IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 December 2022, accepted 13 January 2023, date of publication 19 January 2023, date of current version 25 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3238207

RESEARCH ARTICLE

Semantic Orientation of Crosslingual Sentiments: Employment of Lexicon and Dictionaries

ARSLAN ALI RAZA[1,2], ASAD HABIB[1], JAWAD ASHRAF[1], BABAR SHAH[3], AND FERNANDO MOREIRA[4]

[1]Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 45550, Pakistan
[3]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates
[4]REMIT, IJP, Portucalense University, 4200-072 Porto, Portugal

Corresponding author: Asad Habib (asadhabib@kust.edu.pk)

This work was supported by the FCT—Fundação para a Ciência e a Tecnologia, I.P. [Project UIDB/05105/2020].

ABSTRACT Sentiment Analysis is a modern discipline at the crossroads of data mining and natural language processing. It is concerned with the computational treatment of public moods shared in the form of text over social networking websites. Social media users express their feelings in conversations through cross-lingual terms, intensifiers, enhancers, reducers, symbols, and Net Lingo. However, the generic Sentiment Analysis (SA) research lacks comprehensive coverage about such abstruseness. In particular, they are inapt in the semantic orientation of Crosslingual based code switching, capitalization and accentuation of opinionative text due to the lack of annotated corpora, computational resources, linguistic processing and inefficient machine translation. This study proposes a Heuristic Framework for Crosslingual Sentiment Analysis (HF-CSA) and takes into consideration the NetLingua, code switching, opinion intensifiers, enhancers and reducers in order to cope with intrinsic linguistic peculiarities. The performance of proposed HF-CSA is examined on the Twitter dataset and the robustness of system is assessed on SemEval-2020 task9. The results show that HF-CSA outperformed the existing systems and reached to 71.6% and 76.18% of average accuracy on Clift and SemEval-2020 datasets respectively.

INDEX TERMS Sentiment analysis, lexicon based methods, Urdu language processing, crosslingual orientation.

I. INTRODUCTION

The exponential growth of web-enabled technologies and mobile devices are continuously changing the general trends of online communication and collaboration. Mobile devices and their associated technologies have introduced multi-faceted forms of useful and attractive real world applications for sharing information, facts and sentiments. Consequently, these devices have become integral part of life for users belonging to all segments of society. The online publishers from diverse demographic areas are shifting towards Web 2.0 for communication and collaboration. The Microbloging websites provide one of the most enabling platforms to online users and organizations to generate, update, share and publish sentiments, ideas, suggestions and expressions. The statements and views posted by microbloging users about goods, services and other entities carry vital importance for data-scientists and business organizations.

Market Analyzers, business owners and multinational organizations invest in significant resources to know the feed-back of consumers regarding their products and services.

In past, the business trends and product acceptance rates were evaluated through traditional methods such as surveys. However, the fast and open-for-all social media channels have led to novel scientific techniques of sentiment analysis and user profiling. The statements, suggestions, speculations, ideas moods etc. published online are termed as opinions or sentiments. The techniques of Natural Language Processing and Text Mining are commonly used for Sentiment Analysis (SA). The purpose of this promising field of Data Science is the identification and orientation of users' opinions towards relevant domains [1]. SA techniques offer convenience to

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung.

both product users as well as the business organizations for restructuring their organizational strategies and policies. The business analysts employ opinion mining systems to generate pulse-reports in order to know ''what the general public thinks about the products or services offered by them. The big data analytics, opinion mining and Sentiment Analysis Techniques are expedient in discovering trends and patterns that otherwise remain hidden in the ever-growing online data about diverse domains such as social issues, politics, market, Consumer Confidence Prediction, information retrieval, basket analysis, products and services. These methods provide vibrant insights to market analysts in making fast and intelligent decisions. On the contrary, the contemporary trends in online public interaction have brought about novel and unique challenges to sentiment analysis researchers. In particular, the use of informal and cross-lingual opinionative content lowers the effectiveness and efficiency of SA systems to discover and classify public opinions [1], [2]. The general public may use demographically constrained terms, NetLingo, tags, emoji punctuations and linguistically borrowed tokens during their 'natural conversation'. The code-switching and multi-lingual lexica complement in making the task of sentiment analysis further challenging. Several research studies are conducted and a number of techniques are employed for classification of public opinions shared online in textual form. These techniques can be grouped into three main sentiment analysis paradigms; Supervised sentiment classification, Unsupervised classification and Lexicon-Based methods. The supervised and unsupervised techniques are useful for datasets containing pre-annotated, monolingual and standard texts. However, these methods do not yield satisfactory results for cross-lingual and raw (non-standard) public opinions [2]. It is also observed that the usefulness of generic algorithms is limited when applied to natural languages that suffer from scarcity of computational resources.

It is fact that sentiment analysis of cross lingual text has been an active research area and many researchers have already contributed actively but few language pairs have been lacking a considerable attention due to the scarcity of resources, target language structure and lingual peculiarities. Table 1. presents the significant gaps explored in existing research for cross lingual sentiment analysis of text. Keeping in view the existing research and lingual peculiarities the following two major gaps are considered to make system more significant and novel.

- Consideration of Informal and Code mixed opinions in cross lingual text.
- Contextualized transformation of Urdu opinions along with enhancer and reducer.

This study proposes a lexicon based solution for cross-lingual and informal opinion bearing text. Development of Urdu-English bi-lingual annotated SA lexicon and classifier for cross-lingual sentiment analysis is the novelty of our proposed framework. We intend to verify the effectiveness of our proposed framework on a twitter dataset of informal, Crosslingual, code switched and Anglicized opinion bearing text and robustness of system is assessed by comparing it with existing state of art results for SemEval-2020 task 9 [3], [4].

This article aims to address the following research objectives in order to cope with above mentioned gaps;

- Creation of Urdu-English bi-lingual annotated SA lexicon used in the transformation of Crosslingual sentiment terms.
- Provision of a heuristic framework which can identify, extract and computationally annotate informal as well as cross-lingual opinion bearing tokens along with enhancer, reducer and context shifters.
- Formation of an improved mechanism for text normalization and classification of Urdu-English cross-lingual tokens.

The optimized text normalization and consideration of Urdu/Roman Urdu/ English tokens along with opinion shifter enhances the coverage of sentiment orientation. The experimental evaluation presented in section IV revealed it clearly that consideration of these key objectives played a significant role in semantic orientation of Crosslingual sentiments.

The article is structured as follows. In section II we stated the literature review of Crosslingual SA, Section III presents proposed material and methods, section IV demonstrates the experimental outcomes and Section V concludes the work.

II. LITERATURE REVIEW

Sentiment Analysis is the identification and recognition of feelings and expressions of users, publically shared in textual form. Nasukawa and Yi [5] first coined the term 'Sentiment Analysis (SA)' in 2003 in their research study on detecting the favorability via public sentiments. Second variation used is fast text pre-trained Urdu embedding to obtain word vectors, which are available online [6]. It pro-vides word vectors of 300 dimensions trained using CBOW model. They used rule-based NLP techniques and ML algorithms on opinionative data for analysis and review of organizations and their services. Similarly, the term ''Opinion Mining (OM)'' was first appeared in 2003 for Opinion orientation and semantic analysis of product reviews [7]. Semantic Orientation of user generated contents has great importance not only to observer and analysts of numerous organization but socio monitoring can benefit in reshaping the Decision Making Process [8], [9], [10]. Natural Language Engineering proved to be the key source in the process of sentiment analysis and text normalization for formal as well as informal opinion bearing text [11], [12]. Beside the existence of formal and standard opinionative contents, the contemporary user collaboration contains colloquial as well as non-standard multilingual content, which poses a number of challenges in mining feelings and moods [13].

The extraction and summarization of these cross-lingual opinions is a complicated task. As mentioned above, the three focal paradigms for digital orientation of users' sentiments into generic clauses (positive, negative and neutral) are

TABLE 1. Crosslingual sentiment analysis: A perspective of its past and present.

