Telematics and Informatics 35 (2018) 2173–2183

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains

Neelam Mukhtar[a,*], Mohammad Abid Khan[a], Nadia Chiragh[b]

[a] Department of Computer Science, University of Peshawar, KPK, Pakistan
[b] University of Agriculture, Peshawar, Pakistan

ARTICLE INFO

Keywords:
Supervised Machine Learning approach
Lexicon-based approach
Urdu Sentiment Lexicon
Urdu Sentiment Analyzer

ABSTRACT

Web is facilitating people to express their views and opinions on different topics through reviews and blogs. Effective advantages can be reaped from these reviews and blogs by fusing the sentiment knowledge. In this research, Sentiment Analysis of Urdu blogs from multiple domains is done by using the two widely used approaches i.e. the Lexicon-based approach and the Supervised Machine Learning approach. Three well known classifiers i.e. Support Vector Machine, Decision Tree and K Nearest Neighbor are used in case of Supervised Machine Learning approach whereas a wide coverage Urdu Sentiment Lexicon and an efficient Urdu Sentiment Analyzer are used in Lexicon-based approach. In both the approaches the information are fused from two sources to successfully perform Sentiment Analysis. In case of Lexicon-based approach, the two sources are the wide coverage Urdu Sentiment Lexicon and the efficient Urdu Sentiment Analyzer. In case of Supervised Machine Learning approach, the two sources are the un-annotated data and annotated data along with important attributes. After performing Sentiment Analysis using both the approaches, the results are observed carefully and on the basis of experiments performed in this research, it is concluded that the Lexicon-based approach outperforms Supervised Machine Learning approach not only in terms of Accuracy, Precision, Recall and F-measure but also in terms of economy of time and efforts used.

1. Introduction

Sentiment Analysis (SA) is a hot area of research in these days that uses different techniques of text mining, Natural Language processing and computational linguistics for the identification and extraction of subjective information from different sources for example social networks (Farooq et al., 2017, V-Díaz et al., 2018). Extracting relevant opinions of interest from a huge amount of text is not an easy task (Ali et al., 2015). For performing Sentiment Analysis different methods and approaches are used by different researchers. For extracting feature's opinion from reviews and computing polarities, fuzzy domain ontology with SVM is proposed by Ali et al. (2016). Fuzzy ontology based Sentiment Analysis along with semantic web rule language (SWRL) rule-based decision-making is proposed by Ali et al. (2017) for monitoring transportation activities. Hybrid approach based on sentiment lexicon, semantic rules, negation handling and ambiguity management is used by Appel et al. (2017).

Among different approaches, the two commonly used approaches in this area of research i.e. the Lexicon-based approach and the Machine Learning approach. Lexicon-based approach is also called unsupervised approach. Machine learning approach can either be supervised (e.g. classification) or unsupervised (e.g. clustering).

* Corresponding author.
E-mail address: sameen_gul@yahoo.com (N. Mukhtar).

https://doi.org/10.1016/j.tele.2018.08.003
Received 23 May 2018; Received in revised form 13 July 2018; Accepted 1 August 2018
Available online 04 August 2018

A lexicon of opinion words is required along with the sentiment score in the Lexicon-based approach. In Supervised Machine Learning (SML) approach, a large amount of labeled data, to be annotated by annotators, is required for the training of classifiers.

A large number of people share their opinions/sentiments on the web using Urdu. So, SA of Urdu blogs is focused in this comparison of the two widely used approaches i.e. Lexicon-based and SML are used for the SA of Urdu blogs and are compared to find out that which one is more effective. For Lexicon-based approach, a series of steps are taken to create Urdu Sentiment Lexicon (SL). After the creation of the Urdu SL, an effective algorithm for performing SA of Urdu blogs is developed and implemented to get Urdu Sentiment Analyzer.

For testing the Urdu Sentiment Analyzer, a corpus of 6025 sentences from 151 blogs belonging to 14 different genres is collected. These genres along with the number of blogs considered per genre are:

Current affairs (32 blogs), Ethics (10 blogs), General Knowledge (13 blogs), Personalities (3 blogs), Religion (16 blogs), Sports (5 blogs), Politics (17 blogs), Humor and Satire (5 blogs), Technology (10 blogs), Economy (6 blogs), Psychology (10 blogs), History (8 blogs), Health (10 blogs) and Language and literature (6 blogs). Two human annotators are hired to classify the sentences in corpus as positive, negative or neutral. For tied sentences, when there is disagreement between the two annotators, a third annotator is hired. After successful completion of this step, 1876 sentences are annotated as positive, 2753 sentences as negative and 1388 sentences as neutral (Mukhtar and Khan, 2018). Sentences are processed by the analyzer one by one and are classified as positive, negative or neutral sentences. Urdu Sentiment Analyzer achieves an overall accuracy of 89.03%.

In case of SML approach, Three equal samples with 1800 sentences each are selected and are run on the three classifiers Support Vector Machine (Lib SVM), Decision Tree J48 and K Nearest Neighbor (KNN; IBK) with 10-fold cross validation using Weka[1], a collection of machine learning algorithms, basically used for mining data/text. IBK with 67.02% accuracy emerged as the best classifier among the three.

A detailed comparison is done between the two approaches in terms of Accuracy, Precision, Recall and F-measure. Although in SML approach, IBK proved to be the best classifier than Lib SVM and J48 but achieved only 67.02% Accuracy, 0.68, 0.67 and 0.67 Precision, Recall and F-measure respectively. On the other hand, the Lexicon-based approach achieved 89.03% Accuracy with 0.86 Precision, 0.90 Recall and 0.88 F-measure, after handling negations, context-dependent words and intensifiers. Thus, in this research, the Lexicon-based approach proved to be a better approach than SML approach in case of SA of Urdu blogs in multiple domains.

