Normalization of Unstructured and Informal Text in Sentiment Analysis

Muhammad Javed[1], Shahid Kamal[2]
Institute of Computing and Information Technology, Gomal University,
Dera Ismail Khan, K.P.K, Pakistan.

***Abstract*—Sentiment Analysis is problem of natural language processing which deals with the extraction and analysis of public sentiments shared about target entities over microbloging websites. This field has gained great attention due to the huge availability of decision making textual contents. Sentiment Analysis has enormous application areas such as; Market Analysis, Service Analysis, Showbiz analysis, Movies, sports and even the popularity and acceptance rate of political policies can also be predicted via sentiment analysis systems. Although tremendous volume of opinionative text is available but it is unstructured and noisy due to which sentiment classifiers can't achieve good outcomes. Normalization is the process used to clean noise from unstructured text for sentiment analysis. In this study we have proposed a mechanism for the normalization of informal and unstructured text. Proposed mechanism is comprised of four essential phases; Noise Reduction, Part of Speech Tagging, Stop Word Removal stemming and Lemmatization. Numerous experiments are performed on twitter data set with unsupervised lexicons and dictionaries. Python and Natural language toolkit is used for performing all four essential steps. This study demonstrates that utilization and normalization of informal tokens in tweets improved the overall classification accuracy from 75.42 to 82.357.**

Keywords*—*Informal; normalization; opinion mining; roman; sentiment analysis; text preprocessing

I. INTRODUCTION

Text Mining is computer assisted process introduced to help business organizations by providing effective decision making answers and future trends. Text Mining is the method of mining high quality information from text using patterns with additional knowledge of linguistic rules. Text Mining fulfils the needs of government, research and business e.g. E-discovery, Scientific Discovery and National Security [1]. The extraction and recognition of pattern is performed through the intersection of Machine Learning, Artificial Intelligence and Database system [2]. The Social sites are used widely for socio communication like Blogs and Microblogs. Blogs are utilized for publishing articles, news, or any other topic that is of interest. Mostly organization like Exact, Themovieblog, ESOMAR (European Society for Opinion and Market Research), AAPOR (American Association for Public Opinion Research) and individuals have their own blogs for communication. Blogging sites provide immediate feedback of reviewers about their products, articles and publications. On the other hand, the sites that allow short text for chatting, communication, exchanging views about their interests are considered as Microblogs. These sites allow the posting of short text and messages. Microblogs are designed for expressing real world actions in an instant environment. The common characteristics of microbloging sites are: (i) Short Text (ii) Instant Messaging (iii) Pictorial symbols (iv) Slang terms (v) Real time [3]. The well-known microbloging sites are Tumblr, Plurk, Friendfeed and Twitter. Twitter is the most popular microbloging site that allows its users to publish short messages (tweets) for communication. The emergence of social media sites has changed the public communication style so the research directions are shifted from information retrieval to "Opinion Mining". Online users share bulk of opinionative information over these social networking websites so observers and analysts are taking advantage of these available information by collecting and summarizing concerned opinionative information for the sake of monitoring authors' moods about their launched products, services and even political policies for better decision making. Socio Monitoring is performed by means of Sentiment Analysis and Opinion Mining. Sentiment Analysis or Opinion Mining is the novel field of text classification and problem of Natural Language Processing. Opinion Mining is the computational study of public sentiments, feelings and opinions shared in the form of text over social media sites. Extracting public opinions from user generated content is not a big matter instead the identification, summarization and strength of opinions about desired entity is the challenging task. Efficient classification of opinions requires knowledge of machine learning and classification algorithm with appropriate linguistic rules. The rapid growth of socio communication devices and channels produced newer challenges for observers and analysts. Online users publish their views and opinions in distinctive and informal way which is not directly translatable for machine learning system. Additionally they adopt acronyms, emotion icons and other microbloging features for communication. Sentiment Analysis task can't be performed directly on these published reviews instead it requires massive effort of input text preparation. In past numerous experiments have been performed for text normalization and preprocessing. Text Normalization is task of data mining in which text is cleaned from undesired tags and symbols. Normalization (a.k.a. preprocessing) is process of cleaning user generated text for analysis and prediction [4]. One can't extract actual opinion without assessing opinionative text precisely so quality of decision directly depends on the quality of text. In past preprocessing is performed via many different supervised [5], semi supervised [6] and unsupervised [7] methods. Sentiment Analysis is applicable in almost every field of life. In this research we have decided to normalize user generated contents in political domain for making valuable dataset for

the sake of analysis. We have offered a mechanism in which text is cleaned using four key steps; Noise Reduction, Part of Speech Tagging, Stop Word Removal, Stemming & Lemmatization. Python Natural language toolkit is used for performing all four necessary steps. This study demonstrates that utilization and normalization of informal tokens in tweet can improve the overall classification accuracy. The rest of article is comprised of; Section 2 presents related work, section 3 method, section 4 results and discussion and section 5 presents Conclusion and Future work.