Author	Languages	Method	Gap Analysis/ Remarks
Kia Dashtipour et al. [14].	English German Italian French Arabic Chinese	Unsupervised Semantic Orientation based Pointwise mutual information & hybrid combination of lexicons and corpora is utilized.	Low Accuracy is observed due to lack of lexical resources.
Lucas Brönnimann et al. [15].	English German Italian French	Lexicon based Method	Additional linguistic processing is required to understand the figures of speech.
Mohamed Abdalla [16]	English Chinese Spanish	Logistic Regression and Linear SVM	Single linear transformation mishandled the text due to inaccurate machine translation.
Ruifeng Xu et al. [17]	English Chinese	Multi Kernel SVMs	Transfer learning lacks the implicit information.
Amiri et al. [18]	English Informal	Supervised	Low Accuracy is observed for informal tokens due to Scarcity of target language resources.
Bilal, M et al. [19]	Roman Urdu	Naïve Bayes, Decision Tree, KNN	Inefficient outcomes as Code Switching, Emoticons, emphasizer and formal English opinions are ignored
Pelicon, A et al. [4]	Slovenian Croatian	Deep Learning Based mBERT model	Requires Fine Tuning Inapt for code switched & Anglicized Text
Ghulam, H et al. [20]	Roman Urdu	Deep Neural NW-LSTM	Good on sequential data but inapt at informal and Crosslingual contents.
Kanclerz, K et al. [21]	English Polish Russian Spanish	Convolutional Neural Network Bidirectional Long Short Term Memory	Translated version of language lack coverage of target language context.
Singh, P et al. [22]	English Hindi	Transfer Learning with word embedding	Domain Specificity & Post processing is required with monolingual embedding to reach robust performance.
Almansor, M et al. [23]	English Arabic	Clustering-based bee-colony-sample selection method	Target based feature weighting is expensive in context of resource poor language.

based on the supervised, semi-supervised and unsupervised approaches. Kiritchenko et al. [9] tested a supervised ML system on informal SMS and tweets in which tweet based opinionative lexica are generated through emoticons and tags. Amiri and Chua [18] proposed an optimization algorithm for SA of urban and slang terms. Sarker and Gonzalez [24] handled non-standard subjective tweets using a Support Vector Machine (SVM) based classifier.

Lately, the multi-lingual and cross-lingual SA and negation handling gained noticeable attention due to popularity of the microblogging sites spawned by diverse areas of users [25]. Machine Translation (MT) systems, bilingual vector space embedding and multilingual lexica have been tested for multilingual SA. Single Linear Transformation (SLT) technique is employed to identify the cross-lingual sentiments in English, Spanish and Chinese [16]. The extracted text is first translated into English and then SA is carried out. Lucas Brönnimann [15] performed SA on multi-lingual tweets of Swiss politicians using dictionaries and universally comprehensible emoticons that do not depend upon any specific natural language. Dashtipour et al. [14] tested 11 methods on two corpora and related their low precision to the lack of information provided in the existing studies and datasets. Mozetič et al. [26] developed classifiers using manually annotated tweets in 13 languages and tested them with 6 different classification algorithms. Their study concluded that; i) there is no explicit difference in the performance of classifiers and, ii) the classification results are proportional to the refinement and volume of training data. In stark contrast to the manual annotation of tweets, Adel et al. [27] proposed a model to handle the cross-lingual SA of Arabic text without tagging. They employed feature reduction mechanism to find the desired solutions. A SentiUnit is a baseline approach used in sentiment analysis of product and movies reviews, in which the linguistic structure, grammar, morphology and technical aspects of Urdu language are highlighted [28]. Bilal et al. [19] used Naïve Bayes, Decision Tree and KNN for the semantic orientation of Roman Urdu Opinions in which they concluded that NB produced more efficient result than DT and KNN but they ignored the other formal and informal English language opinionative contents. Ruifeng et al. [17] used Multi-Kernel SVMs in cross-lingual SA of English and Chinese. They reported that opinion holder extraction is a prominent indicator in cross-lingual SA.

Similarly, Translation of one human language into another via computer mediated devices requires proper computational knowledge as well as linguistic resources. Parallel corpora are one of such resource to many applications of natural language processing which play prodigious role in machine translation. Although its availability is scarce due to limit of size and quality of vocabulary coverage but it plays vital role in MT system. Initially text for statistical machine translation (SMT) tasks were composed opportunistically via web resources [29]. Now, it is indispensable to utilize parallel text corpora with few semantic rules in order to improve SMT systems. Such corpora are the prerequisites of many research activities like machine translation, multilingual analysis, and multilingual high range lexicons creation [30].

Despite the fact that parallel corpus plays significant role in sentiment orientation of publics' attitudes but it is dif-

ficult to create a corpus for scare resource language [31]. It has been observed that few languages are considered as rich due to the availability of parallel corpus along with other lingual resources whereas on the other side languages lacking these lingual materials are termed as resource poor language [32]. Gale and church [33] in 1991 compiled the first ever parallel corpus for English-French language pair. Similarly, few high quality parallel corpora are publically available for Hungarian-English, Dutch-English, Bulgarian-English, Czech-English, Greek-English, Spanish-English, Italian-English, Portuguese-English, Romanian-English, German-English and Swedish-English [31], [34], [35], [36]. Although there is massive amount of parallel corpora for resource rich languages however it is lacking for under-resource language such as Laos, Vietnamese and Urdu, further machine translation for such language pair is suffering due to unavailability of these useful resources [37]. Machine translation depends not only on the quantity of parallel corpora but quality is also a big factor in translating one language into another. In fact, an average machine translation system needs up to 100K sentences and parallel corpus of 50M-1000M words [38]. Machine translation can be performed using rule based, example based, statistical and hybrid technique [22]. Rule based technique uses syntactic and semantic rules with the utilization of lexicons, dictionaries and corpora, whereas statistical and example based techniques use parallel corpora while on the other hand hybrid method combines the best features of both statistical rule based technique.

Table 1 summarizes the comprehensive literature review conducted to unfold the gaps of Sentiment Analysis in Crosslingual setting in order to make things more rational and transparent. It provides the list of gaps, methodologies and languages covered so far in cross lingual sentiment analysis. Existing work for cross lingual clearly explored that experiments on CLSA are conducted only for few language pairs as it mainly takes English as source and Chinese, Spanish, German, Japanese and French as target language which ultimately limits the orientation and promotion of CLSA for scarce resource language such as Urdu, Hindi and Punjabi etc. Further it is observed that the accentuation of opinions, enhancers, reducers and context shifters were also ignored in sentiment orientation of Crosslingual setting. Keeping in view the identified gaps a heuristic framework is proposed to cope with the consideration of informal, accentuated and code-mixed opinions in Crosslingual setting which is proved as significant solution in identification and orientation of cross lingual as well as informal contents of English-Urdu language pair.

III. MATERIALS AND METHODS

A framework is proposed to find the public sentiments of cross-lingual text by employing the lexicon and dictionary based methods. Figure 1 illustrates the flow of proposed heuristic framework for Crosslingual sentiment analysis. HF-CSA is comprised of following essential steps;

A. DATASET PREPARATION AND LEXICON COMPILATION

The input text for classification purpose is extracted from social media sites but extracted text is not always in desirable form of classification, as there are numbers of tags, symbols and undesired data associated with it. Text must be preprocessed before going towards classification process. The efficiency of sentiment classification is based on the quality of text. Text Preprocessing is the essential phase of each sentiment classification process. We can't mine public moods accordingly if the source text is not clean. Therefore, dataset preparation is performed in the following manners;

1) DATASET PREPARATION

It is mentioned earlier that the major objective of this study is to explore the cross lingual, informal and code-mixed contents of English-Urdu language pair. So, the tweets and sentences relevant to English, Urdu and roman Urdu opinion bearing sense along with opinion enhancers and reducers are marked in the inclusion criteria. Therefore, search queries for above mentioned inclusive criteria is applied and we reached to SemEval-2020 (Hinglish) and Clift datasets. The description of Datasets used in the assessment of HF-CSA is elaborated as follows;

1) ClIfT_Dataset (Crosslingual and Informal Text)
2) SemEval-2020 Task 9

1) ClIfT_Dataset (Crosslingual and Informal Text).
Data having formal, informal and cross-lingual based domain specific text is extracted using Twitter streaming APIs and Web Crawlers. Open Source Python Package (OSPP) ''Tweepy'' is used to get opinion bearing text from twitter and publically available ''Twint'' an OSINT tool is used for advance scraping of tweets. Similarly, Octoparse is employed to get opinion bearing contents from blogging websites. A search term list having domain related tokens is created to gather relevant contents. Different combinations of search terms are applied to reach the in-depth relevancy of input text and initial noise like URLs, etc. are removed at the time of extraction by just integrating regular expressions. Further Search Term List is appended by embedding NetLingua, Crosslingual and informal tokens to get more pertinent dataset.
A dataset (having multiple CSV files) is generated according to domain specific search queries, we named this dataset as ClIfT_Dataset (Crosslingual and Informal Text). A total of 15486 tweets covering (Jan 2019 to July 2019) are extracted for product reviews including Mobile Phones (Samsung, Vivo, Oppo and IPhone). Review sites and microblogs are used as data source for extracting the desired information whereas keywords, hashtags and accounts are used to target the desired domains.
The input text is not always in desirable form of classification, as there are numbers of tags, symbols and undesired data associated with it. Text must be

TABLE 2. Class wise distribution of SemEval 2020-task 9 dataset [2].

