Corpus of Aspect-based Sentiment for Urdu Political Data

Ehsan ul Haq[1],Sahar Rauf[2],Sarmad Hussain[3],Kashif Javed[3]
Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore
[1]{ehsan.ulhaq},[2,3]{firstname.lastname}@kics.edu.pk

Abstract

We present a corpus of Urdu political data annotated at aspect and sentiment level. The corpus contains 8760 tweets regarding four different aspects (Members, Projects, Party and Actions) of three political parties (PTI, PMLN and PPP) of Pakistan. We also present the results of a baseline system developed using the corpus for analyzing its reliability. It can be seen that the classifiers have achieved reasonable scores for aspects categorization and sentiment classification tasks.

1. Introduction

Sentiment analysis is defined as a task of automatically identifying opinions expressed in a text [1]. The text to be analyzed can be a feedback regarding a product or service, a political review or a social media comment [2]. The sentiment analysis can be done at document level, sentence level and also at aspect level [3]. In case of document level sentiment analysis, the task is to assign an overall sentiment to the document. On the other hand, sentence level sentiment analysis assigns a polarity value to each individual sentence of the document. The aspect level sentiment analysis provides an in-depth analysis by assigning sentiments to different aspects of entities mentioned in a text.

There are different approaches that have been used for performing sentiment analysis. These approaches include lexicon based approach; machine learning based and hybrid techniques [4]. In lexicon based approach, a sentiment lexicon is used for assigning sentiment to text by aggregating the sentiment scores of words present in the text. In machine learning based approaches, a sentiment tagged corpus is used to train machine learning models using supervised learning approach. After training models, they are used for assigning sentiments to input text. Another commonly used approach is using a hybrid technique. In hybrid techniques, a combination of machine learning models and lexicon is used for performing sentiment analysis.

Nowadays sentiment analysis has become an important task in the area like business intelligence (BI) in which a company wants to know the sentiment of customers towards their products or services. They use this analysis in decision making and overcoming their weaknesses. Another important area in which sentiment analysis is used is called social media monitoring (SMM) in which social posts are analyzed for finding opinions towards different entities [5]. Sentiment analysis is also used in disease surveillance systems for monitoring social media content mentioning symptoms, prevention and fear regarding a disease in different areas [6]. Another important area in which sentiment analysis is used is analysis of political data [7].

In this work, we are presenting a corpus for aspect based sentiment analysis (ABSA) task for Urdu political data which can be used as a gold-standard for automatic aspect-based sentiment annotation.

The rest of paper is organized as follows. Section 2 contains related work. Section 3 explains the methodology used for corpus development and its statistics. Section 4 contains baseline results and Section 5 is based on conclusion and future work.

2. Related Work

A corpus for ABSA in political debates has been presented in [7]. The corpus consists of transcribed speeches from the two presidential debates of the 2016 US election. The authors have annotated the corpus and provided baseline results for aspect based sentiment analysis using Support Vector Machine (SVM) algorithm.

The authors in [8] have presented an Italian corpus for aspect based sentiment analysis of movie reviews. The corpus contains sentences that have been manually annotated according to various aspects of movies and also polarities expressed toward them.

Two French language datasets for the purpose of development and testing of aspect based sentiment systems have been presented in [9]. The first dataset consists of 457 restaurant reviews (2365 sentences).The second contains 162 museum reviews (655 sentences). Both datasets were developed as part of SemEval-2016 Task 5 "Aspect-Based Sentiment Analysis" where seven different languages were represented, and are publicly available for research purposes.

A Turkish sentiment corpus comprised of user reviews annotated using semi-automatically is constructed in [10]. The corpus contains Turkish hotel reviews dataset which has 1000 reviews and 5364 sentences. The corpus also contains root forms of words, their usage, POS tags and sentiments.

An Arabic Laptops Reviews (ALR) dataset [11] for ABSA has been prepared according to the annotation scheme of SemEval16-Task5. The annotation scheme addresses two problems: prediction of aspect category and sentiment polarity label prediction. An evaluation procedure that extracts n-grams' features and uses a Support Vector Machine (SVM) classifier has also been described. in order to allow researchers to gauge and compare the performance. The results of evaluation show that there is a need for improvements in the performance of the SVM classifier for the aspect category prediction problem. On the other hand, the SVM's accuracy is actually high for sentiment polarity label prediction.

The work in [12] provides a human annotated Arabic dataset (HAAD). HAAD comprises of books reviews in Arabic which have been annotated by humans with aspect terms and their polarities. The paper also reports a baseline results with common evaluation techniques for the purpose of future evaluation of ABSA systems.

3. Urdu ABSA Corpus for Political Domain

This section describes the methodology that has been used for developing the corpus.

