels from Target Entities.

We experiment with this task using BiLSTM+CRF model implemented using PyTorch [32] with the help of torchnlp[6]. We use a batch size of 8 with the embedding size of word and character of 300 and 30 respectively along with the CRF for joint decoding [24], [25]. We use LSTM of hidden size 100 which is randomly initialized and a dropout rate of 0.5 is used on initial word embeddings only. Unidirectional LSTM is used for character embeddings. We use Adam optimizer to learn the parameters with 0.05 initial learning rate.

For this task, the BERT model performed just slightly better compared to BiLSTM+CRF, however, we believe training BERT from scratch based on English and Nepali corpus will help to perform much better rather than on pre-trained multilingual BERT.

B. Sentiment Polarity Classification

We train basic SVM, CNN, BiLSTM and BERT models to classify sentiment polarity [0, 1] of each aspect categories in every sentence. The dataset statistics used for this task is shown in table IV. Since one sentence can have multiple aspect categories, this dataset can have the same sentence but with different aspect terms and categories with its corresponding sentiment polarity. Table VIII shows the experiment between the *Concatenated* vs *Not concatenated* embeddings. *Concatenated* embeddings is the concatenation of feature vector of words from sentence and aspect terms, while *Not concatenated* represents only the embeddings of sentence to identify the sentiment polarity. However, aspect category as a word is always concatenated with the word embeddings from sentence since the sentiment polarity is associated with aspect category.

We use a batch size of 8 and Adam optimizer to learn the parameters with 0.05 initial learning rate for both CNN and BiLSTM. For CNN training, we use 100 filters of sizes 3, 4 and 5. For BiLSTM training, we use LSTM of hidden size 256 and a dropout rate of 0.5 on both LSTM cells and the hidden state representation.

Similarly, we fine-tune multilingual BERT for sentiment classification task. A slight difference in this task is that we use the LSTM layer on top of regular BERT embedding. And, for SVM training, we train linear kernel with CountVectorizer to vectorize the text. N-grams consists of combination of unigram and bigram word vectors as a feature. SVM performed relatively good score with n-grams as a feature.

Overall, the models trained on concatenated embeddings performed better compared to the non-concatenated version in terms of accuracy and F1 score. This might be due to the extra information received from aspect terms in the concatenated version. We can imply that concatenating embeddings is beneficial in achieving a better score. However, for this task, the BiLSTM model performed better compared to the multilingual BERT model.

VII. DISCUSSION

All of the scores presented are an average score from 5-fold cross-validation experiments. From table V, we observe low F1 score on aspect term extraction compared to target entity recognition because aspect terms are

[4]https://github.com/attardi/wikiextractor
[5]https://dumps.wikimedia.org/enwiki/20200301/
[6]https://github.com/kolloldas/torchnlp/

Tasks	Aspect Term Extraction			Sentiment Classification			
Aspects	**P**	**R**	**F1**	**P**	**R**	**F1**	**Acc**
Feedback	32.44	34.46	33.19	75.30	82.40	76.70	82.40
General	41.74	42.17	41.86	84.80	84.90	84.70	84.90
Location	82.78	77.78	79.37	-	-	-	-
Miscellaneous	63.97	51.87	56.87	-	-	-	-
Organization	66.63	69.58	66.78	-	-	-	-
Person	85.93	90.24	88.00	-	-	-	-
Profanity	56.45	48.86	52.11	73.00	74.60	73.50	74.60
Violence	33.58	39.59	36.20	57.70	58.50	56.50	58.50
Total	58.14	57.88	**57.98**	80.50	81.60	80.60	**81.60**

Table VII
ASPECT-WISE F1 SCORE FOR ASPECT TERM EXTRACTION AND WEIGHTED F1 SCORE FOR SENTIMENT CLASSIFICATION TASK

Features	Not concatenated				Concatenated			
Model	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
BERT	0.789	0.782	0.779	0.782	0.800	0.804	0.800	0.799
BiLSTM	0.805	0.806	0.805	0.805	0.815	0.816	0.816	**0.816**
CNN	0.787	0.788	0.788	0.787	0.811	0.812	0.811	0.811
SVM	0.684	0.691	0.684	0.678	0.714	0.716	0.714	0.712