Language	Split	Total	Positive	Neutral	Negative
Hinglish	Train	14,000	4,634 (33.10%)	5,264 (37.60%)	4102 (29.30%)
	Validation	3,000	982 (32.73%)	1,128 (37.60%)	890 (29.67%)
	Test	3,000	1,000 (33.33%)	1,100 (36.67%)	900 (30%)
	Total	20,000	6,616 (33.08%)	7,492 (37.46%)	5892 (29.46%)
Spanglish	Train	12,002	6,005 (50.03%)	3,974 (33.11%)	2,023 (16.85%)
	Validation	2,998	1,498 (49.96%)	994 (33.15%)	506 (16.87%)
	Test	3,789	3,061 (80.78%)	206 (5.43%)	522 (13.77%)
	Total	18,789	10,564 (56.22%)	5,174 (27.53%)	3,051 (16.23%)

preprocessed before going towards classification process. The efficiency of sentiment classification is based on the quality of text. Text Preprocessing is the essential phase of each sentiment classification process. We can't mine public moods accordingly if the source text is not clean. Extracted contents are then passed to next phase for linguistic preprocessing.

2) SemEval-2020 Task 9 Dataset
Code mixing and crosslingualism is common in the region having multiple or at least bilingual speakers. According to a survey there exist 630 million speakers of Hindi and Urdu whereas India and Pakistan has 30 different languages with more than 1 million speakers [36]
In SemEval-2020 task 9 codemixed tweets of Hinglish (Hindi-English) and Spanglish (Spanish-English) language pair are presented [2]. They released two public corpora for research community. The data is extracted from social networking channels and a huge volume of 20k and 19k tweets are collected and annotated for Hinglish and Spanglish dataset respectively. They named this task as Sentimix which aims to predict the sentiment of a given code-mixed tweet. A word and phrase level orientation of positive and negative tweets of both language pair has been performed and text is annotated with lingual as well as polarity labels. Table 2 presents the statistics of SemEval-2020 task 9 datasets. Here in this research Hinglish (Hindi – English) has been considered for the assessment of HF-CSA as proposed study aims to cope with roman Urdu-English language pair and romanized version of Hindi and Urdu has the similar transliteration.

2) LEXICON COMPILATION

The proposed lexicon is novel contribution of this study as sophisticated lexical resource for a combination of informal and Urdu-English crosslingual pair is lacking. In this section lexicon adaptation has been discussed. The contextualized informal and Crosslingual lexicon is compiled using two core steps.

1) Identification and collection of Crosslingual tokens for target language pair.
2) Assignment of contextualized definition polarity labels to each opinionative token via adaptation of existing resources for more relevant scoring of sentiments.

1) **Identification and collection of Crosslingual tokens for target language pair.**

TABLE 3. List of resources utilized in lexicon compilation.

S.No	Resource Description	Reference of Resource
1	A collection of 1.6k Urdu words	https://drive.google.com/file/d/0B9eF-UfzuXjUbF80aXpyck1fQ1k/edit?resourcekey=0-0gjJtYLCiQQ5Ti2oKlm8Q
2	Urdu words in six different domains by CLE	https://www.cle.org.pk/software/ling$_r esources/UrduHighFreqWords.htm$
3	Urdu WordNet by CLE	https://www.cle.org.pk/software/ling$_r esources/UrduWordNetWordlist.htm$

The major consideration adopted in the identification and collection of Crosslingual opinionative tokens is utilization of frequently used Urdu English opinionative terms from social media sites. In addition to these available social networking channels, a huge collection of NetLingua, Crosslingual and informal opinion bearing terms are fetched through existing resources. Table 3 presents the links of resources.

$$LCT = CLT1(CLW1, CLW2 \ldots CLWn) \ldots CLTn(CLW1, CLW2 \ldots CLWn) \quad (1)$$

$$LST = ST1(SW1, SW2 \ldots SWn) \ldots STn(SW1, SW2 \ldots SWn) \quad (2)$$

$$LIT = IT1(IW1, SW2 \ldots SWn) \ldots ITn(SW1, SW2 \ldots SWn) \quad (3)$$

where, LCT = List of Crosslingual Terms, LST = List of Slang Terms, LIT = List of informal Terms Similarly, CLT = Tweet having Crosslingual terms, ST = Tweet having slang terms IT = Tweet having informal terms. A unified list of eq.1, 2 and 3 is created by merging the extracted Crosslingual, slangs and informal words in order to assign appropriate definition and annotation.

2) Assignment of improved definition and polarity labels to each opinionative token via adaptation of existing resources for more relevant scoring of sentiments.
The assignment of improved definition is ensured via lingual preprocessing and part of speech tagging whereas scoring and polarity labels are assigned through the adaptation of publically available sentiment lexica SentiWordNet SWN [39].
The entries of proposed lexica are mapped with SWN in the following manner.

$$Ei =< T, Trans, swn.id, PoS, Scr, Pol > \quad (4)$$

where T, trans, swn.id and PoS represents term, transliteration, SentiWordNet-ID and Parts of Speech respectively. Similarly, Scr and Pol represents score and polarity respectively. The objective (a.k.a non-opinionated) tokens are excluded in subjectivity classification. A word or phrase can be marked as subjective if it conveys some positive or negative sentiment otherwise it is treated as objective. Part of speech tags and subjective information are employed to filter and classify the opinionative contents. Table 4 presents the partial list of lexical entries of proposed Crosslingual lexicon.

TABLE 4. Partial list of lexical entries of proposed Crosslingual lexicon.

S.No	Term	Transliteration	PoS	Score	Polarity
1	Aala	Awesome	[('awesome', 'JJS')]	0.5	Positive
2	Zabaradast	Great	[('great', 'JJS')]	0.5	Positive
3	Behtreen	Best	[('best', 'JJS')]	0.75	Positive
4	Kamal	Super	[('super', 'NN')]	0.5	Positive
5	Bakwas	worst	[('worst', 'JJS')]	-0.5	Negative
6	Farigh	useless	[('useless', 'NN')]	-0.375	Negative
7	Fazool	waste	[('waste', 'NN')]	0.125	Positive
8	Umda	Nice	[('nice', 'JJ')]	0.875	Positive
9	kharab	Spoiled	[('spoiled', 'VBN')]	-0.5	Negative
10	Tabahi	Amazing	[('amazing', 'VBG')]	0.875	Positive

In order to associate semantic orientation to each opinionative category we adopted SentiWordNet due to its high volume of opinion categories and frequent updates of words along with its senses. SWN associates three polarity labels: Positive, Negative and Neutral to each opinion category based on the orientation and semantic value of opinion word. The semantic orientation ranges between 0.0 to 1.0 and average semantic orientation for term having multiple senses is computed in the following manners;

$$Pos\_Score(word) = \frac{1}{ns}\sum_{(i=0)}^{n} Pos(i) \tag{5}$$

$$Neg\_Score(word) = \frac{1}{ns}\sum_{(i=0)}^{n} Neg(i) \tag{6}$$

$$Neu\_Score(word) = \frac{1}{ns}\sum_{(i=0)}^{n} Neu(i) \tag{7}$$

where Pos_Score, Neg_Score represents the positive, negative score of word and ns denotes the number of senses appeared in SWN against each target (Searched) term. Similarly, dominant semantic orientation of SWN is accessed to reach the relevant polarity of a term and then it is labeled as positive, negative and neutral based on its semantic score as shown in Eq. 5, 6, 7 and 8. In addition to this, proposed lexica is refined via exclusion of the neutral (non-opinionative) tokens on the basis of dominant objective value.

$$SemO^{swn}(w) = \begin{cases} SemO^{+} & \text{if } max(SemO^{+}, SemO^{-}, SemO^{n}) \\ & = SemO^{+} \\ SemO^{-} & \text{if } max(SemO^{+}, SemO^{-}, SemO^{n}) \\ & = SemO^{-} \\ \text{else} SemO^{n} \end{cases} \tag{8}$$

where SemOswn represent the semantic orientation of each word accessed from SWN using eq. 5, 6, 7. Similarly SemO+, SemO− SemOn presents semantic orientation of positive, negative and neutral words.