II. Related Work

The increasing growth of electronic document on World Wide Web has changed the way of analysis dramatically. The social media is growing rapidly due to the availability of millions of online user generated opinionative contents on social sites. The expressed views and suggestions are considered as sentiments and opinions. These sentiments are mined for better decision making and also for the purpose of analysis and evaluation. Sentiment Analysis or Opinion Mining is the computational study of public moods. Dave el al [8] in 2003 used the term "Opinion Mining" for the first time. Opinion mining or sentiment analysis is the problem of NLP. Sentiment analysis on twitter is new and challenging area, reasonable efforts have already been done in this area but due to increasing ratio of online users this research area is rising day by day for analyzing various entities but the quality of text is big issue for observers and analysts.

Sentiment Analysis for politics is the hot topic, in past a lot of work has been done on predicting elections or political events. In fact Microbloging sites are becoming the most popular platform for political arena [9]. The use of internet was limited to exchange of information with each other before the election of USA in 2008 but it was changed dramatically when Barack Obama started his campaign on social media [10]. It was the first political campaign ran on social media. The social media became one of the most valuable source for political conversation after that campaign. Kim,D [11] investigated that Twitter was highly focused for seeking political information during the Korean Election 2010. Gaffney [12] tracked the #IranElection hashtag for studying the use of microbloging sites more specifically twitter during 2009 Iran election, due to maximum usage of twitter the maintenance of twitter was stopped at one stage on the order of US state department [13]. Political incumbents and challengers used twitter for political benefits during the US midterm elections held in 2010 [14]. Although there exist many systems for the extraction of public sentiment but informal nature of text is big hurdle for all of them. Cleaning or normalizing opinionative text is challenging issue of sentiment analysis. In last few years data cleaning and preprocessing is viewed as an important action and topic of research, as various supervised and unsupervised methods have been experimented for number of domains for analysis purpose. Hariharakrishnan, J. et al [15] reviewed numerous techniques of text preprocessing and highlighted the significance and multi-aspects of preprocessing such as Noise reduction, outlier identification, and inconsistent data. They raised a very logical point that most of the experiments for text preprocessing are performed either in data collection phase or for homogeneous data. There is lack of efforts for heterogeneous text preprocessing and also there is no such system which detect and clean data during classification. They are planning to develop a hybrid system for cleaning homogenous data in different situations. Haddi, E et al [16] explored the significance of text preprocessing for extracting public opinions from social media contents. They performed various experiments over movie reviews dataset with supervised algorithm and concluded that SVM significantly achieved better accuracy in comparison with other algorithms. They used three different features like Feature Presence (FP), Feature Frequency (FF) and Term Frequency Inverse Document Frequency (TF-IDF) and achieved 93 % of overall accuracy. They stated that sentiment analysis is harder problem and one can't achieve promising outcomes without cleaning text. Singh, T. et al [17] proposed a system for efficient preprocessing of text for twitter sentiment analysis. They actually explored the importance of slang words in sentiment analysis by combining these with existing features. Various experiments are performed in which SVM was used as base classifier. Their results demonstrate that proposed system achieved promising outcomes with the combination of conditional random field with n-grams. They achieved 94 % of average accuracy after normalization of text. Hemalatha, I. et al [18] offered a three step preprocessing strategy in which they removed URL as first step, Special and repeated characters are removed in second step while third step was introduced to remove question words. They claimed that with this preprocessing algorithm one can easily perform sentiment analysis with any machine learning algorithm. Angiani, G. et al [19] compared various existing machine learning methods for text preprocessing and stated that appropriate preprocessing can improve and gain the valuable information knowledge from available text. They evaluated the performance of numerous preprocessing strategies over twitter data and concluded that using a dictionary is not a useful idea for upgrading the classification performance. Additionally, they suggested that combining different preprocessing filters can positively improve the classification accuracy. Although sentiments and opinions are mined and analyzed in English like languages but it is observed that opinionative contents in other languages are also available at high rate over social networking websites. Duwairi, R. & El-Orfali,M [20] presented their work for Arabic text in which they investigated Arabic text from three perspective; first one is the multi-aspect of text representation such as significant role of stemming n-gram and features correlation for Arabic language. Secondly the performance of three existing machine learning classifiers; NB, SVM and K-NN was examined and the last perspective was to analyses the impact of different characteristics of dataset over sentiment analysis whereas experiments were performed on two datasets; Manually compiled dataset for politics and existing corpus of movie reviews. Their results demonstrate that Naïve Bayes classifier outperforms the other two on both domains (Politics and Movies) by achieving 96.6% and 85.7% accuracy respectively. They stated that preprocessing and behavior of dataset has great impact on sentiment analysis accuracy. Dos Santos & Ladeira, M [21] performed experiments on Portuguese reviews for presenting the role of text preprocessing in sentiment analysis.

