Feature Extraction and Opinion Mining of Gujarati Language text

Himadri Patel[#], Bankim Patel[#], Kalpesh Lad[#]

[#] *Uka Tarsadia University*

***Abstract* – The field of opinion mining has gained much popularity in last few years. Many new techniques and methods are being developed in different languages like English, Hindi etc. However, it is observed that there is no significant progress in the field of Opinion Mining for languages like Gujarati. The presented work uses a deep learning approach for the Opinion Mining of Gujarati language text. The paper also discusses feature extraction which is one of the most important steps in machine learning or deep learning method.**

***Keywords* – Feature Extraction, Opinion Mining, Gujarati, Deep Learning, CNN.**

I.INTRODUCTION

Opinion Mining is considered as a classification problem that identifies if the opinion is positive or negative and is done using the classification methods like SVM, KNN, CNN etc [1]. The literature contains work which proves the SVM to be better performing method among the traditional ML methods SVM and Naïve Bayes [2, 3]. The deep learning methods are proven successful in literature of Opinion Mining for Persian language and Arabic language [4, 5]. However, the Gujarati language or other Indian languages are not explored to use use the deep learning methods.

Opinion Mining performed using classification methods uses features which needs to be extracted from the dataset. A good feature is said to be the features that are more expressive, domain dependent, occur rarely and are selected based on document frequency [6]. One of the first ever work done in the Opinion Mining using Machine Learning technique uses TFiDF for vectorization and “Bag-of-words” as its features [7]. They claim that the order of word is very less significant while using SVM as the ML technique. However, other work done for the Opinion Mining introduces an enhanced method called Delta TFiDF and claims that this method best suits for Opinion Mining as their assigned weight is biased towards one corpus either positive or negative. Other methods used for the vectorization are Word2Vec and Doc2Vec in [8] to vectorize the articles on trending topics that updates every hour. They claim that Doc2Vec outperforms TFiDF as it supports dynamically changing vocabulary. However, Doc2Vec or Word2Vec needs the use of Language Model as its prerequisite.

In Opinion Mining, the vectorization is performed on term or tokens which are then converted into features like Lemma, N-gram, POS tagging, Syntactic Dependency Tree etc [9]. One of such feature selection technique uses combination of Lemma with its POS tags adjective, adverb, noun and verb from the sentence [9]. They claim that this combination of feature outperforms the common N-gram technique as it captures the dependency. Another work of Opinion Mining of News headlines is done using the SVM which experiments with unigram and bigram features with count vectorization and TFiDF vectorization [3]. They show that bigram technique generates a greater number of useful features compare to unigram and also claim that TFiDF with the use of bigram performs best. Work has also been done to identify the sentiment of writer by differentiating the semantic features and syntactic features where they used SVM as the ML technique [6]. The Word2Vec or Doc2Vec cannot be used for Gujarati language due to unavailability of Language model and for the other vectorization methods, TFiDF is proven better than count vectorization. Authors use TFiDF for vectorization and the combination of POS tagging, N-gram and Rule Based method for the extraction of features.

S. Rana and et.al [2] explored sentiment orientation using the SVM and Naïve Bayes method using the film reviews. They claimed that SVM provides best accuracy. J. Chaudhary and et.al [3] also experimented with SVM along with other NLP tools on headlines taken from the newspapers and claimed 91.52% accuracy with bi-gram features. This shows that SVM is proven to perform well for the Opinion Mining tasks.

However, the more recent approaches are using the deep learning algorithms. S. Zobeidi and et.al [4] used CNN for the Persian language Opinion Mining using Word2Vec vectorization and showed that the two class classification achieved 95% and multi-class classification achieved 92% accuracy. H. Elzayady [5] also experimented the Opinion Mining task using CNN for Arabic langauge text and achieved 86.88% accuracy. This shows the successful use of the CNN algorithm as well.

Identifying the strength of the opinion is considered as an important challenge in the Opinion Mining and which is addressed by assigning the weight to the opinion bearing words [10]. Whenever N-gram or Bag of Words method is used for feature extraction, the whole feature needs to be given a weight to identify the weight of overall sentence which at the end gives the strength of the opinion. This challenge is addressed in work done on English language opinions. A very common approach for weighing opinion is to use SentiWordNet. SentiWordNet is a database comprising of nouns, verbs, adjectives and adverbs of language in sets of synonyms and antonyms [11] and are also assigned weight by training the synset to generate scores. The SentiWordNet is developed for a few Indian languages [12] but it does not include the SentiWordNet for Gujarati language. Another approach used is to prepare a list of positive word and negative words, assign same weight to all the positive and negative words and use the average to assign weight to the opinion. Above this, the distance between feature and the opinion word is also used to calculate the weight by authors in [8] which enhances the performance by relating the feature with its opinion. They also use rules that add the solution to context dependency issue. This work is enhanced in one of the Opinion Mining tools that adds more weight to the opinions found in the title of the Opinion [13]. So, this approach is useful when the dataset contains both opinions and their titles.