B. LINGUISTIC PREPROCESSING

Linguistic Preprocessing and Text Normalization is performed on extracted contents in order to produce quality input. Formal and informal opinion indicators such as verbs, adverbs, adjectives, NetLingo, anglicized, enhancer, reducer, cross-lingual Urdu terms and emoji are considered in the scope of desired input. The stop words and other lexemes are discarded as noise. The normalization is carried out in the following manner; Tokenization of Extracted Text: The first and foremost step is to tokenize the extracted content into sentences and subsequently into tokens. Python toolkit is used for tokenization of data. Stop Words Removal (SWR): Frequently used terms having no significance in sentiment orientation are removed for the sake of fast processing. In addition to the commonly used stop words, each domain has its own list of stop words. Table 5 presents the example of initial preprocessing steps. The Python Natural Language Toolkit containing corpora in multiple languages is used for the desired stop words removal. Lemmatization: The tagged sentences/tokens are passed to lemmatization phase for removing ambiguities of inflected forms. Lemmatization is the process of converting inflected forms into root and base forms. In this phase, all the inflected tokens are converted into base form using WordNet Lemmatizer.

TABLE 5. An example of initial preprocessing steps.

Sentence	@umar S21 an awsm mbl of Samsung with aala camera , greatttttt features
Step I Noise Removal	S21 an awsm mbl of Samsung with aala camera greatttttt features
Step II Stop Words Removal	S21 awsm mbl Samsung with aala camera greatttttt features
Step III Tokenization	'S21', 'awsm', 'mbl', 'Samsung', 'with', 'aala', 'camera', 'greatttttt', 'features'
Sentence	@umar S21 an awsm mbl of Samsung with aala camera , greatttttt features

C. LINGUISTIC PROCESSING AND SEMANTIC ORIENTATION OF CROSSLINGUAL SENTIMENTS

Linguistic processing involves the annotation and classification of formal, informal, NetLingo and Crosslingual opinion bearing terms. It involves the following essential steps;

Cross-lingual, implicit and Informal Text Identification (CLTI): In this phase, beside the consideration of formal English opinion words, cross-lingual words and sentences are identified and categorized for appropriate classification. A cross-lingual (English and Urdu) lexicon is employed in the identification of tokens which remain unidentified in the previous phases of normalization. The extracted tokens from cross-lingual corpora are included in the preprocessed dataset for subsequent experimentation. Similarly, slangs, NetLingo and anglicized terms are defined via utilization of manually compiled informal resources. The Python, Natural Language toolkit (NLTK) is used to import cross-lingual corpora and lexicons in order to identify the Urdu language text. Crosslingual term identification for Urdu and informal text is one of the core contribution of proposed HF-CSA.

D. IDENTIFICATION AND LABELLING OF POLARITY ENHANCERS AND REDUCERS

Twitter data comprises of opinion enhancers as well as reducers. Frequently used emoticons having positive expressions and character emphasization of target opinion terms are treated as enhancers whereas negators and context shifters act as polarity reducer. This study followed explicit lists of

TABLE 6. Description of opinionative features.

S.No	Opinionative Feature	Example Token	Remarks
1	Formal Opinion	Love & Worst etc.	Meet the English standards
2	Informal Opinion	Awsm & grt etc.	Nonstandard in nature
3	Crosslingual Opinion	Aala & kamal etc.	Other than English terms
4	Anglicized Opinion	Owesome & Sooper	Pronunciation specific misspell
5	Emphasized Opinion	Greatttt & supppper	Intensified English terms
6	Emoticons	:-) & :-(	Language Neutral Punctuations

frequently used enhancer and reducer and emphasized score is used for the target accentuated term.

E. SYNTACTICAL TAGGING (PoS TAGGING)

The role of proper annotation and Syntactical PoS tagging is vital in Word Sense Disambiguation (WSD). This step is significant in sentiment classification as it helps us to decide whether a lexeme or piece of text is opinionative or not. Python The NLTK has its own PoS tagger list but we utilized senses of SWN to reach more relevant orientation of target word.

F. SUBJECTIVITY CLASSIFICATION

Subjectivity classification is performed in order to separate the opinionative contents from non-opinionative one. Target sentence is scanned for subjectivity in which a sentence is marked as subjective if it contains one or more opinionative tokens either formal or informal including Urdu language terms, otherwise the sentence is marked as objective. Table 6 presents the description of key opinionative features. Subjectivity lexicon is used for identification and classification of subjective and objective contents.

G. SENTIMENT SCORING OF FORMAL, INFORMAL AND CROSS-LINGUAL TEXT

Semantic orientation of sentiments is attained via lexicon based strategy. Lexicon based algorithm carry out classification on the basis of sentiment lexica in which the sentiment scores decide the positivity and negativity of a sentiment. In this method of classification, each sentiment word is assigned a positive or negative semantic orientation (SO) value based on the target lexicon's score and a statistical rule based strategy is adopted by assigning score to each opinion indicator in order to calculate the sentence or document level polarity. As in past, Taboada et al. [40] built a lexicon based algorithm namely ''SO-Calculator'' in which lexicons are used for semantic orientation of reviews on the basis of opinion indicators. Similarly, Turney [6] utilized lexicon based semantic orientation algorithm for rating of positive and negative reviews. Their algorithm composed of three steps; (i) Extraction of phrases having adjectives and adverbs, (ii) Estimation of semantic orientation (SO) using PMI-IR (Pointwise Mutual Information and Information Retrieval) (iii) Classification on the basis of average SO. The preprocessed subjective text is the input of cross-lingual identification phase for semantic orientation of formal, informal and Crosslingual sentiments. However, the sentiment resources, and Crosslingual lexica is used for assigning score to individual tokens by searching the sentiment score of SentiWordNet (SWN). As mentioned earlier SWN is publically available lexical resource used in scoring of sentiments and opinions [39]. It is actually the extension of WordNet. It assigns three labels; Positive, Negative and Neutral to each synset of WordNet. Our experimental setup revealed that lexicon based algorithm has high technical viability and coverage over supervised learning in identifying the abstruseness of text such as enhancers, reducers and cross- lingual sentiment analysis.

H. CROSS-LINGUAL BASED OPINION SUMMARY

In the last phase of sentiment analysis, the aggregated polarity is computed for a complete sentence or Tweet.

$$Tweet_{Score} = \sum_{(i=1)}^{n} SO(Std_{Ops_i}) + SO(N_{std_{Netlingos_i}}) + SO(Crosslings_i) \quad (9)$$

where Std_{Ops}, $N_{stdNetlingos}$ and Crosslings represent Standard, Non-Standard, Informal and Crosslingual tokens respectively. Eq.9 shows that the tweet Sentiment Score is the sum of scores of each standard, non-standard, NetLingo and cross-lingual opinionative indicators appeared in each Tweet.

$$Crosslingual_{Opinion_{Summary}} = (Tweet_S core)/SC \quad (10)$$

where SC denotes the Sentiment Counter which is actually the count of sentiments in each target tweet. It parallelly counts the number of sentiment indicators appeared in each tweet in order to regulate the score with given range as score for each tweet ranges between -1 and $+1$.

Crosslingual and NetLingo score is calculated using SentiWordNet, the term is first searched in SWN then relevant score is calculated as per following criteria;

$$Pos\_Score(word) = 1/n \sum_{(i=0)}^{n} Pos(s_i) \quad (11)$$

$$Neg\_Score(word) = 1/n \sum_{(i=0)}^{n} Neg(s_i) \quad (12)$$

where n denotes the number of senses appeared in SWN against each target (Searched) term.

