Resource Creation and Evaluation of Aspect Based Sentiment Analysis in Urdu

Sadaf Rani
Department of Computer Science,
Comsats University Islamabad,
Lahore Campus
1.5 km Defence Road, Off Raiwind Road,
Lahore, Punjab, Pakistan
sdf.rani0@gmail.com

Muhammad Waqas Anwar
Department of Computer Science,
Comsats University Islamabad,
Lahore Campus
1.5 km Defence Road, Off Raiwind Road,
Lahore, Punjab, Pakistan
waqasanwar@cuilahore.edu.pk

Abstract

Along with the rise of people generated content on social sites, sentiment analysis has gained more importance. Aspect Based Sentiment Analysis (ABSA) is a task of identifying the sentiment at aspect level. It has more importance than sentiment analysis from commercial point of view. To the best of our knowledge, there is very few work on ABSA in Urdu language. Recent work on ABSA has limitations. Only predefined aspects are identified in a specific domain. So our focus is on the creation and evaluation of dataset for ABSA in Urdu language which will support multiple aspects. This dataset will provide a baseline evaluation for ABSA systems.

1 Introduction

We are living in a world where web interaction is increasing. People share their emotions and express their feelings on different platforms through internet. A lot of work has been done to obtain valuable information from these reviews.

Sentiment analysis is a task of identifying people's emotions, feelings and opinion about a particular object from their reviews (Khan et al., 2018a). It is divided into three levels. Document level, Sentence level and Aspect level. "Document level" classifies the opinion of a complete document or paragraph into positive, negative, conflict or neutral. "Sentence level" classifies the sentiment of a sentence into positive, negative, conflict or neutral. "Entity/aspect level" It obtains the sentiment of a sentence at aspect (features) level (Patil and Yalagi, 2016) . Sentiment analysis has obtained importance in different areas like business intelligence and social media monitoring etc. Another area of interest is monitoring political data (Gold et al., 2018).

There has been very little research in the Urdu language since it is a low resourced language (Syed et al., 2010) Urdu has different morphological structures and linguistic features, so sentiment analysis intended for the English language can't be utilized for this language (Khan et al., 2018a) , (Syed et al., 2010). Many approaches and difficulties are discussed for the English language but very few for the Urdu language (El-Masri et al., 2017). Urdu is a mixture of many languages, so one of the challenging tasks is to handle different languages with their orientation (Khan et al., 2018b).

Aspect based sentiment analysis is a process of extracting the opinions expressed about a specific entity.

The sentiment analysis task becomes more challenging when there are multiple sentiments in a review on different aspects (Al-Smadi et al., 2015). Recently, the sentiment analysis research has moved towards having all the more fine-grained approaches thinking about the Urdu language aspects. But still, Urdu language is struggling at the aspect level. Recent work on the Urdu language at aspect level is not much more fine-grained as they deal only 4 predefined aspects. (ul Haq et al., 2020).

We aim to improve the current work by dealing the multiple aspects. An aspect can be a single word(e.g., Runs) or multi word (e.g., T 20 Series). A sentence may have one or more aspects that belongs to the same category. For this purpose clustering is used to group the related aspects into a single cluster. In this way we have a more fine-grained work at aspect level. We present a corpus of "Cricket" and "Football" domain in the Urdu language for ABSA. Which will provide a baseline evaluation and can further used for measuring the results of ABSA systems.

The rest of the paper is sorted out as follows. Section 2 describes related work. Section 3 contains a problem statement. Section 4 has

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 79–84
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

a proposed methodology. Section 5 contains a description of the data collection and annotation procedure. Section 6 provides information on guidelines for annotation. Section 7 has a conclusion.

2 Related work

The task of ABSA was first presented by SemEval in 2014 for English language (Pontiki et al., 2014). They provided data set of restaurant and laptop reviews for training and testing. (Al-Smadi et al., 2015) a benchmark dataset of Arabic language for aspect based sentiment analysis named as "HAAD" is prepared by them. They manually annotated the data at sentence level. Annotators used an online tool (BRAT) for annotation. This dataset provides the baseline evaluation to the four tasks i.e.; Aspect Extraction, Aspect Polarity, Aspect Category identification, Aspect Category Polarity.

(Rehman and Bajwa, 2016) a lexicon based approach has been adapted for sentiment analysis in Urdu language. They collected 124 comments from news sites. In preprocessing, tokens were generated. Polarity of each word is calculated by comparing with the sentiment lexicon. Accuracy of their approach was round about 66%.

(Arif et al., 2016) performed analysis on roman Urdu and provided the results by applying different classifiers on the dataset. Machine learning algorithms are applied on selected features for binary classification. Multiple classifiers are used on sparse matrix to evaluate the performance. Tf-idf is the term weighting model which gives the best overall accuracy with machine learning classifiers. SVM performs better then all of the classifiers by giving 96% accuracy.