Another approach used to weight the opinion is to estimate the weight using Information Gain method which also uses the sentiment dictionary to add the semantic grading to the opinion [14]. Their semantic dictionary gives weight to the opinions based on its strength. For example, “Brilliant” gets weight 1 while “Very good” gets weight 0.8. They also relate the opinions with the feature using the distance which allows multiple features to be evaluated in a single opinion. This work involves assigning weight manually to the words which needs the linguistic expertise. The linguistic rules are used in [15] in order to solve the problem of semantic orientation in the opinion. It creates segments of sentence based on the BUT phrases used in the opinion and find the segment specific semantic score for the opinion using its distance from the object word. Their linguistic rules are based on the conjunctions found in the sentence like ‘and’, ‘but’, ‘however’. The presented paper uses a suitable method to calculate weight to the opinions written in the Gujarati language. Experiments are done to conclude about which method is suitable.

The work of Opinion Mining is found in Indian languages like Hindi, Odia, Telugu, Tamil, Marathi [13, 8] using the Machine Learning based approach, SentiWordNet based approach, Lexicon based approach. Hindi Opinion Mining is developed using Supervised method [14] in which they used Unigram, Best Word and Best Word + Chi Square as a feature and claimed 87.1% accuracy. However, they didn’t experiment with Bigram or Trigram. A Punjabi Opinion Mining is also carried out using the Naïve Byes approach [15] where they used N-gram approach with the variety of unigram, bigram and trigram. But they kept each N-gram feature to be unique during training and testing. Also, the context dependency is not considered in their work. Another work of Opinion Mining is done in Devnagari uses TFiDF and Count Vectorizer as vectorization with no other specific method for feature selection [13]. English SentiWordNet [16] is used in [17] to generate a Sinhala SentiWordNet by translating English words to Sinhala words and their synonyms. They made a few assumptions that the sense of the word, POS of the word and opinion score of word of Sinhala language to be same as of English language.

A Gujarati language is very less explored in this field. The effort to mine Gujarati tweets are done using the SVM method [18] claims 92% accuracy but the tool is tested on less than 50 tweets which is not considered as a good dataset. Another effort in the field of Opinion Mining is done in [19] which only focuses on identifying object from the Opinion using the POS tagger. The nouns or noun phrase found in the Opinion is considered as object but this work doesn’t focus on identifying the polarity. Lata Gohil et al. [20] developed a lexical sentiment resource, Gujarati SentiWordNet(G-SWN), using the synonym relations of Hindi SentiWordNet(H-SWN) and IndoWordNet(IWN). The G-SWN can be used for Gujarati opinion mining of Gujarati text. They also used Gujarati corpus. The limited work in this field is due to the challenges faced by researchers [20].

This shows that some work is available for the field of Opinion Mining for Gujarati language but it doesn’t contain extensive comparison of features so as to choose the best feature selection method. Also, the existing work in Gujarati language is very limited to work with nouns and adjective with a very limited dataset. Hence, there is a wider scope of work in Opinion Mining for Gujarati language text. This paper discusses about the experiment done for extracting features for the Opinion Mining tool in Gujarati language text. It uses the English SentiWordNet to prepare a raw Gujarati SentiWordNet, processes it to add the language dependent element. The extracted features and the weight are then experimented using a Machine Learning technique to identify the best feature. Authors decided to experiment using CNN and SVM to conclude on the best performing method as SVM is traditionally proven best method and CNN is very less explored still have successful use for the Opinion Mining applications. A corpus is developed for Gujarati language text in Education domain to be used for this research.

The work only focuses on the feature extraction of the Opinion Mining and doesn’t focus on an overall framework of opinion mining in broader scope which may also include detailed description of the preprocessing steps.

II. SENTIWORDNET OF GUJARATI LANGUAGE

Gujarati SentiWordNet is developed by translating the English synsets with the help of Google translation tool. The assumption is that the score for English language remains same when translated to Gujarati language. Before processing translation with the English SWN, its detailed analysis was performed in order to decide the further steps so that the gap between the English and Gujarati can be carefully filled.