Additionally this research presented a large corpus of 759 thousands reviews as their contribution. They concluded that text preprocessing has insignificant role in text classification and sentiment analysis. They stated that accuracy and performance of sentiment analysis systems depends on the nature of datasets and sentences/reviews used in the dataset because sometimes preprocessing lower the accuracy by removing the valuable and necessary information from target text. Toman, M [22] proposed a lemmatization system which uses multilingual semantic thesaurus Eurowordnet. They evaluated the performance of proposed system on two different corpora. Their findings suggest that the proposed system achieved promising outcomes and they concluded that conversion of inflected forms into their roots does not affect the classification accuracy while on the other hand Christopher, D.M et al. [23] stated that stemming lowers the precision. Dařena, F., & Žižka, J. [24] reviewed the existing preprocessing methods and application for non-standard short text and unfold several informative patterns. They stated that smaller datasets are inappropriate and produces inefficient outcomes. Additionally, this study explored that preprocessing results highly based on the language of data and algorithm. The positive point about preprocessing is that it reduces dictionary size of data collection. Noise and unclear data is not the issue of English language but all languages used over internet based social media require preprocessing of text. Infact preprocessing experiments are performed for almost all human languages; Arabic, French, Hindi, Chinese, Japanese and Turkish. Hidayatullah, A. F et al [25] experimented with Indonesian language in order to clean the text more specifically tweets for further analysis. They divided these experiments into two parts; common preprocessing and specific preprocessing. Their results demonstrate that they achieve good results with specific text preprocessing tasks. Additionally, they suggested preprocessing process can be improved by introducing novel algorithms and system for automatic recognition of non-standard words. The rapid advent in web came with newer communication indicators such as # tags, @ tags and emotion icons. Ignoring such symbols during preprocessing can affect the quality of dataset. One can't directly remove meaningful punctuations during preprocessing, because meaningful punctuation (Emoticons) convey opinion towards target entities. Wegrzyn-Wolska, K et al [26] compared three emoticon's preprocessing methods; Emotion deletion (emodel), emoticons 2-values translation (emo2label) and emoticon explanation (emo2explanation). Emoticon weight lexicon was used with Naïve Bayes Classier in order to assess the effect of emotion icons. They concluded that emotion icons act as verbal indicator of sentiment and can enhance the sentiment analysis accuracy. They achieved 78% of average accuracy with Naïve Bayes Classifier. Gull, R et al [27] proposed an approach for the analysis of qualitative and quantitative data in specified time. They transformed the extracted text into structured format and prepared politics dataset for the analysis of political party using linguistic features and classifiers. Two classifiers Naïve Bayes and SVM are employed and they concluded that SVM produced better outcomes. In future, they are planning to analyze multilingual text. Stemming is one the key phase of preprocessing because identifying and converting inflected forms of opinionative terms may increase quality of dataset. Arjun et al. [28] compared two stemming techniques; Porter's and Krovetz algorithm. Their findings suggested that both algorithms have few limitations in some specific scenario Porter's algorithm [29] is context based and also it leads to large degree of conversion whereas Krovetz algorithm [30] produced inefficient results with large datasets. In past, preprocessing is performed in a sequential manner using a pipeline of preliminary tasks but there exist few systems which utilizes unified phase for all essential tasks. Clark, A [31] came with the design and implementation tool for the preprocessing of noisy corpora. He coped with typographical errors, white space issue using trainable stochastic transducer model over 100 million word corpus of Usenet news. He stated that preprocessing process can be improved by merging various models for sort of typographical errors. Bao, Y et al [32] explored the significance of text preprocessing for twitter sentiment Analysis. They unfolded the impact of URLs, Negation, repeated letters, stemming and Lemmatization. The experiments were performed on Stanford twitter dataset. Their findings show that handling URLs, negations and repeated characters can improve accuracy while on the other hand stemming and lemmatization decreases the classification accuracy. They achieved an average 85.5 % accuracy with original feature space. Petz, G et al. [33] compared various existing preprocessing techniques for sentiment analysis in real world situations. They stated that to achieve satisfactory outcomes three tasks of preprocessing are essentials; sentence tokenization, replacement of slang symbols and stemming of inflected forms. They achieved 83.42 % of average F1-score for eight different techniques. Krouska, A et al. [34] reviewed the recent research of text preprocessing and sentiment analysis. They revealed the in-depth analysis of preprocessing techniques by performing various experiments over manually compiled twitter dataset and stated that appropriate feature selection and proper representation can improve classification accuracy positively. They compared four key classifiers NB, SVM, KNN and C4.5 over three different datasets OMD, HCR and STS-Gold. Raza, A et al [35] reported that modern linguistic style has number of variant features such as use of romans, slangs, Urdu language terms and sentences for expressing their likes, dislikes about hundreds of real world entities. Additionally, system of one domain and language is inapplicable over other languages and domains. Therefore, Text normalization is important to cope with many tasks such as Plagiarism Detection, pattern discovery, Sentence Recognition, Topic Modelling and information retrieval. Liu, B. and Zhang, L [36] demonstrated that sentiment analysis is performed at three granularity levels; document, sentence and phrase level. Previous research showed that document level sentiment analysis has gained much focus as in past [37, 38] performed document level sentiment analysis using minimum cuts algorithm. Kian, K.D et al [39] performed preprocessing experiments to determine the qualitative difference among high and low agreement data. Raza, Raza, A et al [40] used preprocessing and lexicon based sentiment analysis system to capture public opinion shared about political protest over twitter. They stated that effective preprocessing can enhance the classification accuracy. Yu et al [41] proposed a system for sentence level subjectivity classification. In this research

we have proposed a mechanism for text preprocessing at sentence level for sentiment analysis of political contents using existing and manually built dictionaries and lexicons.