The rest of the paper is organized as follows. Section 2 provides an overview of the comparison that is done by other researchers with respect to these two approaches. In Section 3, the methodology adopted is discussed in detail. In Section 4, results from both the approaches are presented. Comparison between the two approaches is presented in Section 5. In Section 6, a conclusion is presented.

2. Comparison of the two approaches by the research community

Both approaches are compared by the research community. This comparison is presented one by one.

A large amount of training data, annotated by humans, is required for SML in order to get satisfactory results which is not only a difficult task but also very time consuming and costly (Zagibalov, 2010). SML approaches are generally more accurate, if tested in single domain for which they are trained but Lexicon-based approaches are less time consuming and show less memory complexity. A lot of time is consumed in SML approaches while annotating (labeling) the data. Memory complexity also increases in this approach as a huge amount of labeled data is to be saved in the memory. SML approaches are much slower than Lexicon-based approaches (Augustyniak et al., 2016). Lexicon-based approach is efficient enough to be applied for text analysis at document, sentence and entity level (Zhang et al., 2011). The efficiency of this method entirely depends on the presence of opinion words in the lexicon. A lexicon with a poor coverage (i.e. small sized lexicon with lesser number of words) may result in low recall no matter how efficient the algorithm for SA is developed. A particular classifier performs very well in one domain and may achieve an excellent accuracy if it is trained in that domain but the same classifier will perform poorly in another domain for which it remains untrained (Aue and Gamon, 2005). SML approach is not scalable for the analysis of twitter which covers different domains as people can express their opinions about anything (Zhang et al., 2011). SML approach generally overcomes the Lexicon-based approach in terms of accuracy (Nakov et al., 2013, Rosenthal et al., 2014), in the particular domain for which it is trained, the Lexicon-based approach on the other hand, avoid the laborious steps that are needed for labeling the data to train the classifiers (Musto et al., 2014).

An advantage of SML method is that it is capable to adapt and then create models for which the classifiers are being trained i.e. for specific context. This method suffers from a main drawback that is the un-availability of labeled data that results in their low applicability to a new data in another domain (Gonçalves et al., 2013). On the other hand, no labeled data is required in case of Lexicon-based approach but practically, it is not very easy to create a lexical-based dictionary that is unique and with such a very wide coverage, that it can be used in different contexts.

Through experimental results, it is shown that SML approach is having a high precision whereas Lexicon-based approach is equally competitive due to less hard work required in this approach and more needed in case of SML approach for classifier training with less sensitivity for the quality and quantity of the data set used for training (Hailong et al., 2014).

The same view as above is provided in (D'Andrea et al., 2015). According to these authors SML approach enjoys the ability for adapting and creating trained model for specific context and purpose. It suffers from a major limitation i.e. its applicability to new data on which it is not trained previously. Lexicon-based approach on the other hand has broader coverage (of domain-independent words) but suffers from the drawback i.e. of the lexicon with a finite number of words and fixed sentiment score assigned to the words

[1] http://www.cs.waikato.ac.nz/ml/weka/.

in the lexicon.

In case of SML approach, first the data is collected from different domains then this collected data is labeled manually. In case of Lexicon-based approach, all this process is not required but a decision is to be taken that is either to save the stemmed word or the actual word along with its derivatives, prefixes, affixes and suffixes. Saving the actual word increases the accuracy but slows down the learning curve and increase the size of the dictionary (lexicon). In case only the stemmed word is saved, the dictionary size is decreased while speeding up the learning process but the accuracy level is also decreased (Shoukry, 2013).

Shortcomings of both the Lexicon-based approach and SML approach are highlighted in (Hutto and Gilbert, 2014). According to these authors, major shortcomings of the Lexicon-based approach are:

1. They have problems with coverage as important lexical features are ignored.
2. To acquire new set of lexical features along with their valence scores is labor intensive and time consuming process.

Similarly, the major shortcomings that are faced by SML approach are:

1. A huge amount of training data is required, that is sometimes difficult to obtain.
2. The training set is to represent a number of features which is again sometimes difficult in the case of social media with short, sparse text.
3. Computationally, they are more expensive in terms of memory requirement, processing time and the time required for training/classification.

After a number of experiments Choi and Lee (2017), concluded that different data properties i.e. size of the data set, length of the target document and subjectivity of the training/testing data plays important role in improving the performance of the two approaches. According to the authors, for short documents, instead of implementing complex machine learning algorithm and spending a lot of time in collecting a large training data set for Supervised Machine Learning, lexicon-based approach can show good performance than machine learning approach. For the documents with high subjective words, machine learning approach can show good performance.

Lexicon-based approach is performing better than SML in this research because a wide coverage lexicon and efficient Urdu Sentiment Analyzer is developed that can efficiently handle data from different domains. SML approach can handle data from different domains efficiently only if adequate data is collected in each domain which is quite time consuming process.

3. Methodology

For testing in case of Lexicon-based approach and both training and testing in case of SML approach, 6025 sentences from 151 Urdu blogs belonging to 14 different genres are collected from different sources. This data is annotated by three human annotators. For verification, inter-annotator agreement is calculated by using Kappa statistic (Siegel and John Castellan, 1988). Kappa statistic is frequently used to test the inter-rater reliability, where inter-rater reliability is the measurement of the range up to which, same score is assigned to the same variable by different raters (McHugh, 2012). Kappa = 0.78, indicates substantial agreement between the human annotators (Viera and Garrett, 2005).