With the objective to develop the basic version of SWN, the authors aim to develop the SWN in form of a look-up table with unique lemma and their positive and negative sentiment score. However, the SWN of English is a network of synsets which is prone to have same word exists in multiple synset with different score based on the context of the semantic link. It means that at the end of translation process, the look-up table contains words with different scores of a same word as well as multiple entries of a same word with same score. The repeated words having same score were straight forward uniquified and the repeated words having different score were given to the language expert to decide the best score for the word. This entire process resulted in the lookup table with 6033 unique words with their positive, negative or both positive and negative scores for each translated word.

III.DATASET PREPARATION

Development and validation of the ML or AI based tool needs precise and sufficient training data that supports the algorithms to comprehend certain series or patterns of problem outcomes. In this research work reviews written in Gujarati language for the Education domain, which is neither prepared by any researcher till date nor is available on internet. The efforts are made to find online sources where the enough amount of data is available as per the necessities of the presented research or to develop a review or opinion dataset to use in this research. One website[1] has a section where people write their opinion on various topics in Gujarati language but they were not proven very useful as they don't belong to Education domain. There was a need of developing corpus specifically for Education domains in Gujarati language text.

The corpus development task is carried out in two basic ways i) by creating online blogging tool and ii) manual data gathering for those users who are not comfortable using online tool. Approximately 3100 opinions were collected from more than 440 users who gave their opinions on one or more topics from the 12 topic titles. The 3100 opinions were then converted in form of lines which makes 9371 lines where 4799 are positive lines and 4572 are negative lines and is used in this research for the training and validation of Opinion Mining tool.

Corpus development is a crucial process which affect the performance of the tool. So, each step taken during this process needs to be carefully drafted and executed so as to develop a good quality corpus in minimum duration. As the requirements of presented work is to develop corpus for Education domain, the objectives of Education System of India [21] were studied to choose the topics to be covered while collecting people's opinion.

IV.OPINION MINING

The task of opinion mining is carried out using Deep Learning based method. The literature discusses Machine Learning methods which are proven as a better solution for the Opinion Mining task for the English language and other languages for the domains like product review, movie review etc. As the objective of the developed framework is not limited to identifying the polarity (i.e., positive or negative) but to also identify the polarity score of the sentence so as to understand the strength of the opinion, authors considered four classes for the scores, -0.875 and -0.375 for negative sentences and 0.375 and 0.875 for positive sentences.

This work focuses on Opinion Mining of Gujarati language text with the domain Education. The authors experimented with two methods considering the successful use of these methods in literature [4, 5, 3, 2] – one in machine learning and another in deep learning – Support Vector Machine (SVM) and Convolutional Neural Network (CNN) respectively.

A.Support Vector Machine

The SVM algorithm creates a hyperplane that segregate n classes into n-dimensional spaces. The presented framework has four classes in the problem statement so the model prepared by SVM is of four dimensions with three-dimensional hyperplane. The experiments are carried out using Linear and RBF kernels with the same set of data in order to identify the best performing kernel.

B.Convolutional Neural Network

The CNN is a deep learning algorithm which is developed on the concept of creating convolutional layers in the network in order to extract high level features. The presented framework performs max-pooling and flattening while generating model of CNN. The experiments are performed with activation functions like relu and sigmoid, kernel size, batch size and number of epochs with the purpose of identifying best suited hyper-parameters for the network.

V.FEATURE EXTRACTION

The data used as an input for mining opinion are in form of text and are processed step by step in order to identify the polarity of the sentence. One of the majorly used methods to process this text is Machine Learning (ML) methods. ML methods contain two core phases: training phases and testing/evaluation phase. The tool developed using the ML method is trained based on relevant features and the same feature extraction is used during testing phase or when the tool gets actual text to be evaluated. The more accurate the feature extraction is, the more accurate the machine will get trained [22]. Hence, the feature extraction is the heart of the entire ML based tool [9]. All the methods of machine learning give best result when the extracted features are quality features.