III. METHODOLOGY

Sentiment Analysis is process of acquiring users sentiments shared on social networking websites. There exists two main methods of assigning polarities to public sentiments; Supervised & Unsupervised. Whatever the method is used for sentiment analysis it always needs quality text in decision making process.

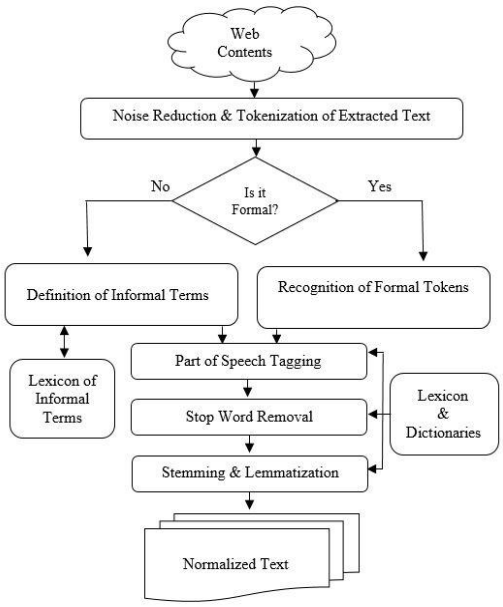

Fig. 1. Mechanism for Normalization of Informal Text.

Today socio sites produce numerous challenges for gathering quality text. In this research a mechanism is proposed for the normalization of user generated opinionative contents in order to perform optimized analysis. Proposed mechanism is comprised of following essential phases as depicted above in Fig. 1.

A. Normalization

Normalization in sentiment analysis is referred as the process of cleaning or removing irrelevant data from a huge collection of extracted data. The extracted data is full of noise containing URLs, tags, links etc. Data preprocessing is performed to remove such noise from extracted text to make it more clear and consistent. In every text mining process data must be preprocessed before going to analysis phase, so we preprocessed the extracted data for further processing. The URLs and tags are removed from extracted data; generally these URLs have no use in sentiment analysis process. Following tasks are involved in text preprocessing process:

B. Noise Reduction

The text extracted from social media sites is full of noise. This text contains URLs, Symbols, undesired punctuations and some special communication symbols i.e. @, RT and < > etc. These symbols and tags have no role in sentiment classification tasks so all such kind of punctuations must be eliminated before mining and analysis. In this phase of preprocessing we removed these undesired symbols and tags using HTML parser.

C. Definition of Informal Tokens

In past extracted text is preprocessed using English repositories but it is observed that today social sites have provided bulk of opinionative data in informal style and words of other language too, so in order to capture opinion from numerous geographical areas there is a dire need to collect and summarize opinion of different styles (formal & Informal) and format. Twitter and other microbloging services allow users to share short informal slangy terms which are not easily detectable for machine and sometimes even for a human reader. So in this study we have captured non-English opinionative words used in English sentences for the sake of efficient sentiment analysis. We first detected all the slang terms and then proper definition is assigned to each extracted token using manually compiled list of slangs and non-standard terms. Slangs refer to misspell English language opinionative terms whereas non-standard used to represent Roman Urdu terms shared in English sentences for expressing positivity and negativity about concerned entity. A list of Roman Urdu opinionative terms is created for effective identification of both slang and non-standard terms. Python Natural Language Toolkit is used for cleaning formal and informal opinionative tokens.

D. Part of Speech Tagging

The noise free text is passed to part of speech tagging phase for labeling appropriate parts of speech tags to each target token. Part of Speech (POS) Tagging is the process of assigning parts of speech to each desired term. In this research tokenized text is labeled according to grammatical nature i.e. adjective, verb, adverb and noun etc. we use python NLTK for assigning part of speech tags to extracted text.

E. Stop Word Removal

The words having high frequency or most frequently used terms are considered as stop words like “for” ”the” “a” etc. In sentiment Analysis, stop words are removed to obtain more concise and desired text for analysis. So we removed all such stop words from extracted data by providing tokens to python NLTK. Python NLTK [42] is a collection of built-in libraries and software for Natural Language Processing. A corpus having list of words of various language is the part of python NLTK so stop words from extracted text is removed through the utilization of corpus with Python NLTK.