Different methodologies are adopted for Lexicon-based approach and SML approach. Both types of methodologies are discussed one by one in the subsections below:

3.1. Methodology adopted for Lexicon-based approach

In case of Lexicon-based approach, first Urdu Sentiment Lexicon is created and then Urdu Sentiment Analyzer that uses this Urdu Sentiment Lexicon is created for performing Urdu Sentiment Analysis.

3.1.1. Urdu Sentiment Lexicon creation

For the creation of Urdu Sentiment Lexicon (SL), first Positive and negative words are collected from two different sources.[2,3] The collected words are further expanded using the Urdu lughat[4] Negative words are kept in one file and positive words are kept in another file. Parts of Speech (POS) are assigned to each word by using an Urdu tagger[5]. The tagger uses the tag set developed by Center for Language Engineering[6]. For maximum reliability, the POS assigned by the tagger to each word is again validated manually by reusing the Urdu lughat.

The developed Urdu SL is thus developed with 11,739 negative words, 9578 positive words, along with separate files for intensifiers, context-dependent words and negations.

[2] http://chaoticity.com/urdusentimentlexicon/.
[3] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment.
[4] http://urdulughat.info/.
[5] http://www.cle.org.pk/.
[6] http:// www.cle.org.pk/software/langproc/POStagset.htm.

3.1.2. Urdu Sentiment Analyzer for Lexicon-based approach

After the development of Urdu SL, a rule-based Urdu Sentiment Analyzer is developed. This rule-based Analyzer is developed by utilizing the developed corpus of 6025 sentences plus taking commonly used examples from the daily life as well. The main steps of the algorithm that work behind the Urdu Sentiment Analyzer are given below:

1. If a word occurs in positive word file, it gets +1 polarity value. If it is found in the negative word file, −1 is assigned to it.
2. If a negative word comes immediately before or immediately after a positive word, then −1 is assigned collectively to both words.
3. A positive word gets negative polarity collectively and a negative word gets positive polarity collectively, if there is negation before or after this word.
4. The positive or negative word gets the same polarity (i.e. +1 for positive word collectively and −1 for negative word collectively) if there is صرف (sirf i.e. only) between the negation i.e. negation followed by صرف e.g. نہ صرف (nah sirf i.e. not only) and the positive or negative word.
5. A positive word gets negative polarity and a negative word gets positive polarity, if there is a negation word after it, even in the presence of an intensifier.
6. If there is an amplifier (intensifier that increases the degree/effect) before/after a positive word its polarity separately is +1. This distance between words may vary from one word to many. Similarly, if there is an amplifier before/after a negative word its polarity separately is −1.
7. The intensity of "زیادہ بہت" (bohat ziyadah i.e. very much) is much more than "زیادہ " (ziyadah i.e. more) (Syed et al., 2010). Thus, if there are two amplifiers before a positive word, then the word and each amplifier is separately assigned +1 polarity. Similarly, if there are two amplifiers before a negative word, then each of the three words get −1 polarity.
8. If a diminisher/down-toner (intensifier that decreases the effect/degree) is preceded or followed by a positive word, then −1 is assigned collectively to them. However, if a diminisher/down-toner is preceded or followed by a negative word, then +1 is assigned collectively to them.
9. If a context-dependent word is preceded or followed by a positive word, then +1 is assigned to the context-dependent word. However, if a context-dependent word is preceded or followed by a negative word, then −1 is assigned to the context-dependent word.
10. If the same word is repeated twice it is assigned polarity only once i.e. either +1 or −1 depending upon that word, whether it is positive word or negative word.

These steps are implemented using Java JDK 6, net bean environment, Tee Chart for Java that resulted in an efficient Urdu Sentiment Analyzer.

This Urdu Sentiment Analyzer takes an Urdu blog with a number of sentences as input. Words in each sentence are first searched in different files. After this step, words are assigned polarities according to specific rules. When the end of the sentence is reached, the polarities assigned to the words in each sentence are added.

If the sum of polarities exceeds 0, the sentence is declared as a positive sentence. If the sum of polarities is less than 0, the sentence is identified as a negative sentence otherwise the sentence is considered as a neutral sentence i.e. either there is no opinion in the sentence or the sentence is with mixed opinion (i.e. both positive and negative). At the end of this process, a final conclusion is displayed which shows total number of positive sentences, negative sentences and neutral sentences. Fig. 1 shows the diagrammatic representation of these steps.

3.2. Methodology adopted for SML approach

In case of SML approach, different steps are taken which are discussed below:

3.2.1. Stop word removal

Urdu stop words play an important role in sentence completeness. Removal of stop words are based on the idea that removal of such words results in reducing the feature space of the classifiers that aids in producing accurate results (Silva and Ribeiro, 2003). The stop words in the Urdu corpus are not removed earlier at the time of annotation phase, as the annotators may face problems in predicting the correct classification of the sentence due to incomplete sentence. After the successful annotation phase, stop words are no more required as they are not playing any role in the correct classification of sentences rather increasing the computation time taken by the classifiers. So, they are removed manually from the data for getting maximum accuracy.

3.2.2. Classification of sentences

Classification of the sentences as positive, negative or neutral is done using Weka. As per requirement, first the data is converted to Attribute-Relation File Format (ARFF). The training data is then used by the three selected classifiers. In the next step, 10-fold cross-validation is used for the evaluation of the classifiers by taking each classifier one by one.