In NLP, the feature extraction method contains two basic steps: vectorization and feature extraction. Text documents typically contain a string of characters which need to be transformed to the representation which an ML algorithm can understand and train itself [7]. In the process of vectorization, a number is given to each root/stem of a word used in the document. The methods used to assign the appropriate numeric value to the root are Term Frequency method, Term Frequency inverse Document Frequency(TFiDF) method, Count Vectorization, Information Gain, Mutual Information, Chi-square approach etc [23]. These methods also perform feature reduction at some level [7]. The vectorized text is then

[1] *http://opinionmagazine.co.uk*

used to extract the useful features during the feature extraction process. For the NLP applications, the usual feature extraction methods used are bag-of-words [7], Word2Vec and Doc2Vec [8], n-gram [9] etc. which is then used by the ML method for further processes of training, testing and evaluation. The methods Word2Vec and Doc2Vec need a language model which is not available for the Gujarati language. Hence, authors experimented with n-gram approach where the text features are used and experimented up-to three-gram.

Another approach to extract features from the text is to identify the important characteristics of text which plays important role to determine the opinion of the sentence. In Opinion Mining, the words bearing POSs adjective and nouns plays crucial roles. Other than that, the present positive words and negative words in the sentence are also decisive part of the sentence. Considering this fact, authors extracted score of each such word from the SentiWordNet and created 13 number features like sum of score of all adjectives, sum of score of all nouns, sum of score of all positive words, sum of score of all negative words, average of scores of adjectives, nouns, positive words and negative words and count of adjectives, nouns, positive words and negative words.

The process of feature extraction takes input as a line which is given to the POS tagger [24] to tag each word. Stop words are removed from the line before giving it to the tagger. Each word is then checked for the inflections using Lemmatizer [25] and the incorrect spelling using the Spell Checker [26] in order to achieve the root word. Once the root words from the text are extracted, they are further used in the process of feature extraction and these features are used for the training and testing/evaluation of the Opinion Mining tool developed using CNN.

VI. RESULT ANALYSIS

The authors experimented with text features and number features using the methods SVM and CNN and used precision, recall, f-measure and accuracy to compare the result in order to draw the conclusion. **Error! Reference source not found.** shows the comparison of precision, recall, f-measure and accuracy using SVM method. The result shows that the best performing feature is a combination of one-gram and two-gram with 68% accuracy and two-gram as well as three-gram performed poor. The result also shows the difference between precision and recall is large which should be minimum.

Table 1 Feature comparison using SVM

Sr. No.	Feature details	Precision	Recall	F-measure	Accuracy
1.	Text (one gram)	73	60	63	67
2.	Text (two gram)	70	57	59	64
3.	Text (three gram)	67	51	52	59
4.	Text (one gram, two gram, three gram)	71	58	61	65
5.	Text (one gram, two gram)	74	61	64	68
6.	Text (one gram, three gram)	72	59	62	66
7.	Text (two gram, three gram)	70	57	59	64
8.	Number features	69	60	62	65

The **Error! Reference source not found.** shows the comparison of precision, recall, f-measure and accuracy using CNN method. It shows that n-gram features performed poor with best performance of one-gram having 64% accuracy. However, the number feature performed best with 75% accuracy.

Table 2 Feature comparison using CNN

Sr. No.	Feature details	Precision	Recall	F-measure	Accuracy
1.	Text (one gram)	66	58	59	64
2.	Text (two gram)	57	52	51	58
3.	Text (three gram)	64	52	52	59
4.	Text (one gram, two gram, three gram)	61	54	54	60
5.	Text (one gram, two gram)	65	58	58	63
6.	Text (one gram, three gram)	60	54	53	60
7.	Text (two gram, three gram)	61	54	54	61
8.	Number features	79	70	72	75

It is observed that the precision and recall does not vary much with CNN with 79% and 70% respectively. It shows that CNN performed better compared to SVM and also that the number feature performed better compared to n-gram for CNN.

VII. SUMMARY

The authors developed Opinion Mining framework and SentiWordNet for Gujarati language text. The experiments are carried out with n-gram features and number features using SVM and CNN methods. The authors also developed a dataset containing 9371 lines with 4799 positive lines and 4572 negative lines for the training and testing/evaluation of the tool. The experiments shows that the CNN method performed best with number features extracted from the text and it's score with 75% accuracy.

VIII. FUTURE SCOPE

The Opinion Mining framework discussed in this paper can be enhanced to handle the challenges like negation or deviation in the sentence. The work can also be tested on LSTM algorithm on a larger dataset. The work can also be enhanced to experiment this approach on another domain's dataset.

REFERENCES

[1] M. Husnain, M. Missen, N. Akhtar, M. Coustaty, S. Mumtaz and S. Prasath, "A systematic study on the role of SentiWordNet in opinion mining," *Front. Computer Science,* 2019.

