Dataset for Aspect Detection on mobile reviews in Hindi

Ayush Joshi Pruthwik Mishra Dipti Misra Sharma
Language Technologies Research Center
Kohli Center On Intelligent Systems, IIIT Hyderabad
{ayush.joshi, pruthwik.mishra}@research.iiit.ac.in,,dipti@iiit.ac.in

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. While relevant research work has been done in aspect detection in resource rich languages like English, we are trying to do the same in a relatively resource poor Hindi language. Here we present a corpus of mobile reviews which are labelled with carefully curated aspects. The motivation behind Aspect detection is to get information on a finer level about the data. In this paper we identify all aspects related to the gadget which are present on the reviews given online on various websites. We also propose baseline models to detect aspects in Hindi text after conducting various experiments.

1 Introduction

Over the last decade people tend to search for products online rather than physically on stores. This has resulted in a surge of online forums where reviews are available on various products, electronic gadgets being one of the more popular ones. But reading so many long reviews is very time consuming and there is no uniformity on the parameters of reviews. To solve this, research work has been done in this area in the form of Aspect Detection which helps to point out the key specifications of the product in a structured format. But the work is limited to only worldwide languages as English and French. For a multi-lingual country like India, we are still far away in getting these information in the native language. We aimed at creating a dataset in Hindi which has the highest number of native speakers. The dataset is annotated with aspects for mobile reviews.

An aspect is a word in a sentence which has some polarity associated with it. The aspect should hold major meaning of the sentence. Following examples will state what aspect is:

S1 : शाओमी रेडमी 4ए को पहली बार हाथ में लेने पर यह आपको मेटल बॉडी का बना लगेगा ।
S1 : Xiaomi redmi 4A ko pehli baar hath m lene par yeh aapko metal body ka bana lagega.

Aspect1 : "मेटल बॉडी" (metal body) which falls under the "डिज़ाइन" (design) category. The aspect shows importance by indicating how the mobile is built.

S2: शाओमी रेडमी नोट में 2 गीगाहर्ट्ज़ ऑक्टा – कोर क्वालकॉम स्नैपड्रैगन 625 प्रोसेसर का इस्तेमाल हुआ है।
S2: Xiaomi Redmi note m 2 gigahertz octa-core qualcomm snapdragon 625 processor ka istemal hua hai. Aspect:"ऑक्टा – कोर क्वालकॉम स्नैपड्रैगन 625" (Octa core qualcomm snapdragon) which falls under the "स्पेसिफिकेशन" (specification) category . The aspect tells specifically tells the details of product.

S3: अफसोस यह कि आप उन्हें हटा नहीं सकते ।
S3: Afsos yeh ki aap unhe hata nahi sakte.
Aspect : "NULL" as there is no word which tells about any detail of the product. Hence, it is classified under no aspect category.

2 Related Work

Major work has been done in Aspect Detection when it comes to resource rich languages like English. The work of Aspect Detection has also been followed by Sentiments analysis which plays a major part in Opinion mining. In 2014 SemEval-Task 4, Maria Pontiki (2014) provided the first dataset which con-

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 130–134
Hyderabad, India, December 2019.

Aspect Class In Hindi	Aspect Class In Roman	Count
सॉफ्टवेयर	software	52
स्पेसिफिकेशन और फ़ीचर	Specification aur feature	360
हमारा फ़ैसला	hamara faisla	9
कैमरा और बैटरी लाइफ	camera aur battery life	5
स्पेसिफिकेशन और सॉफ्टवेयर	specification aur software	137
कैमरा	camera	76
परफॉर्मेंस	performance	826
लुक व बनावट	look vah banawat	26
बैटरी लाइफ	battery life	1
हमारा फैसला	hamara faisla	300
कैमरा परफॉर्मेंस	camera performance	16
डिज़ाइन	design	138
डिज़ाइन और लुक	design aur look	168
NULL	NULL	352
स्पेसिफिकेशन	specification	139
डिज़ाइन और बिल्ड	design aur build	390
डिज़ाइन और डिस्प्ले	design aur display	49
स्पेसिफिकेशन , सॉफ्टवेयर और परफॉर्मेंस	specification, software aur performance	40

Table 1: Class Set

sisted of English reviews annotated at sentence level with their aspects followed by their polarity. Some of the systems that emerged who targeted this task were Zhiqiang Toh (2014), Chernyshevich (2014); Joachim Wagner and Tounsi (2014); Giuseppe Castellucci (2014), Shweta Yadav (2015). However, almost all these systems are related to some specific languages, especially English. In 2016, SemEval released new datasets of similar domains(mobile, laptop, restaurant) [1] but in multiple languages. In 2016, the datasets were released in English, Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish.