3.2.3. Attribute/Feature selection

In the beginning, the classifiers resulted in very poor accuracy i.e. the accuracy is below 50%. 154 attributes are then selected from the data. These attributes can also be called features of the sentences. Sentences are taken one-by-one and the attributes/

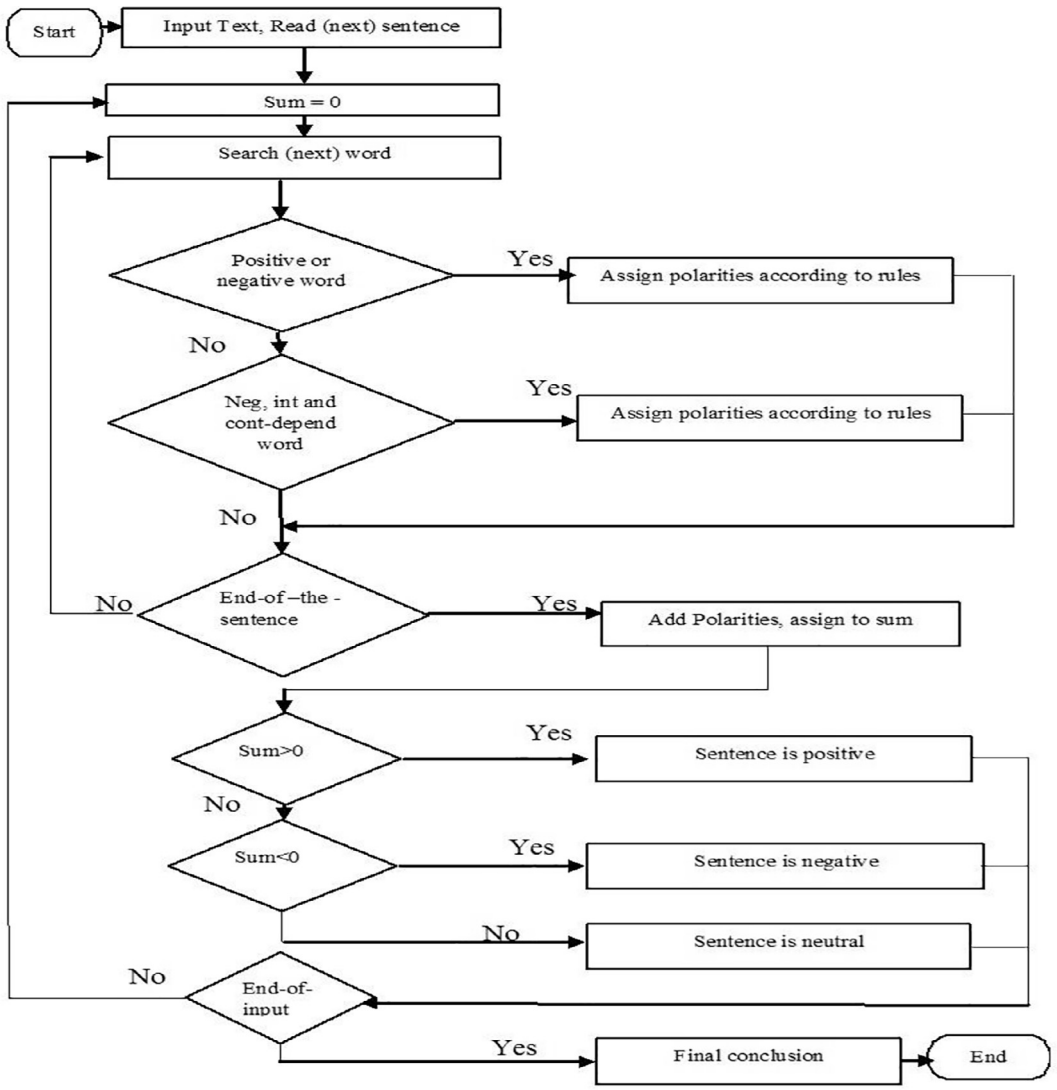

Fig. 1. Main steps of the Urdu Sentiment Analyzer.

features are located. In case of absence of a particular attribute in a sentence, a "?" is placed for the missing value. 10-fold cross-validation is again applied; slight improvement in the performance of Lib SVM is noticed with 50.28% accuracy. The results are still not encouraging as at least 60% accuracy is the target.

Table 1 shows the results with 154 attributes.

3.2.4. Important attribute/feature selection.

There are some attributes out of 154 attributes that are although identified in the previous step but they are having either "no values" or "less than 10 values". By "no value", it is meant that the attribute is not used in any one of the 6025 sentences and by "less

Table 1
Classification results with 154 attributes.

Classifier	Evaluation Metrics			
	Precision (Float)	Recall (Float)	F-Measure (Float)	Accuracy (%)
Lib SVM	0.47	0.50	0.43	50.28
KNN (IBK)	0.39	0.43	0.37	43.29
Decision Tree (J48)	0.19	0.45	0.27	44.47

than 10 values", it is meant that out of 6025, it is used once or more than one time but the total number of occurrences of that particular attribute is still less than 10. All such attributes in this research are termed as "less important attributes "and are discarded. Only 39 "important attributes", that are used 10 or more than 10 times in the collected corpus are retained.