Two datasets of low resourced languages Catalan and Basque for aspect level sentiment analysis are introduced by (Barnes et al., 2018). They performed tokenization, POS tagging and Lemmatization using Lxa-pipes. They trained a linear svc classifier for the classification of the polarity of opinion expressions. They trained a CRF on the standard features for the extraction of opinion holders, targets and expressions. For evaluation, 10 fold cross validation is used on 80% of data and F1 is used for extraction and classification.

(Apidianaki et al., 2016) this paper describes the data collection procedure and annotation guidelines. To increase the applicability and comparability of system, SemEval-2015 Task for English guidelines are used for annotation. They used different systems to identify three types of information (aspect category, opinion target expression and sentiment polarity). The categories returned by a framework are compared with gold annotations and Precision, Recall and F1 is measured.

(Nawaz et al., 2019) a segregational approach has been used for identifying the aspects. They introduced a technique which is consisted on two phases. 1st phase consist on extraction and grouping of target related words for a given objective by utilizing Normalized Google distance (NGD). Aspects are identified using POS tagger which is modified. In 2nd phase, they reduced the redundant and irrelevant aspects using Concept Net. They have applied this strategy on each word in an opinionated sentence to identify either it is a aspect word or non-aspect word.

(Brychcín et al., 2014) participated in the task of SemEval task 4 2014. They have used machine learning approach for constrained system and for unconstrained system they expand the constrained feature set by LDA, semantic spaces and semantic dictionaries. They compared their results with best and averages as well as with the baseline of SemEval and found that performance of their system is quite well.

(Zhao et al., 2014) worked on introducing a new annotation scheme and developed corpus for Chinese sentiment analysis. Elements of their annotation scheme are target entity, aspect, implicit aspect, polarity expression, modifier, Negation, polarity, transition word and compare. For sentiment analysis task, 3 main most relevant elements are ¡object, description, polarity¿. In their annotation scheme, 1st three elements are for object then 4-6 are for description and last three elements can be used to compute polarity. They performed an experiment on target-aspect pair extraction task. For cross validation, they used 10 fold on the new corpus. They applied Ml algorithms and reached the accuracy at 81.80%.

(Clematide et al., 2012) worked on the creation of Multilayered Reference corpus for German sentiment analysis. They used layered approach for annotation. At layer 1, polarity and subjectivity/objectivity is analyzed. At layer 2, they analyzed word and phrase-level annotation. At layer 3, they focused on the annotation of expression-level. Each layer had been analyzed by multiple annotators. To ensure the quality of data, they also calculated the inter-annotator agreement.

(Kumar et al., 2018) worked on the development of Hindi-English corpus for aggression annotation. They defined an annotation scheme in which they divided aggression tag sets into three levels (top-level). Overtly aggressive, covertly aggressive and non aggressive. Each of the top level has further two attributes – discursive role and discursive effect. Discursive effects have 10 kind(s) and are based on the type of aggression. They used a hierarchical approach of 3 top-level tags and 10 level 2 tags for annotation.

(Kılınç et al., 2017) worked on the creation of dataset named as TTC 3600. Stemming is performed on the primary data. For feature selection, two methodologies are used. Correlation-based feature selection and Attribute ranking based feature selection. Thinking about the high dimensionality and over fitting characteristics five classifiers were selected for text categorization (NB, SVM, K-NN, j48 and RF). Three different versions of TTC 3600 (F5-DS, F7-DS and Zemb-DS) were created by removing stop words and using stemmer. RF has the highest accuracy on all the datasets as compared to other classifiers. RF gives the best accuracy 91.03% on Zemb-DS dataset after applying ARFS.

(Tocoglu and Alpkocak, 2018) a dataset for emotion analysis in Turkish language has been prepared by them named as TREMO. They performed a validation process to validate the raw dataset and got two datasets. In preprocessing, they performed Fixed prefix (F5) stemming and Zemberek stemming. Which resulted four versions of dataset, F5, F5-V, Z and Z-V. For feature selection, they used mutual information. Then they applied four different classifiers (CNB, J48, RF and SVM) on all types of datasets and evaluate the performance by Precision, Recall, accuracy and F-measure. Result of SVM was better than the others.

3 Statement of the problem

Rather than classifying the complete sentiment of a sentence into positive, negative or neutral, ABSA allows us to associate specific sentiment with different aspects of a product. Since Urdu language is a low-resourced language, there is very little work on ABSA, especially in Urdu language (ul Haq et al., 2020). Unavailability of Urdu dataset for aspect based sentiment analysis leads to constructing a benchmark dataset of a specific domain. We aim to create a benchmark corpus to facilitate the ABSA systems. Annotation will be performed according to the guidelines of SemEval and performance is measured by using different Machine Learning approaches.