[2] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using svm and naive bayes techniques," in *2016 2nd International Conference on Next Generation Computing*

Technologies (NGCT), IEEE, October, 2016.

[3] J. Chaudhary and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2152-2163, June 2019.

[4] S. Zobeidi, M. Naderan and S. Alavi, "Opinion mining in Persian language using a hybrid feature extraction approach based on convolutional neural network," in *Multimedia Tools and Applications, Springer Science+Business Media, LLC, part of Springer Nature 2019*, August 2019.

[5] H. Elzayady, K. Badran and G. Salama, "Arabic Opinion Mining Using Combined CNN - LSTM methods," *I.J. Intelligent Systems and Applications*, pp. 25-36, August 2019.

[6] R. S. Rahate and M. Emmanuel, "Feature selection for sentiment analysis by using SVM," *International Journal of Computer Applications*, vol. 84, no. 5, pp. 24-32, 2013.

[7] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol. 1398, 1998.

[8] S. Shah and A. Kaushik, "Sentiment Analysis on Indian Indigenous Languages: A Review on Multilingual Opinion Mining," *Preprint*, November 2019.

[9] S. Siddiqui, M. A. Rehman, S. M. Daudpota and A. Waqas, "Opinion Mining: An Approach to Feature Engineering," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 3, 2019.

[10] M. Kumar, S. Amirneni and S. Prabhu, "Sentiment Ranking for Opinion Extraction by Weighted Feature Scheme," *International Journal of System Modeling and Simulation*, vol. 2, no. 1, pp. 7-13, March 2017.

[11] G. A. Miller, "WordNet: a lexical database for English," in *Communications of the ACM*, 1995.

[12] A. Das and S. Bandyopadhyay, "SentiWordNet for Indian languages," in *Proceedings of the eighth workshop on Asian language resouces*, 2010.

[13] S. Shah, A. Kaushik, S. Sharma and J. Shah, "Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models," *Big Data and Cognitive Computing 4*, vol. 3, no. 1, 2020.

[14] V. Jha, N. Manjunath, D. Shenoy, K. R. Venugopal and L. M. Patnaik, "HOMS: Hindi Opinion Mining System," in *Proceeding of IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, 2015.

[15] A. Kaur and V. Gupta, "N-gram based approach for opinion mining of Punjabi text," in *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, Cham, 2014.

[16] E. Andrea and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," *LREC*, vol. 6, pp. 417-422, 2016.

[17] M. Nishantha, S. Shanmuganathan and J. Whalley, "Sentiment lexicon construction using SentiWordNet 3.0," in *11th International Conference on Natural Computation (ICNC), IEEE*, 2015.

[18] V. Joshi and V. Vekariya, "An Approach to Sentiment Analysis on Gujarati," *Advances in Computational Sciences and Technology*, vol. 10, no. 5, pp. 1487-1493, 2017.

[19] H. Patel, R. Mehta, A. Shaikh, R. Mehta, N. Patel and D. Patel, "Object or its feature identification from Mobile Reviews in Gujarati Language," *GIT-Journal of Engineering and Technology*, vol. 9, no. 1, pp. 36-39, 2016.

[20] H. Patel and B. Patel, "A critical study of challenges in educational opinion mining of text written in Gujarati language," *National Journal of System and Information Technology*, vol. 9, no. 1, pp. 25-34, 2016.

[21] D. R. Sirswal, "Philosophy, Education and Indian Value System," 2011.

[22] R. Imran, H. Banka and H. M. Khan, "A Hybrid Feature Selection Approach Based on LSI for Classification of Urdu Text," in *Machine Learning Algorithms for Industrial Applications, Springer*, Cham, 2020.

[23] Saha, S. Kumar, P. Mitra and S. Sarkar, "A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition," *Knowledge-Based Systems*, vol. 27, pp. 322-332, 2012.

[24] G. Information Retrieval Lab DA-IICT, February 2021. [Online]. Available: http://irlab.daiict.ac.in/tools.php.

[25] H. Patel, P. Bankim, L. Kalpesh and S. Jikitsha, "Stemmer based hybrid Gujarati lemmatizer". India Patent Indian Patent No 201821025419 A, 7 July 2018.

[26] H. Patel, B. Patel and K. Lad, "Jodani: A spell checking and suggesting tool for Gujarati language," in *Confluence-2021:11th International Conference on Cloud Computing, Data Science & Engineering*, Noida, India, 2021.

[27] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1, pp. 1-135, 2008.