3.1. Corpus Collection and Preprocessing

We have collected 8760 tweets containing user comments regarding the following three political parties of Pakistan: Pakistan Tehreek-e-Insaaf (PTI), Pakistan Muslim League Nawaz (PMLN) and Pakistan People Party (PPP). The tweets have been collected in Urdu language using tweeter API. For searching relevant tweets, we have used the keywords indentified for each aspect in search queries.

After collecting the data, the next step is preprocessing. In preprocessing step, we have done the following tasks:

- Removal of special characters like hash tags, emoticons and punctuation marks like; '?, !, ;'.
- Resolving segmentation issues like incomplete words, issues of extra spaces between letters of words and presence of Zero Width Non Joiner (ZWNJ).
- Removal of duplicate tweets
- Removal of URLs and hyper links from the tweets.

3.2. Aspect-based tagging

The corpus has been designed for conducting aspect based sentiment analysis of the reviews of people regarding the above mentioned political parties of Pakistan. The reviews are analyzed for finding sentiments regarding the following four aspects:

3.2.1 Member

This aspect refers to the positive, negative and neutral comments of users regarding members of a party. For example, consider the following sentence:
/عمران خان کی نیت صاف ہے/
/ɪmrɑ:n xɑ:n ki: ni:jjət̪ sɑ:f hæ:/

Imran Khan's intentions are pure

Here, the quality of a PTI member has been mentioned in a positive way as the word 'pure' is very positive.

Consider another example:
/عمران خان کی سونامی تباہی ہے/
/ɪmrɑ:n xɑ:n ki: so:nɑ:mi: t̪əbɑ:hi: hæ:/

Imran Khan's tsunami is disastrous

Here, a negative sentiment has been expressed in the form of words 'tsunami' and 'disastrous'.

3.2.2. Projects

This aspect refers to the projects of a party. The followings could be the examples of party's projects; /نیا پاکستان/ /nəjɑ: pɑ:kɪst̪ɑ:n/ /New Pakistan/, / اورنج ٹرین/ɔ:rɪndʒ tre:n/ /Orange train/ etc.
Consider the following review as an example:
/اورنج ٹرین ایک ناکامیاب پروجیکٹ ہے/
/ɔ:rɪndʒ tre:n e:k nɑ:kɑ:mjɑ:b pro:dʒækt hæ:/
Orange train is an unsuccessful project

In this sentence, a negative sentiment is attached with the project of PMLN in the form of word 'unsuccessful'.

3.2.3. Actions

This aspect refers to policies of a party. These policies and actions could be foreign policies, economy policies, price control policies and health policies etc. The following keywords could be examples of action aspect; /مہنگائی/ /mæhŋgɑ:i:/ /Inflation/ and /قرضہ//qərzɑ:/ /Debt/ etc.
Consider the following example:
/ پی ٹی آئی کی حکومت کی وجہ سے مہنگائی بڑھ رہی
ہے/

/pi: ti: ɑ:i: ki: həku:mət̪ ki: vəʤɑ: se: mæhŋgɑ:i: bəɽ^{h} rəhi: hæ:/

Inflation is increasing due to PTI's government

In the above mentioned sentence, a negative sentiment is expressed towards the action of PTI.

3.2.4. Party

This aspect refers to feedback of people regarding party as a whole.

Consider the following example:

/پی ٹی آئی کی کارکردگی اچھی نہیں/
/pi: ti: ɑ:i: ki: kɑ:rkərd̪əgi: əʧ^{h}ʧ^{h}i: nəhi:/

The performance of PTI is not good

This review contains a negative feedback regarding performance of PTI as a whole party rather than individual members.

3.3. Sentiment Polarities

For the purpose of assigning sentiments to each aspect, we have used a five point scale from -2 to 2, where -2 means more negative, -1 means less negative, 0 means neutral, 1 means positive and +2 means more positive.

3.4. Corpus Tagging

A team of expert linguists with Mphil and PhD degrees in the area of Applied Linguistics and Urdu Literature respectively has tagged the corpus. The tested data achieved Inter Annotator Accuracy (IAA) of 75% at aspects and sentiment level tagging.

3.5. Corpus Statistics

This section explains the statistics of corpus that has been tagged as aspect and sentiment level. Table 2 below is presenting the statistics of each aspect in the developed corpus.

Table 2. Statistics of aspects in developed corpus

PARTY	MEMBER	PROJECT	PARTY	ACTION
PTI	1622	445	798	669
PMLN	2215	758	621	308
PPP	1497	306	919	535

The statistics of aspect wise sentiments is given in Table 3 below.

Table 3. Statistics of aspect wise sentiments in the corpus

		Sentiment		
Party	**Aspects**	**POS**	**NEG**	**NEU**
PTI	MEMBER	314	1090	218
	PROJECT	50	391	4
	PARTY	151	521	126
	ACTION	81	570	18
PMLN	MEMBER	758	1257	200
	PROJECT	406	205	147
	PARTY	135	375	111
	ACTION	157	140	11
PPP	MEMBER	295	914	288
	PROJECT	89	126	91
	PARTY	295	501	123
	ACTION	34	489	12

4. Automatic Aspect-Based Sentiment Annotation

The developed corpus has been used for measuring the performance of machine learning based algorithms and baseline results have been reported. This evaluation is useful for indicating the reliability of the developed corpus.