F. Stemming and Lemmatization

The process of reducing all the terms with the same stem to a common form is named as stemming; a stem is a root form. For example the stem for the words “fishing”, “fished”, “fisher” is “Fish” while lemmatization is the process of removing inflectional endings and replacing this inflected word with a base word, and the module used for this process is known as Lemmatizer; the Lemmatizer uses an additional dictionary to replace the inflected forms into its base form. As in stemming the terms “begging” “beggar” “beginning” will be replaced with the terms “beg” while in lemmatization these terms will be replaced with terms “beg” , “beg” and “begin”

respectively. The output generated by Lemmatizer is more accurate as compared to that of stemmer. So we replaced all the inflected form to their base form by using python NLTK Lemmatizer. The Python NLTK Lemmatizer uses WordNet database for finding lemmas of inflected terms. The canonical form of the word is known as lemma.

The preprocessed text is saved in separate file as dataset for classification and analysis. We have evaluated the effectiveness of our preprocessed text using existing classification technique.

IV. Results and Discussion

To evaluate the effectiveness of proposed mechanism comprehensive experiments are performed on twitter data about Pakistan Political Parties and Leaders. The data is extracted about Pakistan politics from publically available reviews of twitter using twitter APIs. Manual annotation is performed to assign polar classes to each extracted tweet so a set of 1400 tweets in which 700 positive whereas other 700 negative are labelled as benchmark in order to evaluate the performance of preprocessing mechanism. Table I. shows the statistics for positive & negative tweets for both formal & informal opinions.

All the necessary steps of normalization mentioned in section 3 are performed using Python natural language toolkit. This study presents a novel mechanism of text normalization in the classification of informal opinion bearing text. In past there exist many methods for text normalization but still there is sufficient gap for improvement so we proposed a mechanism which detect all the opinionative feature in preprocessing phase. Table.II presents the opinionative features which are considered in this study just to increase the classification accuracy.

TABLE I. Human Annotated Dataset of Formal and Informal Opinionative Tweets

MANUALLY LABELED OPINION	**POSITIVE**	**NEGATIVE**
FORMAL OPINION	500	500
INFORMAL OPINION	200	200

TABLE II. Formal and Informal Opinionative Features

S.NO.	**OPINION FEATURES**	**NATURE**	**DEFINITION**
1	**f1:** Adjective	Formal	Qualifying Word: A word that shows the quality of an entity
2	**f2:** Verb	Formal	Action: A word that shows some action on an entity
3	**f3:** Adverb	Formal	An adverb is a word that emphasis an adjective, verb.
4	**f4:** Slangs & Acronyms	Informal	Informal opinionative words that are left misspelled intentionally in some particular context.
5	**f5:** Roman Urdu Terms	Informal	Writing Urdu script with English letter according to its appropriate pronunciation.
6	**f6:** Emoticon	Symbolic	Punctuations and combination of characters used to show facial expressions.

This section presents the experimental findings of proposed mechanism. In order to underline the impact of preprocessing we have presented precision, recall, f-measure & accuracy of tweets collection for both formal and informal opinion bearing terms. We have computed precision, recall f-measure separately for both terms just to emphasize the effectiveness of handling informal opinions in sentiment analysis.

A. Precision

Precision is actually the fraction between retrieved and relevant instances as shown below in equation 1.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

Whereas TP is used for True Positive, FP is for False Positive. TN shows True negative and FN is for False Negative. In this study we used these terms for specifying the following criteria, TP: Correctly identified as positive by the proposed framework, FP: Incorrectly identified as positive. Similarly, for negative, TN is for correctly identified as negative while FN shows the terms which are incorrectly identified as negative.

Precision for **formal** positive instances:

$$\frac{TP}{TP+FP} = \frac{396}{486} = 81.31\%$$

Precision for **formal** negative instances:

$$\frac{TN}{TN+FN} = \frac{409}{513} = 79.72\%$$

The precision for informal positive and negative tweets is as follow;

Precision for **informal** positive instances:

$$\frac{TP}{TP+FP} = \frac{186}{224} = 83.03\%$$

Precision for **informal** negative instances:

$$\frac{TN}{TN+FN} = \frac{162}{176} = 92.04\%$$

Similarly, precision for both formal & informal positive and negative tweets is described as below;

Precision for **both** formal & informal positive instances:

$$\frac{TP}{TP+FP} = \frac{582}{711} = 81.85\%$$

Precision for **both** formal & informal negative instances:

$$\frac{TN}{TN+FN} = \frac{571}{689} = 82.87\%$$

B. Recall

Recall is used to find the numbers of relevant from retrieved instances as shown below in eq.2.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

Recall for formal positive & negative tweets is presented as follow;

Recall for **formal** positive instances:

$\frac{TP}{TP+FN} = \frac{396}{500} = 79.2\%$

Recall for **formal** negative instances:

$\frac{TN}{TN+FP} = \frac{409}{500} = 81.8\%$

Recall for informal positive & negative tweets is shown below;

Recall for **informal** positive instances:

$\frac{TP}{TP+FN} = \frac{186}{200} = 93\%$

Recall for **informal** negative instances:

$\frac{TN}{TN+FN} = \frac{162}{200} = 81\%$

Similarly, recall for both formal & informal positive & negative tweets is shown below;

Recall for **both** formal & informal positive instances:

$\frac{TP}{TP+FP} = \frac{582}{700} = 83.14\%$

Recall for **both** formal & informal negative instances:

$\frac{TN}{TN+FN} = \frac{571}{700} = 81.57\%$

C. F-Measure

F-Measure is also a statistical analysis used in binary classification. It is used to measure the accuracy by considering both precision and recall as shown below in eq.3.

$$\text{F} - \text{Measure} = \frac{2(Precision * Recall)}{Precision + Recall} \quad (3)$$

$$\text{F} - \text{Measure (Positive)} = \frac{2(81.85 * 83.14)}{81.85 + 83.14} = 82.489\%$$

$$\text{F} - \text{Measure (Negative)} = \frac{2(82.87 * 81.57)}{82.87 + 81.57} = \frac{2(6759.70)}{164.44} = 82.21\%$$

D. Accuracy

The accuracy is the degree of correctness; Mathematical representation of accuracy is shown below in eq. 4.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Accuracy} = \frac{582+571}{582+571+129+118} = 82.357\%$$

It is observed that out of 1153 opinionative tweets 1054 tweets are identified as opinionative with formal and informal opinion bearing words whereas rest of the opinionative tweets are identified using informal tokens only, where no single formal opinionative token was present which increases accuracy up to 6.937 from 75.42 to 82.357. Table III shows a subset of tweets collection which are marked as opinionative with formal and informal opinion.

TABLE III. OPINIONATIVE TWEETS HAVING BOTH FORMAL & INFORMAL FEATURES

S. No	OPINIONATIVE TWEETS	OPINION FEATURES	LABEL
2	Riaz has bribed many politicians but he must know he can never bribe me: PTI chief	F2	Positive
3	Once a drbari always darbari	F5	Negative
4	Patwari (zehni ghulam), darbari and bhikari All are in shock That what happened to us.	F2, F5	Negative
5	Chal patwari get lost...	F4, F5	Negative
6	Aala to good zbrdst	F1, F5	Positive
7	Fucking Daghi :-P	F2, F5	Negative
8	That's great janbaz 👍 #Bilawal	F1, F5, F6	Positive
9	Wah,,bahut aala,,,pti linked offshore companies are neat and clean like imran niazi,,,, hahahaha	F1, F4, F5	Positive
10	I have seen KPK hospitals , they are better than Punjab hospitals, 1000 times better and i am a doctor also i know better than you Mr brainless Patwari	F1, F5	Positive

E. Confusion Matrix

Confusion Matrix or Contingency Table is used for the evaluation of proposed system. Confusion Matrix is actually a two dimensional array which is used to visualize the performance of proposed mechanism. Table IV shows the confusion matrix of experimental results. In which rows show the number of manually annotated instances whereas columns show the machine/system annotated instances.

TABLE IV. CONFUSION MATRIX OF FORMAL AND INFORMAL OPINIONATIVE TERMS

POLARITY CLASS LABELS FOR FORMAL & INFORMAL OPINIONATIVE TERMS			MACHINE ANNOTATED LABELS ↓ ↓ Positive	Negative	TOTAL
Formal opinion	**HUMAN ANNOTATED**	Positive	396 (TP)	104 (FN)	500
		Negative	91 (FP)	409 (TN)	500
		TOTAL	487	513	1000
Informal opinion		**CLASS LABELS**	**Positive**	**Negative**	**TOTAL**
		Positive	186 (TP)	14 (FN)	200
		Negative	38 (FP)	162 (TN)	200
		TOTAL	224	176	400
Formal & informal opinions		**CLASS LABELS**	**Positive**	**Negative**	**TOTAL**
		Positive	582 (TP)	118 (FN)	700
		Negative	129 (FP)	571 (TN)	700
		TOTAL	711	689	1400

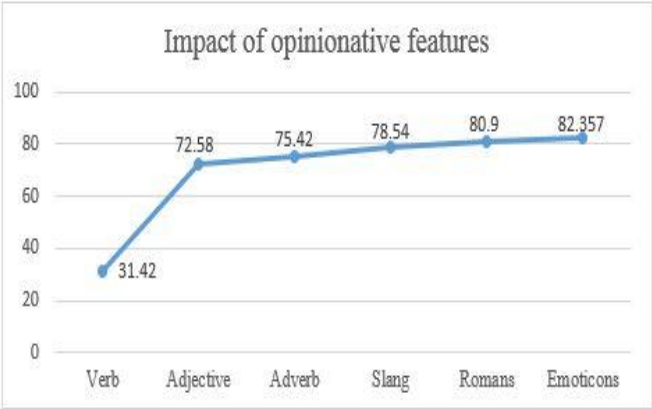

Fig. 2. Impact of Opinionative Features.

