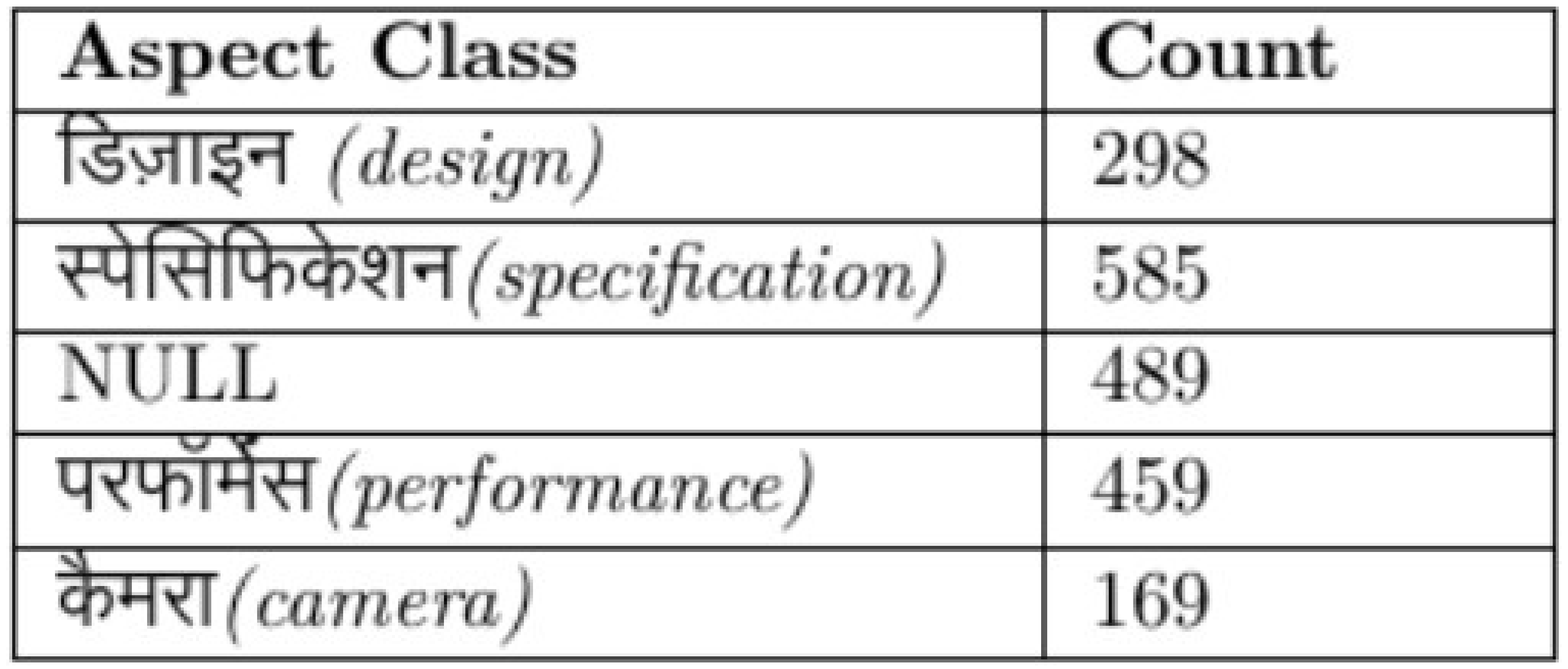
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2020

Dataset for Aspect Detection on Mobile Reviews in Hindi

Abstract

In recent years Opinion Mining has become one of the very interesting fields of Language Processing. To extract the gist of a sentence in a shorter and efficient manner is what opinion mining provides. In this paper we focus on detecting aspects for a particular domain in a relatively resource poor language Hindi. Here we present a corpus of mobile reviews in Hindi which are labelled with carefully curated aspects. We also propose baseline models to detect aspects in Hindi text after conducting various experiments.



Aspect Class	Count
डिज़ाइन *(design)*	298
स्पेसिफिकेशन *(specification)*	585
NULL	489
परफॉर्मेंस *(performance)*	459
कैमरा *(camera)*	169