IV. EXPERIMENTAL EVALUATION AND RESULTS

This section presents the outcomes and comparative analysis of HF-CSA (Heuristic Framework of Crosslingual Sentiment Analysis). The experimental performance is evaluated on ClIfT_Dataset (Crosslingual and Informal Text). A total of 15486 tweets covering (Jan 2019 to July 2019) are extracted for product reviews including Mobile Phones (Samsung, Vivo, Oppo and IPhone) and robustness of proposed system is ensured on SemEval-2020 Task-9.

The detailed description of dataset is mentioned in section III,A and statistics of ClIfT_Dataset and SemEval-2020 Task-9 of Hinglish datasets is shown in Table 7 and Table 8 respectively

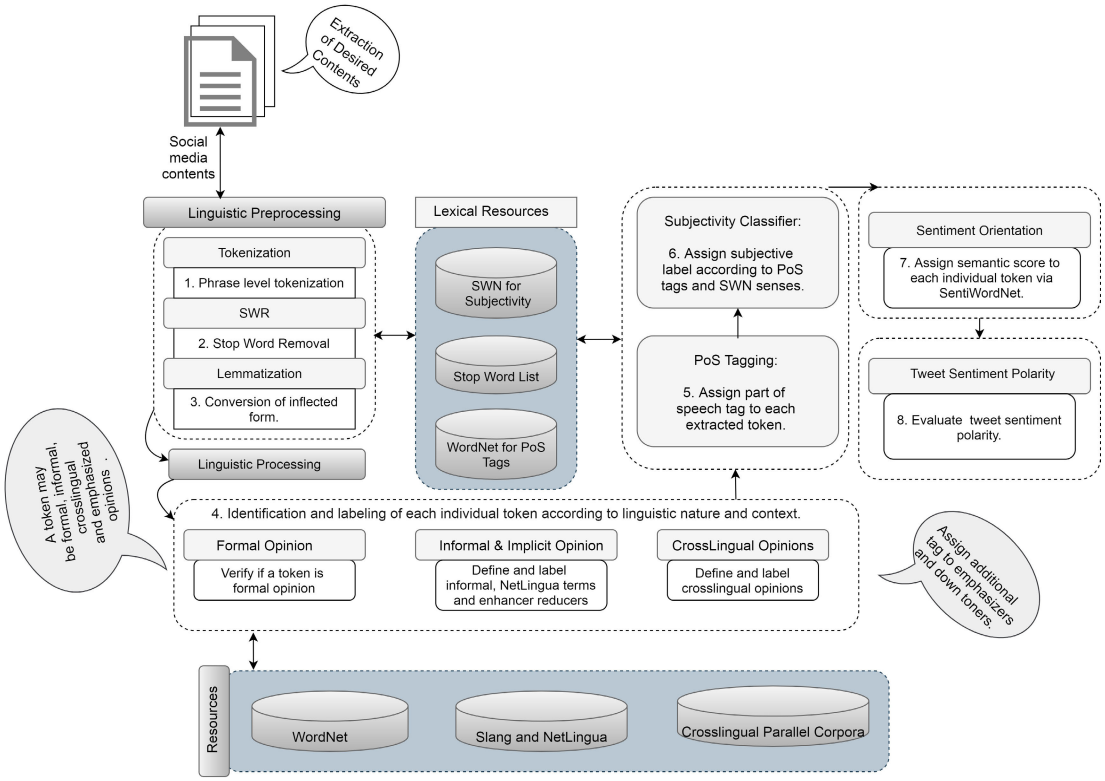

FIGURE 1. Proposed model.

TABLE 7. Statistics of CIIfT_dataset.

Categories	Informal and Anglicized	Informal and Crosslingual text	Total Set of Tweets
Positive	5000	2500	7500
Negative	5000	2500	7500
Total	10,000	5000	15,000

TABLE 8. Statistics of SemEval-2020 Task-9 [2] of hinglish dataset.

Language	Split	Total	Positive	Neutral	Negative
Hinglish	Train	14,000	4,634 (33.10%)	5,264 (37.60%)	4102 (29.30%)
	Validation	3,000	982 (32.73%)	1,128 (37.60%)	890 (29.67%)
	Test	3,000	1,000 (33.33%)	1,100 (36.67%)	900 (30%)
	Total	20,000	6,616 (33.08%)	7,492 (37.46%)	5892 (29.46%)

As mentioned earlier the performance of HF-CSA is assessed on Clift and SemEval-2020 datasets and state of art studies have considered standard evaluation parameters; Precision, Recall, F-Measure and Accuracy. The rationality behind adopting these metrics is to make our comparison more adequate and transparent with existing studies.

A. PRECISION

In sentiment classification precision signifies the probability of relevant set of tweets among the total number of retrieved tweets.

$$PPV = \frac{TP}{(TP + FP)} \; and \; PNV = \frac{TN}{(TN + FN)} \tag{13}$$

where PPV and PNV denotes Precision for Positive and Precision for Negative respectively. The lower precision indicates that high numbers of negatives tweets are labeled as positive while a higher precision means less number of negative tweets are incorrectly labeled as positive.

B. RECALL

In sentiment classification it signifies the probability of retrieved tweets that are relevant.

$$RPV = \frac{TP}{(TP + FN)} \; and \; RNV = \frac{TN}{(TN + FP)} \tag{14}$$

where RPV and RNV denotes Recall for Positive and Recall for Negative respectively. Recall usually measures the correctly classified tweets from total numbers of tweets classified The higher recall indicates that less numbers of positive tweets are incorrectly labeled as negative.

C. F1-MEASURE

F1-Measure signifies the harmonic mean of Precision and Recall mathematically denoted as below;

$$F1 - Measure = \frac{2PR}{(P + R)} \tag{15}$$

where P and R denotes Precision and Recall respectively.

D. ACCURACY

Accuracy in SA signifies the state of being correct in terms of performance.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{16}$$

Algorithm 1 CrossLingual Sentiment Scoring

```
Input: tweet
Output: Semantic Orientation
Emoji List: List of Emoji Icons                 # E.g. = {☺, ☹ ..... ..}
Enhancer List: List of emphasizers              # E.g: = {grnrreat, worssssst, SUPERBBB .....}
Adverbs List: List of adverbs                   # E.g. = {really, very ..... }
Negation List: List of Negators                 # E.g. = {not, aren't, couldn't ..... ..}
Function Semantic Score (text)
        cleaned_text = Preprocessor (tweet)
        words = tokenizer (cleaned_text)
        Tweet_Score = 0
        Sentiment_score = 0
        Crosslingual_Opinion_Summary = 0
        SC = 0                          #SC = Sentiment Counter
        ##Steps
        (i) if word is opinionative then increase sentiment counter
        (ii) if word is in Enhancer_List then enhance the polarity of word
        (iii) if word is in Adverbs_List then emphasize the polarity of word
        (iv) if word is in Negation_List then reverse the polarity of Sentence
        For word in words
                If word exists in sentiment_lexicon then
                        Sentiment_score = sentiment_lexicon (word)
                        Perform step (i) to (iv)
                else If word exists in Netlingo_lexicon then
                        Sentiment_score = Netlingo_lexicon[1] (word)
                        Perform step (i) to (iv)
                else if word exists in crosslingual_lexicon then
                        Sentiment_score = crosslingual_lexicon[2] (word)
                        Perform step (i) to (iv)
                else if word exists in Emoji_List then
                        Sentiment_score = Emoji_List (word)
                        SC= SC+1
                else
                        Sentiment_score = 0
                end if
                        Tweet_Score + = Sentiment_score
        Next
Crosslingual_Opinion_Summary = Tweet_Score /SC
End Function
////////////////////////////////////////////////////////
[1]Function Netlingo_lexicon (word)
        search Netlingo_lexicon
        get definition of word      # replace the NetLingo token with relevant English term
        process and assign score
        return score
End Function
////////////////////////////////////////////////////////
[2]Function crosslingual_lexicon(word)
        search crosslingual_lexicon
        get definition of word      # replace the cross_lingual token with relevant English term
        process and assign score
        return score
End Function
```

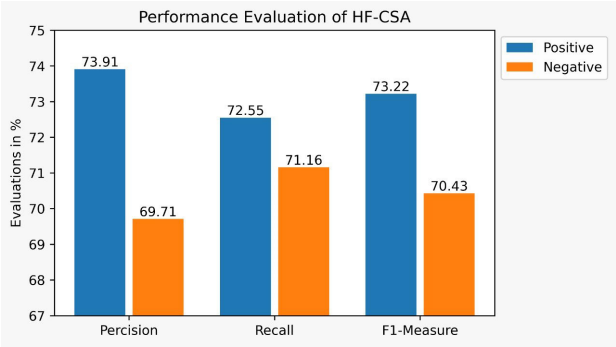

FIGURE 2. Evaluation of HF-CSA on ClIfT_Dataset.