In above table, vertical representations of positive and negative instances show the outcomes of machine whereas horizontal instances indicate manually labeled instances. We have decomposed whole data set into formal and informal so here in Table.4 ternary confusion matrix is shown in order to provide clear picture of informal opinions. Third &last confusion matrix shows the overall results of both formal & informal opinions in which machine labeled 711 opinionative tweets as positive while 689 are marked as negative with accuracy of 81.9 and 82.85 respectively. As in this preprocessing mechanism six opinionative features are considered as shown in table.2. Experimental results demonstrate that all these features have great impact on real world results. Fig.2 shows that few of them has increased the classification accuracy dramatically.

Fig.2 shows that all considered features have increased the average accuracy of sentiment analysis. If we consider only verbs from whole collection the accuracy was noticed as 31.42, and for verb and adjectives, accuracy is jumped to 72.58 and similarly for all features the overall accuracy is achieved as 82.357 which shows significant contribution in sentiment analysis.

Table.V shows the comparative results of proposed preprocessing mechanism, it is noticed that proposed system outperformed the existing systems by achieving an average precision, recall and accuracy of 81.9%, 82.35%, and 82.357% respectively.

TABLE V. COMPARATIVE ANALYSIS OF PROPOSED PREPROCESSING MECHANISM

STUDIES	PRECISION	RECALL	ACCURACY
Pang, B et al. [37]	**83%**	80.58%	81.5%
Haddi, E [43]	63%	60%	60% lexical
Etaiwi, W et al. [44]	51.8%	74.9%	72.96%
Proposed	81.9%	**82.35%**	**82.357%**

V. CONCLUSION AND FUTURE WORK

Sentiment Analysis is computational study of user's opinion about real world entities. Analysis are performed on publically available data over social media sites. Machine Learning algorithms need a well formed quality dataset for analysis so publically available text is first normalized in order to achieve decision making results. One can't mine public opinions accurately without inputting meaningful instances. In fact, quality of analysis directly depends on the size and nature of input data. This research proposes a novel mechanism for normalization of publically available opinionative data for the sake of sentiment analysis. Text normalization is not just a single step, Infact it is the process of performing a flow of essential actions sequentially i.e. Tokenization, Stop word removal, Part of Speech Tagging, Stemming and Lemmatization. In this study we have considered six opinion bearing indicators; Verb, Adjective, Adverb, Slang, Roman Urdu terms and Emoticon as classification attributes. In past, non-standard and unstructured terms are handled at classification phase which sometimes lowers the classification accuracy so in order to overcome this deficiency proposed study provides a proper definition to each extracted informal & non-standard terms at normalization phase. Twitter data is first crawled using twitter APIs and then separate file is generated for performing normalization tasks. The normalization steps namely; Noise reduction, informal definition, parts of speech tagging, stemming and lemmatization are performed in incremental manners. To evaluate the results of proposed mechanism experiments are performed on manually annotated collection of 1400 tweets equally distributed for positive and negative opinion bearing instances. Experimental results demonstrate that informal opinions have great impact on the classification accuracy as we achieved 82.357% accuracy with an increment of 6.937%. Proposed mechanism is robust and can be applied at multidimensional domains. We must encourage future researchers to experiment with novel opinionative features to provide quality datasets for many real world entities.

REFERENCES

[1] Glass, K. and Colbaugh, R., 2012. Estimating the sentiment of social media content for security informatics applications. Security Informatics (a springer open jouranl), Vol. 1, Issue 1.pp 1-16

[2] Lahoti, A.A., 2014. Data Mining Technique its Needs and Using Applications. , IJCSMC Vol. 3.Issue. 4, pp.572-579.

[3] Kumar, A. & Sebastian, T.M., 2012. Sentiment Analysis: A Perspective on its Past, Present and Future.International Journal of Intelligent Systems and Applications, Vol. 4. Issue.10. pp1-14.

[4] Jianqiang, Z. and Xiaolin, G., 2017. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access, 5, pp.2870-2879.

[5] Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), pp.111-117.

[6] Xiang, B. and Zhou, L., 2014. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 434-439).

[7] Karl, M., Bayer, J. and van der Smagt, P., 2016. Unsupervised preprocessing for Tactile Data.

[8] Dave, K. et al., 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th

international conference on World Wide Web, pp.519–528. New York, NY, USA.

[9] Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M., 2011. Election forecasts with Twitter: How 140 characters reflect the political landscape. Social science computer review, 29(4), pp.402-418.

[10] Anderson, D., 2009. How has Web 2.0 reshaped the presidential campaign in the United States? In Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.

[11] Kim, D., 2011. Tweeting politics: Examining the motivations for Twitter use and the impact on political participation. In 61st Annual Conference of the International Communication Association.

[12] Gaffney, D., 2010. iran Election : Quantifying Online Activism. Analysis, (pp.1-8). Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

[13] Pleming, S., 2009. US State Department speaks to Twitter over Iran. Reuters, June, 16.

[14] Cozma, R. and Chen, K., 2011, May. Congressional Candidates' Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?. In 61st Annual Conference of the International communication association.