The three classifiers are again tested in the new scenario where the "important features" are retained. Three samples with 1800 sentences in each sample are taken. Sampling is done without replacement i.e. different samples are selected for maximum reliability. In each sample, equal number of positive, negative and neutral sentences (i.e. 600 each) is taken with 10-fold cross-validation. Improvement is observed in the performance of all the three classifiers, mainly due to discarding "less important attributes" and retaining "important attributes".

4. Results

Encouraging results are achieved from both approaches i.e. Lexicon-based approach and SML approach (taken from Ph.D thesis). These results are discussed one-by-one in the sub-sections:

4.1. Results from the Lexicon-based approach

An efficient algorithm is developed and implemented to get an Urdu Sentiment Analyzer. This Analyzer takes Urdu sentences as input, analyze them and classify them as positive, negative or neutral.

The algorithm for Urdu Sentiment Analyzer is implemented in four phases. Only positive and negative words are considered in phase 1. Apart from adjectives, nouns and verbs are also taken. In case of positive and negative words, both simple and compound words are included. Negations, intensifiers and context-dependent words are not dealt yet. In addition to positive and negative words, negations are handled in phase 2. Handling of intensifier is added in phase 3. Context-dependent words are successfully dealt along with the previous three phases in the last phase i.e. the 4th phase. These four phases collectively resulted in the development of Urdu Sentiment Analyzer. Table 2 shows all the four phases and the phase wise accuracy that is achieved. Each succeeding phase brought more accuracy compared to previous phase(s).

Where accuracy is the measurement of how close the estimated classification is to the actual classification.

$$\text{Accuracy} = (\text{correctly classified sentences}/\text{total number of sentences}) \times 100$$

For evaluating the effectiveness and efficiency, accuracy alone is not sufficient performance metric. Therefore the other three standard metrics i.e. Precision, Recall and F-measure are also considered.

Table 3 shows the Precision, Recall and F-measure of the final phase when 89.03% accuracy is achieved.

It is observed from Table 3, that 0.86 Precision, 0.90 Recall and 0.88F-measure on the average are achieved, which are quite promising results.

4.2. Results from the Supervised Machine Learning approach

Three individual samples are drawn from 6025 sentences by taking equal number of positive, negative and neutral sentences. There are 1800 sentences in each sample, where 600 positive, 600 negative and 600 neutral sentences are taken. A different sample is taken each time. Each of the samples is run by using the three classifiers with 10-fold cross validation. The results with three different samples are averaged. The three classifiers along with their average values are shown in Table 4.

From Table 4, it is clear that IBK is performing better than Lib SVM and J48 in terms of Accuracy, Precision, Recall and F-measure.

5. Overall comparisons between the two approaches

5.1. Error analysis

In case of Lexicon-based approach, sentences are not identified correctly due to the following reasons. For ease in understanding, each Urdu example is first presented in Roman-Urdu to know the pronunciation of each word and then translated in English for clarifying the meaning of sentence.

1. A sentence is considered as a neutral sentence when the number of positive words and negative words are equal whereas in fact

Table 2
Accuracy level at different phases.

Specifications	Incorrect sentences (Num)	Correct sentences (Num)	Accuracy (%)
Positive and negative words	1574	4451	73.88
Positive, negative words along with negation handling.	1306	4719	78.32
Positive, negative words along with negations and intensifiers handling.	999	5026	83.42
Positive, negative words along with handling negations, intensifiers and context- dependent words.	661	5364	89.03

Table 3
Performance of Urdu Sentiment Analyzer.

Class	Evaluation Metrics		
	Precision (Float)	Recall (Float)	F-measure (Float)
Positive (Float)	0.89	0.95	0.93
Negative (Float)	0.95	0.79	0.87
Neutral (Float)	0.75	0.97	0.84
Average	**0.86**	**0.90**	**0.88**

Table 4
Performance of the three classifiers.

Classifier	Evaluation Merrics			
	Precision (Float)	Recall (Float)	F-Measure (Float)	Accuracy (%)
Lib SVM	0.67	0.65	0.65	65.00
KNN (IBK)	**0.68**	**0.67**	**0.67**	**67.02**
Decision Tree (J48)	0.66	0.63	0.60	62.50

Bold values highlights better performance than other classifiers in comparison.

the sentence has positivity or negativity. Consider the following examples.

i. مایوسی کو تقویت کس نے دی؟
Mayoosi ko taqwiyat kis ne di?
By whom the frustration was powered?

ii. یہ چوک پر کھڑے ہو کر حکومت کوسنے سے بہتر ہے
Yeh chowk par kharray ho kar hukoomat kosnay se behtar hai.
It is better than cursing the government while standing on the square.

In (i), مایوسی (Mayoosi i.e. frustration) is a negative word. After being searched in the Urdu Sentiment Lexicon, it receives −1 polarity. تقویت (taqwiyat i.e. powered) is a positive word, so +1 polarity is assigned to it. The sum of polarities is 0. Thus the sentence is actually negative but it is identified as neutral by the software. In (ii) کوسنے (kosnay i.e. cursing) is a negative word, therefore receives −1 polarity and بہتر (behtar i.e. better) is a positive word with +1 polarity. The sum of polarities is 0. Overall the sentence is positive but it is identified as neutral. Both the above sentences are incorrectly identified as neutral as number of positive and negative words is equal in each of the sentence.

2. A sentence is considered as a neutral sentence when there is no positive or negative word in it. The fact is that, a sentence may be positive or negative without the presence of any positive or negative word.

i. کچھ بولنے سے پہلے کم از کم سوچ ہی لیا کرو
Kuch bolnay se pehlay kam-az-kam soch hi liya karo.
At least think before you speak something.

ii. انسان آج بھی بکتے ہیں
Insaan aaj bhi bektay hain.
Even today, people give themselves up for money.