But this area of field is largely unexplored in Indian languages due to the unavailability of high quality datasets and other tools and resources required. The datasets which were created by research groups mainly by Aditya Joshi (2010); Balamurali A R (2011, 2012) were very less in size and low in quality. Also Google transolator was used to create data in Indian languages (Akshat Bakliwal, 2012) but dataset created was not rich enough to perform aspect detection with high efficiency. Moreover the datasets available in Hindi were not domain specific which also added to poor results in past.

3 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 extracted the text from the HTML data with the help of BeautifulSoup library [2] in python. As our language was Hindi, online reviews were very less for which we tried both dynamic and manual crawling of data.

After crawling over 8 websites, we were able to get over 381 reviews. 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. There was no uniformity in the reviews, even after extraction and tokenization of these reviews,

[1] http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

[2] https://pypi.org/project/beautifulsoup4/

Unclean reviews	381
Unclean sentences	37410
Clean reviews	294
Clean sentences	2000
Total tokens	34359

Table 2: Corpus Details

Many reviews had proper headings like specifications, performance, price, design under which two-three paragraphs of text was present. But there were many reviews without any headings. To make it uniform and bring it to sentence level rather than paragraph level, we assigned the heading as labels to every sentence appearing under that heading in the review. This was our first annotation strategy. While assigning heading as aspects, there were certain sentences which had no heading above them. Such sentences were labelled as NULL. After this initial annotation, we had 18 classes of aspects in total. After doing analysis on our 18 classes, we observed a lot of overlapping between different classes. Some classes had the same name, but due to spelling variations they were assigned different labels. Table 3 gives a clear picture about the overlapping between different classes. We show the counts of highly frequent overlapping class pairs.

Class1 and Class2	**Overlap Count**
स्पेसिफिकेशन और फ़ीचर, स्पेसिफिकेशन और सॉफ्टवेयर *(specification aur feature), (specification aur software)*	441
डिज़ाइन और लुक, डिज़ाइन और बिल्ड *(design aur look), (design aur build)*	418
स्पेसिफिकेशन और फ़ीचर, स्पेसिफिकेशन *(specification aur feature, specification)*	410
डिज़ाइन, डिज़ाइन और बिल्ड *(design), (design aur build)*	387
स्पेसिफिकेशन स्पेसिफिकेशन और सॉफ्टवेयर *(specification), (specification aur software)*	336

Table 3: Overlapping Between Initial Classes

The following decisions to club different classes and provide them a single label were taken based on the percentage of overlapping.

- सॉफ्टवेयर*(software)*, स्पेसिफिकेशन और फ़ीचर *(specification aur feature)*, स्पेसिफिकेशन और सॉफ्टवेयर *(specification aur software)*, स्पेसिफिकेशन, स्पेसिफिकेशन , सॉफ्टवेयर और परफॉर्मेंस *(specification, specification, software aur perfomance)* clubbed under one single class called स्पेसिफिकेशन *(specification)*.
- कैमरा और बैटरी लाइफ *(camera aur battery life)*, कैमरा *(camera)*, , कैमरा परफॉर्मेंस *(camera performance)* were clubbed under a class कैमरा *(camera)*.
- लुक व बनावट *(look wh banawat)*, डिज़ाइन *(design)*, डिज़ाइन और लुक *(design aur look)*, डिज़ाइन और बिल्ड *(design aur build)*, डिज़ाइन और डिस्प्ले *(design aur display)* categorized under one class डिज़ाइन*(design)*.
- कैमरा *(camera)*,कैमरा परफॉर्मेंस *(camera performance)*, कैमरा और बैटरी लाइफ*(camera aur battery life)* were categorized under one class कैमरा *(camera)*.
- NULL and हमारा फैसला*(hamara faisla)* were merged as into a single class NULL.