4 Proposed Methodology

As our proposed methodology can be seen from figure 1.

We crawled Urdu tweets of Cricket and Football domain from Twitter using Twitter API's. Then we performed preprocessing on raw data. In preprocessing, special characters, hashtags, links and punctuation marks are removed using Sci Kit Learn and regular expressions. After preprocessing, we prepared annotation guidelines for annotating the data. These guidelines are according to the standards of SemEval. After preparing guidelines for annotation, three annotators will manually annotate the dataset. In the annotation process, four types of information have to identify, i.e., aspect, aspect category, aspect polarity and aspect category polarity. Annotators will identify aspects and assign polarity. To ensure the quality of the dataset, we will compute the inter-annotator agreement. For feature extraction, we will apply Tf-idf vectorizer and n-gram models. Next, we will apply Baseline models of Machine Learning like naive Bayes, Random Forest, KNN. Then we will evaluate the performance by Precision, Recall and F1 measure.

5 Data Collection and Annotation

This dataset is consists of two different domains "Cricket" and "football". We collected 7000 tweets of cricket domain and 3000 tweets of football domain from Twitter using Twitter API's. Three annotators will manually annotate the data. They have to identify four types of information, which is discussed in section 4.

5.1 Data description

This section describes the data statistics that have been annotated in the developed corpus. Table 1 shows the aspect category statistics.

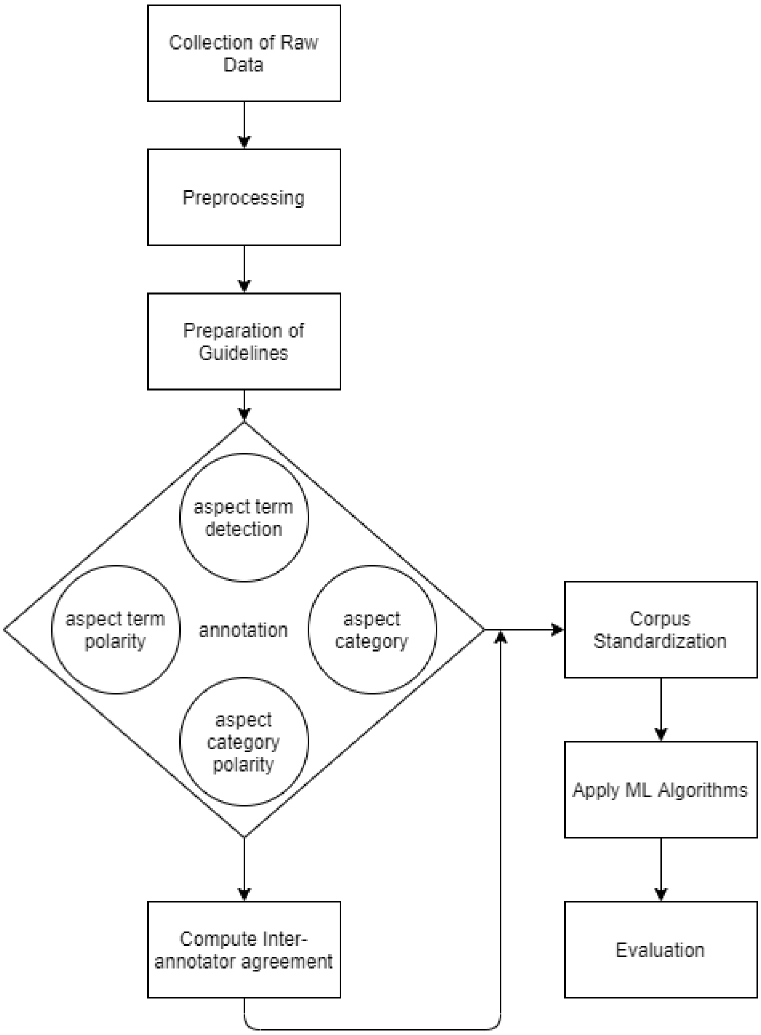

Figure 1: Research Methodology

6 Annotation guidelines

Initially, we have prepared some guidelines to annotate the data. The aim of these guidelines is to identify the aspects and sentiment polarity in sentences. We have to identify four types of information for annotation.

- Aspect / Entity
- Aspect Polarity
- Aspect Category
- Aspect Category Polarity

6.1 Guidelines for Aspect / Entity Extraction

Usually Noun / Nominal phrases indicate aspect term / aspect terms.

- Words / phrases that are expressing opinion (subjectivity indicator) are not considered aspect term.
- If an aspect term occurs more than once, then annotate all of them.
- If an identified aspect is misspelled, it should be annotated.
- Annotate aspect terms even if it is in quotation marks or brackets.
- Implicit aspects are not annotated.