Table VIII
ACCURACY AND WEIGHTED F1 SCORE IN SENTIMENT POLARITY CLASSIFICATION OF A GIVEN TEXT WHEN CONCATENATED (TEXT + ASPECT CATEGORY + ASPECT TERM) VS NOT CONCATENATED (TEXT + ASPECT CATEGORY)

Sentence	Aspect Term	Aspect Category	Sentiment Polarity
डा. सुरेन्द्र के.सि र रमेश खरेल सर हरूलाई पनि राख्नु पर्यो *You should bring Dr. Surendra KC and Ramesh Kharel*	राख्नु पर्यो should bring	FEEDBACK	0
रबि लामिछाने नेपालि जन्ता को हिरो हुन *Rabi Lamichhane is a Hero for Nepali people*	हिरो हुन Hero	GENERAL	0
भर्स्ट जति जेलमै ठोक्नु पर्छ ग्यनेन्द्र शाहिको समर्थ छ *Corrupt people should be punished in jail, we support Gyanendra Shahi*	ठोक्नु पर्छ should be punished	VIOLENCE	1

Table IX
FEW EXAMPLES OF END-TO-END INFERENCE USING OUR BEST MODEL FROM SEQUENCE LABELLING AND CLASSIFICATION TASK. ASPECT TERM AND ASPECT CATEGORY ARE IDENTIFIED BY ASPECT TERM EXTRACTION MODEL, WHICH IS THEN COMBINED AND FED INTO SENTIMENT CLASSIFICATION MODEL.

highly subjective and annotators had difficulty agreeing on the boundary of a noun phrase.

Additionally, from table VII, we can imply that since *General* and *Feedback* category is very ambiguous especially in their positive cases, they both have relatively lower score. Among the aspect term category, *Profanity* has a comparatively higher score because the majority of the profane terms are unigrams and less confusing but *Violence* has the lowest score because of thin line in noun phrase boundary during annotation.

Similarly, we see a low score on *Miscellaneous* category which is ambiguous in nature as we did not fix the type of common noun that fits under this category. Example: नेता, पत्रकार, राजा Translation: *Politician, Journalist, King* . Additionally, we tagged only such *Miscellaneous* target terms that have a direct relationship with aspect terms, which means all *Miscellaneous* should be targeted entities. We can see a relatively higher score for *Location* and *Person* as they are relatively less ambiguous for annotators and contains many overlapping samples.

Moreover, as the sentiment polarity is not associated with any target entity terms, we have not shown its score in table VII. The sentiment score shown in table VII is the weighted average F1 score of *0* and *1* of each aspect categories. We see higher sentiment score for *General* because of the higher number of unigrams and bigrams creating a strong feature vector when concatenated.

We found that a more number of false positives or negatives are in *General* category. Overall, this is especially caused by the high volume of unknown words. The unknown words are mainly due to grammatical errors, spelling errors, transliterated words, syntactical error and neologisms in social media data.

VIII. Conclusion

This paper mainly focuses on the creation of a new Targeted Aspect Based Abusive Sentiment Analysis dataset from social media data in the Nepali language. This is the first TABSA dataset in the Nepali language to be released publicly. We found the agreement score to be high for target entity identification task but less for aspect term identification task because of it being highly subjective. However, it can be dealt with strict guidelines with the help of Nepali linguists. We found that there were lots of grammatical and syntactical errors in social media data which is another reason for low score because

the word embeddings were trained on Wikipedia data and National Corpus (books, magazines and online news portal), where such errors are negligible. Despite the annotation challenges, we provide promising baselines using BERT (multilingual) for the Aspect Term Extraction task and BiLSTM for Sentiment Classification Task achieving 57.978% and 81.60% F1 score respectively. We plan to use POS tags and LASER embeddings[7] as a future work. We plan to extend this annotation schema to other languages in South Asia.