Hence, for the purpose of analyzing that whether the corpus can be used for developing an ABSA classifier, we have trained classifiers for aspects recognition and sentiment tagging using an SVM library, namely LIBSVM in Weka [13].

We have used One-Vs-All approach and trained classifiers for each aspect and sentiment separately. So, we have trained 12 models for aspects and 12 models for sentiments classification. We have evaluated the results by performing 10-fold cross validation.

We have also experimented with different n-gram features with n ∈ {1,2,3}. For the purpose of vectorizing the data, we have used a binary scheme in which the vector contains 1 if word is present and 0 otherwise.

The results for sentiments classification system are presented in Table 3 below.

Table 4. F1-scores for sentiment models using CV

		Parties		
Features	**Aspects**	**PTI**	**PMLN**	**PPP**
Unigram	MEMBER	70.7	70.6	65.2
	PROJECT	85.7	62.8	60.7
	PARTY	77	65.3	77.5
	ACTION	80	75.1	89.9
Bigram	MEMBER	64	70.1	65.7
	PROJECT	84.1	62.8	60.9
	PARTY	90.1	64.2	71.9
	ACTION	80.3	73.6	90.2
Trigram	MEMBER	63.1	63.9	61.1
	PROJECT	82.6	47.6	59.9
	PARTY	61.1	53.3	60.4
	ACTION	78.9	59.8	89.9

The results of aspects classifiers are presented in Table 4 below.

Table 5. F1-scores for aspects models using CV

		Parties		
Features	**Aspects**	**PTI**	**PMLN**	**PPP**
Unigram	MEMBER	88.5	86.3	90
	PROJECT	96.2	94.6	98.3
	PARTY	90.7	92.9	91.6
	ACTION	91.2	94.1	95.6
Bigram	MEMBER	89.9	87.7	88.4
	PROJECT	96.8	95.6	98.9
	PARTY	92.9	92.7	93.7
	ACTION	90.6	94.6	94.1
Trigram	MEMBER	79.1	76.1	77.1
	PROJECT	94.3	91.3	92.3
	PARTY	90.7	90.6	89.9
	ACTION	88.9	94.7	85

5. Conclusion

A corpus for ABSA for Urdu political data has been presented in this paper. The developed corpus has been tagged at aspect and sentiment level with IAA score of 75%. A baseline system has also been developed using the corpus for analyzing its reliability in future use. It can be seen from the above mentioned results that the aspect classifiers have achieved a F1-score of more than 90% in most of the cases. Moreover, the sentiment classifiers have also achieved a F1-score of more than 70% in many cases

6. References

[1] Walaa, A. Hassan, and H. Korashy Medhat, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[2] Maria, D. Galanis, H. Papageorgiou, I. Androutsopoulos and S. Manandhar Pontiki, "Semeval-2016 task 5: Aspect based sentiment analysis," in Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), 2016, pp. 19-30.

[3] D. Mohey and El-Din Mohamed Hussein, "A survey on sentiment analysis challenges," Journal of King Saud University-Engineering Sciences, vol. 30, no. 4, pp. 330-338, 2018.

[4] "Automated sentiment analysis in tourism: Comparison of approaches," Journal of Travel Research, vol. 57, no. 8, pp. 1012-1025.

[5] Muhammad, M.Diab, and S.Kübler Abdul-Mageed, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," in Computer Speech & Language, 2014, pp. 20-37.

[6] V. Kumar, and S.Kumar. Jain, "Effective surveillance and predictive mapping of mosquito-borne diseases using social media.," Journal of Computational Science, pp. 406-415, 2018.

[7] M.Bexte and T.Zesch D.Gold, "Corpus of Aspect-based Sentiment in Political Debates," in Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018), 2018.

[8] Antonio, et al. Sorgente, "An italian corpus for aspect based sentiment analysis of movie reviews," in CLICIT2014, 2014.

[9] Marianna, X.Tannier, and C.Richart Apidianaki, "Datasets for aspect-based sentiment analysis in french," in Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016.

[10] S. İlhan, E. Ekinci, and H. Türkmen Omurca, "An annotated corpus for Turkish sentiment analysis at sentence level.," in In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP).IEEE., 2017, pp. 1-5.

[11] Al-Ayyoub and Mahmoud, "Aspect-Based Sentiment Analysis of Arabic Laptop.," , 2017.

[12] Mohammad, et al. Al-Smadi, "Human annotated arabic dataset of book reviews for aspect based sentiment analysis," in 2015 3rd International Conference on Future Internet of Things and Cloud. IEEE, 2015., 2015.

[13]https://www.cs.waikato.ac.nz/~ml/weka/index.html.