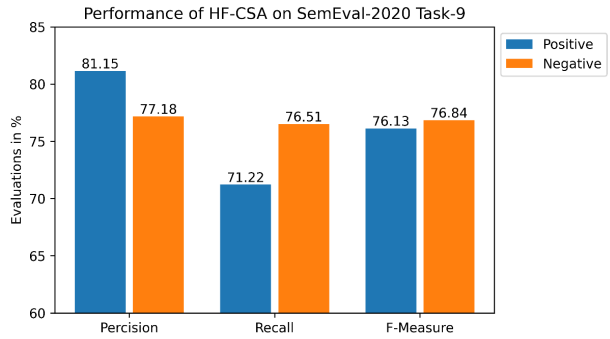

FIGURE 3. Evaluation of HF-CSA on SemEval-2020 Task-9.

E. SENSITIVITY AND SPECIFICITY

In sentiment classification, sensitivity signifies the true positive rate whereas specificity signifies the true negative rate.

$$Sensitivity = \frac{TP}{(TP + FN)} \; and \; Specificity = \frac{TN}{(TN + FP)} \tag{17}$$

The comparative outcomes of HF-CSA are shown below in terms of Precision, Recall, F-Measure, Sensitivity, Specificity and Accuracy.

Figure 2 depicts that HF-CSA attained 73.91% Precision for Positive Instances, 69.71% for Negative and similarly for recall positive it achieved 72.55% and 71.16% for negative instances on ClIfT_dataset whereas F1-Measure for positive instances is achieved as 73.22% and for negative it is 70.43%.

Figure 3 depicts that performance of HF-CSA on SemEval-2020 Task-9 achieved 85.15% Precision for Positive Instances, 77.18% for Negative and similarly for recall positive it attained 71.22% and 76.51% for negative instances on SemEval-2020 dataset, whereas F1-Measure for positive instances is achieved as 76.13% and for negative it is 76.84%.

The system performance is also validated via making a qualitative comparison with machine learning and deep learning based methods on SemEval-2020 dataset. As bidirectional LSTM and XLM-RoBERTRA is utilized in assessing Hindi and Romanized text of SemEval-2020 datasets [3], similarly Wu et al. [4], employed a fine tune BERT (Bidirectional Encoder Representation from Transformers) with multitask learning over SemEval-2020 dataset and three different

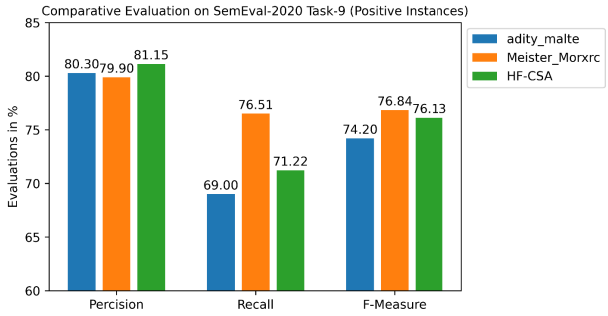

FIGURE 4. Comparative performance of HF-CSA on SemEval-2020 Task-9.

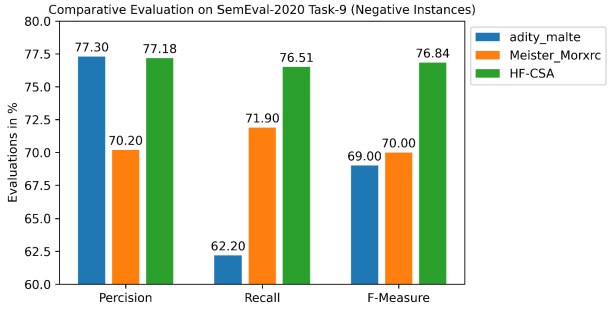

FIGURE 5. Comparative performance of HF-CSA on SemEval-2020 Task-9.

embedding; word position, encoding word and sentence level encoding has been performed.

Figure 4 depicts the comparative performance of HF-CSA on SemEval-2020 Task-9 with ML and deep learning based systems and it is observed that 80.30%, 69.0%, 74.20% precision, recall and f-measure has been observed by adity-malte [3], similarly 79.9%, 76.51% 76.84% precision, recall and f-measure has been observed by Meister_Morxrc [4], Whereas HF-CSA reached to 81.15%, 71.22% and 76.13% of precision, recall and f-measure on positive instances of SemEval-2020 Task-9.

Figure 5 depicts the comparative performance of HF-CSA on negative instances of SemEval-2020 Task-9 with state of art systems and it is noticed that 77.30%, 62.20%, 69.0% precision, recall and f-measure has been observed by adity-malte [3], similarly 70.2%, 71.90% 70.0% precision, recall and f-measure has been observed by Meister_Morxrc [4], Whereas HF-CSA reached to 77.18%, 76.51% and 76.84% of precision, recall and f-measure on negative instances of SemEval-2020 Task-9.

Similarly,Figure 6 presents the confusion matrix of HF-CSA on Clift dataset. HF-CSA attained 71.60% and 76.18% accuracy on ClIfT and SemEval-2020 datasets respectively. Table 9 presents the sensitivity, specificity of HF-CSA on Clift dataset and Table 10 fine-grained interpretability of opinion wise accuracies respectively. It is clearly visible that incorporation of informal and Crosslingual features has improved the accuracy as it raised from 54.92 to 71.60.

As mentioned earlier, Crosslingual sentiment analysis is a learning paradigm which transfers the information of one

TABLE 9. Evaluation of HF-CSA in terms of sensitivity and specificity.

Datasets	Sensitivity	Specificity
Clift	72.5	71.1
SemEval-2020	71.2	76.5

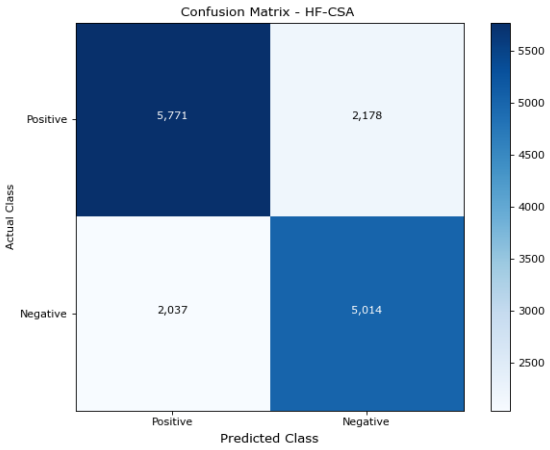

FIGURE 6. Confusion matrix of HF-CSA.

TABLE 10. Attention based improvement in accuracy.

Method	Formal Opinions	Formal & Informal	Formal, Informal & Crosslingual	Formal, Informal, Crosslingual & Emphasizers
Data in %	38%	38 + 21 = 59%	38 + 21 +18= 77%	38 + 21 +18+ 23=100%
HF-CSA	54.92	61.36	67.47	71.60

human language into another in order to help the resource poor language for addressing scarcity of data. Here in this study roman Urdu, informal NetLingua terms are mapped and transformed into resource rich form.

Although there exists a preprocessing model for each BERT encoder but it doesn't handle informality and abstruseness of text, further Romanization of Urdu text is also lacking. Proposed system is capable of handling informality and abstruseness of text.