Acknowledgment

We would like to express sincere thanks to UMBC CARTA lab[8] for providing us NVIDIA Tesla P100 GPU on IBM Minsky server and also many thanks to UMBC High Performance Computing Facility[9] for providing us NVIDIA Tesla V100 GPU to perform our experiments.

References

[1] S. R. Das and M. Y. Chen, "Yahoo! for amazon: Sentiment extraction from small talk on the web," *Management Science*, vol. 53, no. 9, pp. 1375–1388, 2007. [Online]. Available: http://www.jstor.org/stable/20122297

[2] C. P. Gupta and B. K. Bal, "Detecting sentiment in nepali texts: A bootstrap approach for sentiment analysis of texts in the nepali language," in *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*, March 2015, pp. 1–4.

[3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004

[4] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 486–495.

[5] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.

[6] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2014, pp. 49–54.

[7] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1546–1556. [Online]. Available: https://www.aclweb.org/anthology/C16-1146

[8] M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Aspect based sentiment analysis in Hindi: Resource creation and evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2703–2709. [Online]. Available: https://www.aclweb.org/anthology/L16-1429

[7] https://engineering.fb.com/ai-research/laser-multilingual-sentence-embeddings/

[8] https://carta.umbc.edu/

[9] https://hpcf.umbc.edu/

[9] M. A. Rahman and E. K. Dey, "Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation," *Data*, vol. 3, p. 15, 2018.

[10] D. Ekawati and M. L. Khodra, "Aspect-based sentiment analysis for indonesian restaurant reviews," in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, Aug 2017, pp. 1–6.

[11] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 347–354.

[12] T. A. Wilson, "Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states," Ph.D. dissertation, University of Pittsburgh, 2008.

[13] N. Farra, K. McKeown, and N. Habash, "Annotating targets of opinions in Arabic using crowdsourcing," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 89–98. [Online]. Available: https://www.aclweb.org/anthology/W15-3210

[14] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 151–160.

[15] M. Mitchell, J. Aguilar, T. Wilson, and B. Van Durme, "Open domain targeted sentiment," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1643–1654. [Online]. Available: https://www.aclweb.org/anthology/D13-1171

[16] D.-T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[17] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 3087–3093.

[18] C. P. Gupta and B. K. Bal, "Detecting sentiment in nepali texts: A bootstrap approach for sentiment analysis of texts in the nepali language," in *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*. IEEE, 2015, pp. 1–4.

[19] L. B. R. Thapa and B. K. Bal, "Classifying sentiments in nepali subjective texts," in *2016 7th International conference on information, intelligence, systems & applications (IISA)*. IEEE, 2016, pp. 1–6.

[20] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "brat: a web-based tool for NLP-assisted text annotation," in *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association for Computational Linguistics, April 2012.

[21] B. K. Bal and P. Shrestha, "A morphological analyzer and a stemmer for nepali," 2004.

[22] O. M. Singh, A. Padia, and A. Joshi, "Named entity recognition for nepali language," in *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, Dec 2019, pp. 184–190.

[23] E. F. Tjong Kim Sang and F. De Meulder, "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 - Volume 4*, ser. CONLL '03. USA: Association for Computational Linguistics, 2003, p. 142–147. [Online]. Available: https://doi.org/10.3115/1119176.1119195

[24] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074. [Online]. Available: https://www.aclweb.org/anthology/P16-1101

[25] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "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*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://www.aclweb.org/anthology/N16-1030

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[27] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[28] Y. Yadava, A. Hardie, R. Lohani, B. Regmi, S. Gurung, A. Gurung, T. Mcenery, J. Allwood, and Pat, "Construction and annotation of a corpus of contemporary nepali," *Corpora*, vol. 3, 11 2008.

[29] P. J. Ortiz Suárez, B. Sagot, and L. Romary, "Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures," in *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, Jul. 2019. [Online]. Available: https://hal.inria.fr/hal-02148693

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.