Data Creation

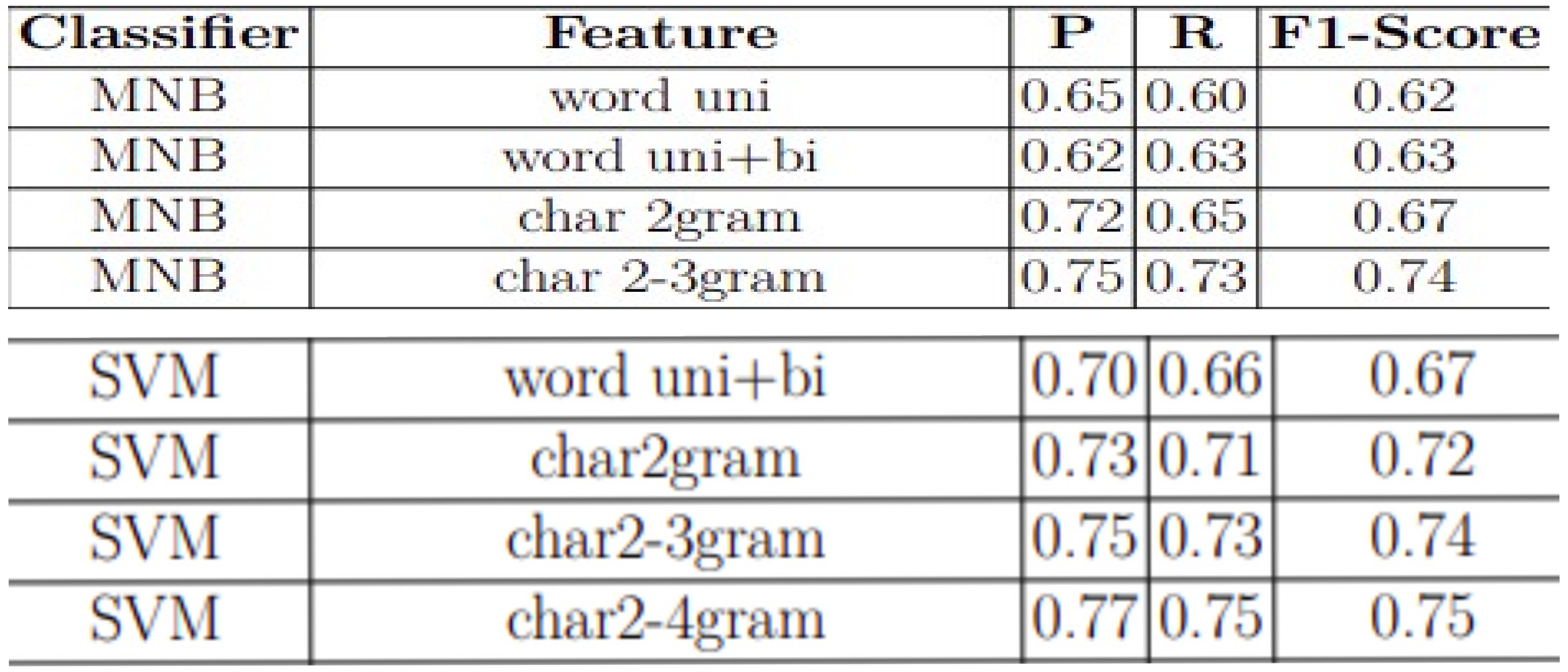
•As mentioned, earlier our work is on a specific domain. To build our corpus we scrapped data from various online forums with reviews on mobile phones. We retrieved 294 mobile reviews(37410 sentences) in a HTML format after extensive removal of noisy reviews. We had 294 HTML files which had raw data between different HTML tags. After initial annotation of assigning headings as the aspects, we had 18 classes of aspects in total. After eliminating all redundancies, we finally had 5 classes or aspects for our mobile reviews.

•The main task was to predict aspects in ev ery sentence in a review. We used different classifiers for the prediction task. We mostly experimented with machine learning models with 5-fold cross-validation as we had limited amount of data at our disposal. Our featureset consisted of Word N grams and Character N grams.

Results

•We observed that both the classifiers equally perform well on the data. We also observed that character n-grams models are superior than word n-gram models. Combination of word and char n-gram TF-IDF vectors do not improve the performance significantly.

•From the values of confusion matrix,we observed that class has स्पेसिफिकेशन (specification) overshadowed classes NULL and कै मरा(camera). It shows that our model is not able to predict between the umbrella class and the child class accurately.



Classifier	Feature	P	R	F1-Score
MNB	word uni	0.65	0.60	0.62
MNB	word uni+bi	0.62	0.63	0.63
MNB	char 2gram	0.72	0.65	0.67
MNB	char 2-3gram	0.75	0.73	0.74
SVM	word uni+bi	0.70	0.66	0.67
SVM	char2gram	0.73	0.71	0.72
SVM	char2-3gram	0.75	0.73	0.74
SVM	char2-4gram	0.77	0.75	0.75

Conclusion and Future work

•We annotated aspects for mobile reviews written in Hindi as a part of this work. We also presented baseline models for automatic aspect identification in mobile reviews.
•The baseline models will help us to annotate more reviews semiautomatically and can then be integrated to improve our systems. We will explore more into neural network architecture and word embeddings.
•The next task in this area would be to annotate polarity of the aspects. We can also explore identifying the most informative reviews

Authors: Ayush Joshi, Pruthwik Mishra, Dipti Misra Sharma Research Center: MT&NLP Lab, LTRC