Figure 3, 4, 5 presents it clearly that the outcomes of HF-CSA are better in comparison with baseline systems over informal, anglicized and Crosslingual contents. Existing Literature of deep learning methods, Urdu transliterations systems [41], [42], [43], [44], [45], [46], [47], [48] and experimental setup revealed that unsupervised lexicon based systems generate satisfactory outcomes for standard, formal, informal as well as multilingual text of resource poor languages.

One another solution to multilingualism is facial recognition and it has been reported that facial recognition based sentiment analysis has gained noticeable attention due to the advanced technology in commercial and industrial applications such as smart Master card for online transaction, health related devices, character recognition, IoT, Post Pandemic World, pain detection, criminal identification and security surveillance [49], [50], [51], [52], [53], [54], [55]. HF-CSA makes it possible to sort and utilize unstructured text of resource poor languages in order to assess customer support issues and to support consumers' satisfaction, reputation management, brand monitoring, decision support systems and market analysis.

V. CONCLUSION

Sentiment Analysis of users' opinion has become a de-facto skillset for many organization and companies. Beside the challenges of detecting formal and resource rich languages it is also noticed that social media publishers are adopting informal, Crosslingual, code-switched, anglicized and emphasized nature of opinion bearing terms. Semantic orientation of Crosslingual text is rising research topic of sentiment analysis but for resource poor languages it is less researched due to resource scarcity and aforementioned linguistic peculiarities.

This study proposes a Heuristic Framework for Crosslingual Sentiment Analysis (HF-CSA) in order to improve the efficacy of sentiment classification for resource poor languages. Crosslingual and informal dataset is compiled for English, Urdu and anglicized Opinionative text to assess the performance of proposed HF-CSA. The contribution of this study is the employment of linguistic processing along with lexicon based method for resource poor language and additional contribution is the compilation of Urdu-English bilingual annotated SA lexical resource for anglicized, informal and Crosslingual contents. Experimental setup determined that incorporation of informal text normalization and linguistic processing of Crosslingual contents is proved as backbone for HF_CSA.

The robustness of HF-CSA is ensured on SemEval-2020 task 9 and results articulate that deep learning based methods are still inferior solution on informal data of resource poor language due to inefficient linguistic processing and limit of small size input window.

The results show that HF-CSA achieved better outcomes in comparison with existing systems on resource poor, informal and Crosslingual text but few systematic deficiencies and limitations are still there;

- Scarcity of large scale coded corpora for informality of Urdu contents.
- Imbalance morphological complexities of source and target language.
- HF-CSA lacks the handling of ironic, sarcastic and aspect based Orientation.
- Urdu parts of speech tagging and dependency parsing have not been fully explored.
- Sophisticated lexical annotation and proficient word sense disambiguation of Urdu is still lacking.

Keeping in view the above mentioned limitations and systematic deficiencies following key point can be treated as future directions.

- Large scale sentiment lexicon and parallel corpora of Urdu needs to be extended for proficient coverage of target language text.
- Best range of source language needs to be explored to minimized imbalance gap between source and target language.

- Provision of ironic, lexicographical and morphological information can improve the efficacy of Crosslingual Sentiment Analysis and it also helps in the orientation of Emotional Intelligence.

ACKNOWLEDGMENT

This work was supported by the FCT—Fundação para a Ciência e a Tecnologia, I.P. [Project UIDB/05105/2020].

AUTHOR CONTRIBUTIONS

Conceptualization, A.A.R., A.H. and J.A.; methodology, A.A.R., and A.H.; Model, A.A.R and A.H.; validation, A.A.R., A.H. and J.A.; formal analysis, A.A.R.; resources, A.A.R, A.H and F.M.; writing—original draft preparation, A.A.R.; writing—review and editing, A.A.R, A.H., B.S.,J.A. and F.M.; visualization, A.A.R.; supervision, A.H., J.A. and F. M.; funding acquisition, B.S and F.M. All authors have read and agreed to the published version of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.
[2] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. Pykl, B. Gambäck, T. Chakraborty, T. Solorio, and A. Das, "SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets," in *Proc. 14th Workshop Semantic Eval.*, 2020, pp. 774–790.
[3] A. Malte, P. Bhavsar, and S. Rathi, "Team_Swift at SemEval-2020 task 9: Tiny data specialists through domain-specific pre-training on code-mixed data," in *Proc. 14th Workshop Semantic Eval.*, 2020, pp. 1310–1315.
[4] Q. Wu, P. Wang, and C. Huang, "MeisterMorxrc at SemEval-2020 task 9: Fine-tune BERT and multitask learning for sentiment analysis of code-mixed tweets," 2020, *arXiv:2101.03028*.
[5] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. 2nd Int. Conf. Knowl. Capture*, Oct. 2003, pp. 70–77.
[6] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," 2002, *arXiv:cs/0212032*.
[7] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proc. 12th Int. Conf. World Wide Web*, 2003, pp. 519–528.
[8] A. A. Raza, A. Habib, J. Ashraf, and M. Javed, "Semantic orientation based decision making framework for big data analysis of sporadic news events," *J. Grid Comput.*, vol. 17, no. 2, pp. 367–383, Jun. 2019.
[9] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, Aug. 2014.
[10] S. Kiritchenko, M. Javed, X. Zhu, and A. Raza, "Socio monitoring framework (SMF): Efficient sentiment analysis through informal and native terms," *Int. J. Adv. Appl. Sci.*, vol. 7, no. 12, pp. 113–126, Dec. 2020.
[11] A. Ali, A. Habib, J. Ashraf, and M. Javed, "A review on Urdu language parsing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 93–97, 2017.
[12] A. Habib and A. A. Raza, "Iot-based pervasive sentiment analysis: A fine-grained text normalization framework for context aware hybrid applications," in *Information and Knowledge in Internet of Things*. Cham, Switzerland: Springer, 2022, pp. 201–226.
[13] X. Fu, W. Shi, X. Yu, Z. Zhao, and D. Roth, "Design challenges in low-resource cross-lingual entity linking," 2020, *arXiv:2005.00692*.
[14] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: State of the art and independent comparison of techniques," *Cogn. Comput.*, vol. 8, no. 4, pp. 757–771, Aug. 2016.
[15] L. Brönnimann, "Multilanguage sentiment-analysis of Twitter data on the example of Swiss politicians," Univ. Appl. Sci. Northwestern Switzerland, Windisch, Switzerland, Tech. Rep., 2013.
[16] M. Abdalla and G. Hirst, "Cross-lingual sentiment analysis without (good) translation," 2017, *arXiv:1707.01626*.
[17] R. Xu, L. Gui, J. Xu, Q. Lu, and K.-F. Wong, "Cross lingual opinion holder extraction based on multi-kernel SVMs and transfer learning," *World Wide Web*, vol. 18, no. 2, pp. 299–316, Mar. 2015.
[18] H. Amiri and T.-S. Chua, "Mining slang and urban opinion words and phrases from cQA services: An optimization approach," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, Feb. 2012, pp. 193–202.
[19] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, decision tree and KNN classification techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016.
[20] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for Roman Urdu text," *Proc. Comput. Sci.*, vol. 147, pp. 131–135, Jan. 2019.
[21] K. Kanclerz, P. Miłkowski, and J. Kocoń, "Cross-lingual deep neural transfer learning in sentiment analysis," *Proc. Comput. Sci.*, vol. 176, pp. 128–137, Jan. 2020.
[22] N. V. Son, "Mining parallel corpora for multilingual machine translation system," Ph.D. dissertation, Int. Res. Inst. MICA Multimedia, Inf., Univ. Oxford, Oxford, U.K., 2005.
[23] M. A. M. Almansor, C. Zhang, W. Khan, A. Hussain, and N. Alhusaini, "Cross lingual sentiment analysis: A clustering-based bee colony instance selection and target-based feature weighting approach," *Sensors*, vol. 20, no. 18, p. 5276, Sep. 2020.
[24] A. Sarker and G. Gonzalez, "DiegoLab16 at SemEval-2016 task 4: Sentiment analysis in Twitter using centroids, clusters, and sentiment lexicons," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 209–214.
[25] R. Amalia, M. A. Bijaksana, and D. Darmantoro, "Negation handling in sentiment classification using rule-based adapted from Indonesian language syntactic for Indonesian text in Twitter," *J. Phys., Conf. Ser.*, vol. 971, Mar. 2018, Art. no. 012039.
[26] I. Mozetič, M. Grčar, and J. Smailović, "Multilingual Twitter sentiment classification: The role of human annotators," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0155036.
[27] A. Al-Shabi, A. Adel, N. Omar, and T. Al-Moslmi, "Cross-lingual sentiment classification from English to Arabic using machine translation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, pp. 1–7, 2017.
[28] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in *Proc. Mex. Int. Conf. Artif. Intell.* Berlin, Germany: Springer, 2010, pp. 32–43.
[29] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 224–227.
[30] M. Mohammadi and N. GhasemAghaee, "Building bilingual parallel corpora based on Wikipedia," in *Proc. 2nd Int. Conf. Comput. Eng. Appl.*, vol. 2, 2010, pp. 264–268.
[31] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. Mach. Transl. Summit X, Papers*, 2005, pp. 79–86.
[32] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *J. Artif. Intell. Res.*, vol. 44, pp. 179–222, May 2012.
[33] K. W. Church and W. A. Gale, "Concordances for parallel text," in *Proc. 7th Annu. Conf. UW Centre New OED Text Res.*, 1991, pp. 40–62.
[34] P. Koehn and C. Monz, "Shared task: Statistical machine translation between European languages," in *Proc. ACL Workshop Building Using Parallel Texts*, 2005, pp. 119–124.
[35] P. Koehn, "Pharaoh: A beam search decoder for phrase-based statistical machine translation models," in *Proc. Conf. Assoc. Mach. Transl. Amer.* Berlin, Germany: Springer, 2004, pp. 115–124.
[36] D. Eberhard, G. Simons, and C. Fennig, *Ethnologue: Languages of the World*, 23rd ed. Dallas, TX, USA: SIL International, 2020. [Online]. Available: http://www.ethnologue.com
[37] C. P. Huynh, "New approach for collecting high quality parallel corpora from multilingual websites," in *Proc. 13th Int. Conf. Inf. Integr. Web-Based Appl. Services*, Dec. 2011, pp. 341–344.
[38] C. Boitet, "Corpus pour la ta: Types, tailles et problèmes associés, selon leur usage et le type de système," *Revue Française Linguistique Appliquée*, vol. 12, pp. 25–38, Jan. 2007.
[39] A. Esuli and F. Sebastiani, "SentiWordNet: A high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, no. 1, p. 26, 2007.
[40] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.
[41] Q. Zhang, Z. Zhang, M. Yang, and L. Zhu, "Exploring coevolution of emotional contagion and behavior for microblog sentiment analysis: A deep learning architecture," *Complexity*, vol. 2021, pp. 1–10, Jan. 2021.