After eliminating all these redundancies, we finally had 5 classes or aspects for our mobile reviews.

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

Table 4: Classwise Distribution

Two annotators were involved in this task. We obtained a Fleiss' [3] score of 0.87 for inter annotator agreement.

4 Experimental Setup

The main task was to predict aspects in every sentence in a review. We used different

[3] https://en.wikipedia.org/wiki/Fleiss'_kappa

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
MNB	char 2-4gram	0.74	0.74	0.74
MNB	char 2-5gram	0.73	0.75	0.74
MNB	word uni+char2-5gram	0.74	0.74	0.74
MNB	word uni+bi+char2-5gram	0.73	0.75	0.74
SVM	word uni	0.65	0.64	0.65
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
SVM	char2-5gram	0.77	0.75	0.76
SVM	word uni+char2-5gram	0.74	0.73	0.73
SVM	word uni+bi+char2-5gram	0.75	0.73	0.74

Table 5: Results Of Models After 5-fold Cross Validation

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.

4.1 Feature Engineering

Feature engineering is critical in designing accurate models. The features used in designing our supervised learning models are detailed here.

TF-IDF Vectors

- Word n-grams - This feature deals with the presence or absence of certain sequence of words. The value of n used varied from 1 to 2.
- Character n-grams - This is similar to word n-grams where a sequence of characters is extracted from the text. The value of n used varied from 2 to 5.

4.2 Machine Learning Approach

We created baseline with two classifiers

- Support Vector Machines (SVM)
- Multinomial Naive Bayes (MNB)

These two classifiers were implemented using the sklearn (Pedregosa et al., 2011) library. We used different feature set in both the classifiers.

5 Results

The results are shown in table 5. Classifiers and their corresponding features are detailed in this table. We used precision, recall and macro F1-score as the evaluation metric for checking the performance of our models. The words 'uni', 'bi' refer to the word unigrams and bigrams respectively. char 'a-b' gram denotes the combination of character n-grams where n lies in $\{a, a+1, a+2, .., b\}$

6 Observation

From table 5, 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 significantly improve the performance.

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

7 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 semi-automatically 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 (Mishra et al., 2017).

References

Pushpak Bhattacharyya Aditya Joshi, Balamurali A R. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. In *Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India.*

Vasudeva Varma Akshat Bakliwal, Piyush Arora. 2012. Hindi subjective lexicon : A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC).*

Aditya Joshi Pushpak Bhattacharyya Balamurali A R. 2011. Harnessing wordnet senses for supervised sentiment classification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31.*, pages 1081–1091.

Aditya Joshi Pushpak Bhattacharyya Balamurali A R. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters, COLING 2012, Mumbai, December 2012.*, pages 73–82.

Maryna Chernyshevich. 2014. Ihs rd belarus: Cross-domain extraction of product features using conditional random fields. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) .*, pages 309–313.

Danilo Croce Roberto Basili Giuseppe Castellucci, Simone Filice. 2014. Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland .*, pages 761–767.

Santiago Cortes Utsab Barman Dasha Bogdanova Jennifer Foster Joachim Wagner, Piyush Arora and Lamia Tounsi. 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland .*, pages 223–229.

John Pavlopoulos Haris Papageorgiou Ion Androutsopoulos Suresh Manandhar Maria Pontiki, Dimitrios Galanis. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014).*, pages 27–35.

Pruthwik Mishra, Prathyusha Danda, Silpa Kanneganti, and Soujanya Lanka. 2017. Iiit-h at ijcnlp-2017 task 3: A bidirectional-lstm approach for review opinion diversification. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 53–58.

Fabian Pedregosa, Gäel Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sriparna Saha Shweta Yadav, Asif Ekbal. 2015. Feature selection for entity extraction from multiple biomedical corpora: A pso-based approach. In *In Natural Language Processing and Information Systems, Springer.*, pages 220–233.

Wenting Wang Zhiqiang Toh. 2014. Dlirec: Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, August 23-24*, pages 235–240.