Although (i) is a positive sentence and (ii) is a negative sentence but both are incorrectly identified as neutral due to the absence of any positive word or negative word in each of the sentence.

3. A sentence is incorrectly identified as positive, negative or neutral due to sarcasm e.g. a word may be positive but used in a negative sense. Sarcasm is not handled currently in this research.

i. مولوی ہو کر کھلونے خریدتا ہے ؟
Molvi ho kar khilonay kharidata hai?
Buy toys while being a molvi (religious scholar)?

ii. مولوی آج پارک میں چہل قدمی کر رہا تھا
Molvi aaj park mein chehal-qadmi kar raha tha.
Molvi (religious scholar) was taking a walk in the park today.

Both(i) and(ii) are actually used in negative sense as there is sarcasm but (i) is incorrectly identified as neutral (due to the absence of any positive or negative word) and (ii) is incorrectly identified as positive due to the presence of چہل قدمی (chehal-qadmi i.e. walk) which is a positive word and l receives +1 polarity. The sum is equal to 1 which is greater than 0, therefore is incorrectly identified as positive sentence by the software.

4. A sentence is incorrectly identified as positive as there are more positive words than negative words in it whereas in fact it is negative.
   i. بے شک میرا اردو سے لگاؤ اپنی جگہ درست ہو لیکن میری شدت پسندی اس درست کو بھی غلط کر دے گی
   Be-shak mera urdu se lagao apni jagah durust ho lekin meri shiddat-pasandi is durust ko bhi ghalat kar day gi.
   No doubt, I may be rightly attached with Urdu but my extremism will make it wrong.
   ii. وہ خوشحالی کی روشن راہیں کہاں گم ہو گئیں؟
   Woh khushhali ki roshan rahein kahan gum ho gayeen?
   Where are the bright vistas of prosperity lost?

Both (i) and (ii) are identified as positive sentences whereas in fact both are negative sentences. In (i), four positive words i.e. بے شک (Be-shak i.e. no doubt), لگاؤ (lagao i.e. attached/attachment), درست (durust i.e. right/rightly) and درست (durust i.e. right/rightly) are used. Each word receives +1 polarity. Two negative words i.e. شدت پسندی (shiddat-pasandi i.e. extremism) and غلط (ghalat i.e. wrong) are used. Each negative word receives −1 polarity. The sum of polarities is 4–2 = 2, which is greater than 0. Thus the sentence is declared as positive. In (ii), two positive words i.e. خوشحالی (khushhali i.e. prosperity) and روشن (roshan i.e. bright) and one negative word i.e. گم (gum i.e. lost) are used. Each of the two positive words receives +1 polarity and negative word receives −1 polarity. The sum of polarities in this case is 2–1 = 1, which is greater than 0. The sentence is identified as positive sentence. As there are more positive words than negative words in each of the sentence therefore they are incorrectly identified as positive sentences.

5. A sentence is incorrectly identified as negative as there are more negative words than positive words in it whereas in fact it is positive.
   i. اس کے بعد ہم لوگ لوڈ شیڈنگ پر جلنے اور کڑھنے کے بجائے خوش ہوا اس کے Is ke baad hum log load-shedding par jalne aur kurhne ke bajaye khush howa karen ge.
   After this we will enjoy load-shedding rather than distressing and fretting.
   ii. شیطان کے مکر و فریب سے پناہ میں رکھنا
   Shetan ke makr-o-fraib se panah mein rakhna.
   Keep us safe from the evils (wickedness) of Satan.

Both (i) and (ii) are identified as negative sentences whereas in fact both are positive sentences. In (i), two negative words i.e. جلنے (jalne i.e. distressing) and کڑھنے (kurhne i.e. fretting) and one positive word i.e. خوش (khush i.e. happy (in the above case “enjoy”)) are used. Each of the two negative words receives −1 polarity respectively, whereas the single positive word receives +1 polarity. The sum of polarities is −2 + 1 = −1, which is less than 0. The sentence is therefore be identified as negative one. In (ii), again two negative words i.e. شیطان (shetan i.e. Satan) and مکر و فریب (makr-o-fraib, evils (wickedness)) and one positive word i.e. پناہ (panah i.e. protect (in the above case “safe”)) are used. The sentence is identified as negative after the process as discussed above. As there are more negative words than positive words in each of the sentence therefore they are incorrectly identified as negative sentences.

Thus, five reasons are identified in this section due to which sentences are classified incorrectly by Urdu Sentiment Analyzer.

In case of SML approach the main reason of incorrectly labeling the sentences as positive, negative or neutral is that the data is not from a single domain. If data is from multiple domains then a huge amount of data is needed to train the classifiers in all of these domains. The classifier can identify a new sentence correctly only if it is already trained for it. The classifiers are trained and tested using WEKA. Results are provided in the form of Accuracy, Precision, Recall, F-measure and other statistical measures. Error analysis in the form of sentence by sentence analysis is not possible.