[42] M. Adnan, A. Habib, J. Ashraf, B. Shah, and G. Ali, "Improving M-learners' performance through deep learning techniques by leveraging features weights," *IEEE Access*, vol. 8, pp. 131088–131106, 2020.
[43] A. A. Malik and A. Habib, "Urdu to English machine translation using bilingual evaluation understudy," *Int. J. Comput. Appl.*, vol. 82, no. 7, pp. 5–12, Nov. 2013.
[44] S. Sazzed, "Cross-lingual sentiment classification in low-resource Bengali language," in *Proc. 6th Workshop Noisy User-Generated Text (W-NUT)*, 2020, pp. 50–60.
[45] P. Singh and E. Lefever, "Sentiment analysis for Hinglish code-mixed tweets by means of cross-lingual word embeddings," in *Proc. LREC 4th Workshop Comput. Approaches Code Switching, Eur. Lang. Resour. Assoc. (ELRA)*, 2020, pp. 45–51.
[46] K. Dashtipour, M. Gogate, J. Li, F. Jiang, B. Kong, and A. Hussain, "A hybrid Persian sentiment analysis framework: Integrating dependency grammar based rules and deep neural networks," *Neurocomputing*, vol. 380, pp. 1–10, Mar. 2020.
[47] F. Barbieri, L. E. Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 258–266.
[48] Y. Xu, H. Cao, W. Du, and W. Wang, "A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations," *Data Sci. Eng.*, vol. 7, no. 3, pp. 279–299, Sep. 2022.
[49] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020.
[50] S. E. Morabit, A. Rivenq, M.-E.-N. Zighem, A. Hadid, A. Ouahabi, and A. Taleb-Ahmed, "Automatic pain estimation from facial expressions: A comparative analysis using off-the-shelf CNN architectures," *Electronics*, vol. 10, no. 16, p. 1926, Aug. 2021.
[51] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, "A survey of AI-based facial emotion recognition: Features, ML & DL techniques, age-wise datasets and future directions," *IEEE Access*, vol. 9, pp. 165806–165840, 2021.
[52] A. Darwiesh, M. I. Alghamdi, A. H. El-Baz, and M. Elhoseny, "Social media big data analysis: Towards enhancing competitiveness of firms in a post-pandemic world," *J. Healthcare Eng.*, vol. 2022, pp. 1–14, Mar. 2022.
[53] A. Ghosh, S. Umer, M. K. Khan, R. K. Rout, and B. C. Dhara, "Smart sentiment analysis system for pain detection using cutting edge techniques in a smart healthcare framework," *Cluster Comput.*, vol. 25, pp. 1–17, Jan. 2022.
[54] S. Dhankhar, M. K. Gupta, F. H. Memon, S. Bhatia, P. Dadheech, and A. Mashat, "Support vector machine based handwritten Hindi character recognition and summarization," *Comput. Syst. Sci. Eng.*, vol. 43, no. 1, pp. 397–412, 2022.
[55] S. Bhatia, A. Kumar, and M. M. Khan, "Role of genetic algorithm in optimization of Hindi word sense disambiguation," *IEEE Access*, vol. 10, pp. 75693–75707, 2022.

ARSLAN ALI RAZA is currently pursuing the Ph.D. degree in computer science with the Kohat University of Science and Technology, Kohat, Pakistan. He is also a Lecturer of computer science with COMSATS University Islamabad, Vehari Campus, Vehari, Pakistan. His research interests include natural language engineering, data mining, machine learning, and artificial intelligence.

ASAD HABIB received the Ph.D. degree from the Nara Institute of Science and Technology, Nara, Japan. He is currently an Assistant Professor with the Institute of Information Technology, Kohat University of Science and Technology. His research interests include natural language processing, human–computer interaction, artificial intelligence, learning technologies, mobile learning, and adaptive interface design.

JAWAD ASHRAF received the Ph.D. degree from the Department of Computer Science, University of Leicester, U.K. He is currently with the Institute of Information Technology, Kohat University of Science and Technology, where he is working on partner-based scheduling algorithm for grid workflows in advance reservation environment, K-shortest path variant for routing in advance reservation environment, and novel workflow job selection technique.

BABAR SHAH is currently an Associate Professor with the College of Technological Innovation, Zayed University, Dubai, United Arab Emirates. His professional services include but are not limited to guest editorships, university services, the workshops chair, a technical program committee member, and a reviewer of several reputed international journals and conferences. His research interests include WSN, WBAN, the IoT, churn prediction, security, real-time communication mobile P2P networks, and M-learning.

FERNANDO MOREIRA received the degree in computer science, in 1992, the M.Sc. and Ph.D. degrees in electronic engineering from the Faculty of Engineering, University of Porto, in 1997 and 2003, respectively, and the Habilitation degree, in 2018. He has been a member of the Department of Science and Technology, Portucalense University, Portugal, since 1992. He was the Head of the Department of Science and Technology, from May 2018 to February 2022. He is currently a Full Professor with Portucalense University. He is also a Full Professor and a Visiting Professor with the University of Porto Business School. He teaches subjects related to undergraduate and postgraduate studies. He was the Computation Co-ordinator of the M.Sc. in computation during the last ten years. He supervises several Ph.D. and M.Sc. students. He organized several special issues from JCR journals. He is the coauthor of more than 200 scientific publications with peer review on national and international journals and conferences. His research interests include mobile computing, ICT in higher education, mobile learning, social business, and digital transformation. He serves as a member of the editorial advisory board for several journals and books. He has already regularly served as a member of program and scientific committees for national and international conferences. He is associated with NSTICC, ACM, and IEEE. He was awarded the Atlas Elsevier Award, in April 2019. He holds editorial experience and he is the co-editor of several books.