[15] Hariharakrishnan, J., Mohanavalli, S. and Kumar, K.S., 2017, January. Survey of pre-processing techniques for mining big data. In Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on (pp. 1-5). IEEE.

[16] Haddi, E., Liu, X. and Shi, Y., 2013. The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17, pp.26-32.

[17] Singh, T. and Kumari, M., 2016. Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89, pp.549-554.

[18] Hemalatha, I., Varma, G.S. and Govardhan, A., 2012. Preprocessing the informal text for efficient sentiment analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1(2), pp.58-61.

[19] Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F. and Manicardi, S., 2016. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb.

[20] Duwairi, R. and El-Orfali, M., 2014. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, 40(4), pp.501-513.

[21] Dos Santos, F.L. and Ladeira, M., 2014, October. The role of text pre-processing in opinion mining on a social media language dataset. In Intelligent Systems (BRACIS), 2014 Brazilian Conference on (pp. 50-54). IEEE.

[22] Toman M, Tesar R, Jezek K. Influence of word normalization on text classification. Proceedings of InSciT. 2006 Oct 25;4:354-8.

[23] Christopher, D.M., Prabhakar, R. and Hinrich, S.C.H.Ü.T.Z.E., 2008. Introduction to information retrieval. An Introduction To Information Retrieval, 151, p.177.

[24] Dařena, F. and Žižka, J., 2015. Interdependence of text mining quality and the input data preprocessing. In Artificial Intelligence Perspectives and Applications (pp. 141-150). Springer, Cham.

[25] Hidayatullah, A.F. and Ma'arif, M.R., 2017, January. Pre-processing Tasks in Indonesian Twitter Messages. In Journal of Physics: Conference Series (Vol. 801, No. 1, p. 012072). IOP Publishing.

[26] Wegrzyn-Wolska, K., Bougueroua, L., Yu, H. and Zhong, J., EXPLORE THE EFFECTS OF EMOTICONS ON TWITTER SENTIMENT ANALYSIS. Computer Science & Information Technology, p.65.

[27] Gull, R., Shoaib, U., Rasheed, S., Abid, W. and Zahoor, B., 2016. Pre Processing of Twitter's Data for Opinion Mining in Political Context. Procedia Computer Science, 96, pp.1560-1570.

[28] Arjun Srinivas Nayak , Ananthu P Kanive , Naveen Chandavekar, Bala subramani R. ,2016. Survey on Pre-Processing Techniques for Text Mining. International Journal of Engineering and Computer Science. 5(6),pp. 16875-16879.

[29] Moral, C., de Antonio, A., Imbert, R. and Ramírez, J., 2014. A survey of stemming algorithms in information retrieval. Information Research: An International Electronic Journal, 19(1), p.n1.

[30] Ramasubramanian, C. and Ramya, R., 2013. Effective pre-processing activities in text mining using improved porter's stemming algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(12), pp.4536-8.

[31] Clark, A., 2003, March. Pre-processing very noisy text. In Proc. of Workshop on Shallow Processing of Large Corpora(pp. 12-22).

[32] Bao, Y., Quan, C., Wang, L. and Ren, F., 2014, August. The role of pre-processing in twitter sentiment analysis. In International Conference on Intelligent Computing (pp. 615-624). Springer, Cham.

[33] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S., Schaller, S. and Holzinger, A., 2012. On text preprocessing for opinion mining outside of laboratory environments. Active media technology, pp.618-629.

[34] Krouska, A., Troussas, C. and Virvou, M., 2016, July. The effect of preprocessing techniques on Twitter sentiment analysis. In Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on (pp. 1-5). IEEE.

[35] Raza, A.A., Habib, A., Ashraf, J. and Javed, M., 2017. A Review on Urdu Language Parsing. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 8(4), pp.93-97.

[36] Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer US.

[37] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[38] Pang, B. and Lee, L., 2004, July. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[39] Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., Lalande, A., Bhanderi, S., Belfer, R., Kanagasabai, N. and Sarrazingendron, R., 2018. Sentiment Analysis: It's Complicated!. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (Vol. 1, pp. 1886-1895).

[40] Raza, A.A., Habib, A., Ashraf, J. and Javed, M., 2018. Semantic Orientation Based Decision Making Framework for Big Data Analysis of Sporadic News Events. Journal of Grid Computing, pp.1-17.

[41] Yu, H. and Hatzivassiloglou, V., 2003, July. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 129-136). Association for Computational Linguistics.

[42] Maynard, D. & Funk, A., 2012. Automatic Detection of Political Opinions in Tweets. In CEUR Workshop Proceedings. pp. 88–99, Venezia, Italy.

[43] Haddi, E., 2015. Sentiment analysis: text, pre-processing, reader views and cross domains (Doctoral dissertation, Brunel University London).

[44] Etaiwi, W. and Naymat, G., 2017. The Impact of applying Different Preprocessing Steps on Review Spam Detection. Procedia Computer Science, 113, pp.273-279.