5.2. Performance

1. In case of Lexicon-based approach, negations, intensifiers and context-dependent words are kept in separate files and are handled effectively by the algorithm, which helped in significantly improving the performance.
   In case of SML approach, all these i.e. negations, intensifiers and context dependent words are included in the final 39 attributes or features, but slight improvement in the performance is seen due to the inclusion of these three.
2. After comparing Accuracy, Precision, Recall and F-measure of both approaches as mentioned in Tables 1 and 2, it is clear that the Lexicon-based approach performed much better than the SML approach in this research, when Urdu Sentiment Analysis is performed in multiple domains. In SML approach, IBK is performing better than the other two classifiers that achieved 67.02% Accuracy, 0.68, 0.67 and 0.67 Precision, Recall and F-measure respectively. Whereas 89.03% Accuracy with 0.86 Precision, 0.90 Recall and 0.88 F-measure are achieved from Lexicon-based approach that is much better result than SML approach. Table 5 shows a comparison of performance of the three classifiers and the Urdu Sentiment Analyzer.

Fig. 2 shows a comparison between Accuracy of the three classifiers and Urdu Sentiment Analyzer
Fig. 3 shows a comparison of precision, recall and f-measure of the three classifiers and Urdu Sentiment Analyzer.

Table 5
Performance of the three classifiers and Urdu Sentiment Analyzer.

Supervised Classifiers/Un supervised Urdu Sentiment Analyzer	Accuracy (%)	Precision (Float)	Recall (Float)	F-Measure (Float)
Lib SVM	65.00	0.67	0.65	0.65
Decision Tree (J48)	62.50	0.66	0.63	0.60
KNN (IBK)	67.02	0.68	0.67	0.67
Urdu Sentiment Analyzer	**89.03**	**0.86**	**0.90**	**0.88**

Bold values highlights better performance than other classifiers in comparison.

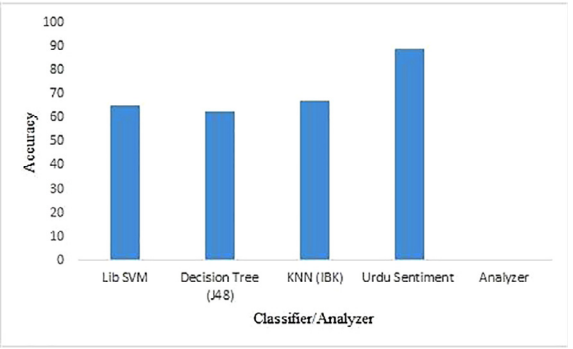

Fig. 2. Comparison of Accuracy.

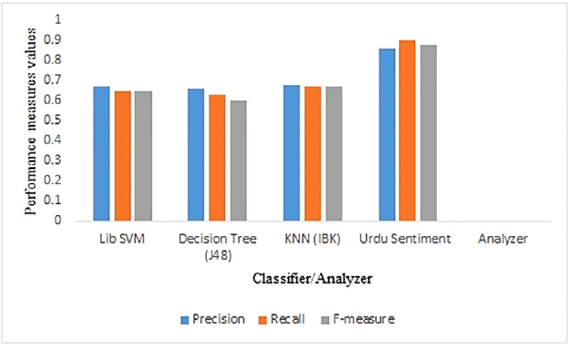

Fig. 3. Comparison of precision, recall and f-measure.

Table 6
A comparison with the previous work in Urdu SA.

Research work	Approach	Size of the tested data	Accuracy
Syed et al. (2011)	Lexicon based, Adjective Lexicon i.e. Urdu Sentiment annotated Lexicon	1350 reviews on two products	74%
Syed et al. (2014)	Same as above, noun phrases associated to Senti Units i.e. (sentiment carrier expressions)	Same as above	82.5%
Daud et al. (2014)	Lexicon based, Roman-Urdu Adjective Lexicon	1620 comments on 3 products.	21.1% comments were falsely categorized. So Accuracy is 78.9%
Rehman and Bajwa (2016)	Lexicon-based, 2607 negative and 4728 positive sentiment words for Urdu.	Text from news websites (Details are not available)	66.00%
Proposed Urdu Sentiment Analyzer	Lexicon based, Urdu Sentiment Lexicon for adjectives, nouns and verbs with 11,739 negative words and 9578 positive words along with negations intensifiers and context dependent words.	151 blogs with 6025 sentences	89.03%

A comparison of this Urdu Sentiment Analyzer and the previous work done in Urdu SA using Lexicon-based approach is provided in Table 6.

5.3. Consumption of resources

For the Lexicon-based approach, the collected data is only used for testing, otherwise it is not a pre-requisite. It can perform very well without any training. In case of SML approach training data is a pre-requisite as the classifiers need training on huge data to perform well.

Collecting a huge amount of data that covers all domains adequately is not an easy task. If the huge data is collected in multiple domains, annotation of data by authentic annotators and then finding the agreement between the annotators is a big task. The corpus with 6025 annotated sentences in this research by three annotators took 4 months only in the annotation phase. Even if a large annotated corpus is developed with reasonable coverage in multiple domains. Then training classifiers on such large corpus to get trained for classification is time consuming (in case of few classifiers).

In case of Lexicon-based approach, the lexicon developed is with a reasonably wide coverage. It took two months in developing this wide coverage lexicon and one month in developing and implementation of the algorithm. So, SML is more time consuming and costly than the Lexicon-based approach, for performing Urdu SA in multiple domains.

6. Conclusion

A comparison between the two approaches is presented in this research, where both the approaches are used. The accuracy of Urdu Sentiment Analyzer is 89.03% with 0.86 precision, 0.90 recall and 0.88 F-measure. The accuracy is raised from 73.88% to 89.03%. The reason is dealing with negations, intensifiers and context-dependent words effectively. These results show that these three aspects are used in Urdu text frequently and therefore should not be neglected while performing Urdu SA. Rather these should be dealt with care.