6.2 Guidelines for Aspect polarity

Assign polarity to each aspect from these (positive, negative, neutral and conflict).

- When a sentence contains positive sentiment then assign positive polarity.
- When a sentence contains negative sentiment then assign negative polarity.
- Assign neutral polarity to aspect when an aspect is neither positive nor negative.

Category	Comments	Pos. Comments	Neg. Comments	Neutral Comments
performance	2117	1299	668	150
general	1678	653	496	529
inquiry Commission	89	14	58	17
management	1099	231	430	438
other	1689	313	1002	374
Total	6672	2510	2654	1508

Table 1: Statistics of aspect categories

- When a sentence has more than 1 aspect and each aspect have a contrast among polarities then assign conflict polarity to the aspects.

6.3 Guidelines for Aspect Category

Identify the aspect category from these predefined categories. Assign ”other” category to those aspects which have implicit aspect.

- management
- performance
- inquiry commission
- general
- other

6.4 Guidelines for Aspect Category Polarity

It is same as Aspect term polarity. Assign the polarity from (positive, negative, neutral and conflict).

7 Conclusion

In this paper, we have presented a benchmark corpus for ABSA in the Urdu language. This dataset has been prepared to cover different areas of research. It will facilitate researchers because we are covering four tasks of sentiment analysis. Which can be further used in the future by adding more domains. We believe that it will be very beneficial for the researcher community because we work in a different language and a specific domain. In future, we aim to improve our work by applying different techniques for feature extraction. We also plan to explore more machine learning algorithms and neural networks for these tasks.

References

M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider. 2015. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 726–730.

Marianna Apidianaki, Xavier Tannier, and Cécile Richart. 2016. Datasets for aspect-based sentiment analysis in french.

Huniya Arif, Kinza Munir, Abdul Subbooh Danyal, Ahmad Salman, and Muhammad Moazam Fraz. 2016. Sentiment analysis of roman urdu/hindi using supervised methods. *Proceedings of ICICC*, 8:48–53.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2018. Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. *arXiv preprint arXiv:1803.08614*.

Tomáš Brychcín, Michal Konkol, and Josef Steinberger. 2014. Uwb: Machine learning approach to aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 817–822.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. Mlsa—a multi-layered reference corpus for german sentiment analysis.

Mazen El-Masri, Nabeela Altrabsheh, and Hanady Mansour. 2017. Successes and challenges of arabic sentiment analysis research: a literature review. *Social Network Analysis and Mining*, 7(1):54.

Darina Gold, Marie Bexte, and Torsten Zesch. 2018. Corpus of aspect-based sentiment in political debates.

Ehsan ul Haq, Sahar Rauf, Sarmad Hussain, and Kashif Javed. 2020. Corpus of aspect-based sentiment for urdu political data. *LANGUAGE & TECHNOLOGY*, page 37.

Khairullah Khan, W Khan, A Rehman, A Khan, and Asfandyar Khan. 2018a. Urdu sentiment analysis. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 9(9).

SS Khan, M Khan, Q Ran, and R Naseem. 2018b. Challenges in opinion mining, a comprehensive review. *Sci. Technol. J.(Ciencia e Tecnica Vitivinicola)*, 33(11):123–135.

Deniz Kılınç, Akın Özçift, Fatma Bozyigit, Pelin Yıldırım, Fatih Yücalar, and Emin Borandag. 2017. Ttc-3600: A new benchmark dataset for turkish text categorization. *Journal of Information Science*, 43(2):174–185.

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Asif Nawaz, Sohail Asghar, and Syed Husnain Abbas Naqvi. 2019. A segregational approach for determining aspect sentiments in social media analysis. *The Journal of Supercomputing*, 75(5):2584–2602.

Priyanka Patil and Pratibha Yalagi. 2016. Sentiment analysis levels and techniques: A survey. *space*, 1:6.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Zia Ul Rehman and Imran Sarwar Bajwa. 2016. Lexicon-based sentiment analysis for urdu language. In *2016 sixth international conference on innovative computing technology (INTECH)*, pages 497–501. IEEE.

Afraz Z Syed, Muhammad Aslam, and Ana Maria Martinez-Enriquez. 2010. Lexicon based sentiment analysis of urdu text using sentiunits. In *Mexican International Conference on Artificial Intelligence*, pages 32–43. Springer.

Mansur Alp Tocoglu and Adil Alpkocak. 2018. Tremo: A dataset for emotion analysis in turkish. *Journal of Information Science*, 44(6):848–860.

Yanyan Zhao, Bing Qin, and Ting Liu. 2014. Creating a fine-grained corpus for chinese sentiment analysis. *IEEE Intelligent Systems*, 30(1):36–43.