In case of SML, KNN (IBK) proved to be the best classifier in the light of experiments adopted in this research work. KNN (IBK) achieved 67.02% accuracy and 0.68, 0.67 and 0.67 precision, recall and f-measure respectively which is much lower than Lexicon-based approach.

Based on the experiments performed in this research, it is concluded that, Lexicon-based approach outperforms SML approach not only in terms of accuracy, precision, recall and f-measure but also in terms of taking less time and effort. Lexicon-based approach is performing better than SML approach in this research because a wide coverage lexicon and an efficient Urdu Sentiment Analyzer (on the basis of experiments performed) is developed that can efficiently handle data from different domains. SML approach can handle data from different domains efficiently only if enough data is collected in each domain which is quite time consuming process.

References

Ali, F., Kim, E.K., Kim, Y.-G., 2015. Type-2 fuzzy ontology-based opinion mining and information extraction: a proposal to automate the hotel reservation system. Appl. Intell. 42, 481–500.

Ali, F., Kwak, K.-S., Kim, Y.-G., 2016. Opinion mining based on fuzzy domain ontology and Support Vector Machine: a proposal to automate online review classification. Appl. Soft Comput. 47, 235–250.

Ali, F., Kwak, D., Khan, P., Riazul Islam, S.M., Kim, K.H., Kwak, K.S., 2017. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transp. Res. Part C: Emerging Technol. 77, 33–48.

Appel, O., Chiclana, F., Carter, J., Fujita, H., 2017. Successes and challenges in developing a hybrid approach to sentiment analysis. Appl. Intell.

Aue, A., Gamon, M. 2005. Customizing sentiment classifers to new domains: a case study. In: proceedings of the Recent Advances in Natural Language Processing (RANLP).

Augustyniak, Ł., Szymański, P., Kajdanowicz, T., Tuligłowicz, W., 2016. Comprehensive study on lexicon-based ensemble classification sentiment analysis. Entropy 18, 1–29.

Choi, Y., Lee, H., 2017. Data properties and the performance of sentiment classification for electronic commerce applications. Inf. Syst. Front.

D'Andrea, A., Ferri, F., Grifoni, P., Guzzo, T., 2015. Approaches, tools and applications for sentiment analysis implementation. Int. J.Comput. Appl. 125, 26–33.

Daud, M., Khan, R., Duad, A., 2014. Roman Urdu opinion mining system (RUOMiS). CSEIJ 4, 1–9.

Farooq, U., Mansoor, H., Nongaillard, A., Ouzrout, Y., Qadir, M.A., 2017. Negation handling in sentiment analysis at sentence level. J. Comput. 12, 470–478.

Gonçalves, P., Araújo, M., Benevenut, F., Meeyoung, C. 2013. Comparing and combining sentiment analysis methods. In: proceedings of the First ACM Conference on Online Social Networks.

Hailong, Z., Wenyan, G., Bo, J. 2014. Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. In: proceedings of the WISA '14, 11th Web Information System and Application Conference.

Hutto, C J., Gilbert, E. 2014. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14).

McHugh, M.L., 2012. Interrater reliability: the Kappa statistic. Biochem Med 22, 276–282.

Mukhtar, N., Khan, M.A., 2018. Urdu Sentiment Analysis using supervised machine learning approach. Int. J. Pattern Recogn. Artif. Intell. 32.

Musto, C., Semeraro, G., Polignano, M. 2014. A Comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. In: proceedings of the DART 2014, Information Filtering and Retrieval, 8th International Workshop on Information Filtering and Retrieval, CEUR Workshop Proceedings Pisa, Italy.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In: proceedings of the SemEval-2013 Workshop (SemEval 2013), Atlanta, Georgia, USA.

Rehman, Z U., Bajwa, I S. 2016. Lexicon-based Sentiment Analysis for Urdu language In The Sixth International conference on Innovative Computing Technology (INTECH 2016), pp. 497–501.

Rosenthal, S., Nakov, P., Ritter, A., Stoyanov, V. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In: proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014), Dublin.

Shoukry, A.M., 2013. Arabic Sentence level Sentiment Analysis (Master thesis). Department of Computer Science and Engineering, The American University in Cairo, Egypt.

Siegel, S., John Castellan, N., 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill.
Silva, C., Ribeiro, B 2003. The importance of stop word removal on recall values in text categorization. In: proceedings of the Neural Networks, 2003. Proceedings of the International Joint Conference.
Syed, A.Z., Muhammad, A., Enríquez, A.M.M., 2010. . Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. The Proceedings of the 9th Mexican International Conference of Artificial Intelligence. MICAI, Berlin Heidelberg.
Syed, A.Z., Muhammad, A., Enríquez, A.M.M., 2011. Adjectival phrases as the sentiment carriers in Urdu. J. Am. Sci. 7, 644–652.
Syed, A.Z., Muhammad, A., Enríquez, A.M.M., 2014. Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text. Artif. Intell. Rev., Springer 41, 535–561.
V-Díaz, PascualEspada, J., GonzálezCrespo, R., G-Bustelo, B.C.P., 2018. An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis. Appl. Soft Comput. 67, 822–833.
Viera, A.J., Garrett, J.M., 2005. Understanding inter observer agreement: the kappa statistic. Family Med 37, 360–363.
Zagibalov, T., 2010. Unsupervised and Knowledge-poor Approaches to Sentiment Analysis (Ph.D. thesis). University of Sussex.
Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B. 2011. Combining lexicon based and learning-based methods for twitter sentiment analysis: Hewlett-Packard Labs Technical Report, HPL-2011-89.