Multilabel Classification for Emotion Analysis of Multilingual Tweets

Lata Gohil, Dharmendra Patel

Abstract**:** ***Emotion Analysis of text targets to detect and recognize types of feelings expressed in text. Emotion analysis is successor of Sentiment analysis. The latter does coarse-level analysis and classify the text into positive and negative categories while former does fine-grain analysis and classify text in specific emotion categories like happy, surprise, angry. Analysis of text at fine-level provides deeper insight compared to coarse-level analysis. In this paper, tweets are classified in discrete eight basic emotions namely joy, trust, fear, surprise, sadness, anticipation, anger, disgust specified in Plutchik's wheel of emotions*** [1]. ***Tweets for three languages collected out of which one is English language and rest two are Indian languages namely Gujarati and Hindi. The collected tweets are related to Indian politics and are annotated manually. Supervised Learning and Hybrid approach are used for classification of tweets. Supervised learning uses tf-idf as features while hybrid approach uses primary and secondary features. Primary features are generated using tf-idf weighting and two different algorithms of feature generation are proposed which generate secondary features using SenticNet resource. Multilabel classification is performed to classify tweets in emotion categories. Results of experiments show effectiveness of hybrid approach.***

Index Terms**:** ***Emotion Analysis, Sentiment Analysis, Affect Analysis, Fine-grained, Hindi Corpus, Gujarati Corpus, Opinion Mining***

I. INTRODUCTION

The Internet users have surged in last few years globally. Worldwide 3.48 billion social media users are reported in year 2019 [2]. The Internet users and thus social media users too are rapidly increasing. People express their opinion and feeling about events, happening through social media posts. Social media users from all over the world post their comments in their regional language apart from English language. To understand sentiment and emotions of social media users, monolingual text analysis would not be sufficient. Thus, multilingual text analysis is the need of the day. This study proposes new methodology for multilabel classification of tweets into eight discrete emotions. Datasets have been developed for Gujarati, Hindi and English languages to evaluate proposed methodology.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author

Lata Gohil*, Computer Science and Engineering Department, Institute of Technology, Nirma University, India Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), CHARUSAT, Changa, India

Dharmendra Patel, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA), CHARUSAT, Changa, India

II. RELATED WORK

It is important to analyze multilingual social media content as it allows to understand how people from different geographical area and cultural background view various events and happenings [3].

Early research work on emotion analysis was predominantly focused on English language. However increasing amount of non-English languages content on Internet leads to need of analyzing non-English languages and multilingual content too.

Social media users use emoticons and emoji characters to express their feeling. This emotion tokens express emotions regardless of the language of the post on social media. The study in [4] has used emoticons and emoji characters to classify tweets into different emotions with the assumption that emoticons represent overall emotion expressed in the tweets. While study in [5] has mentioned that "a few emoticons are strong and reliable signals of sentiment polarity but a large group of the emoticons conveys complicated sentiment hence they should be treated with extreme caution".

The work in [3] has done pilot study for multilingual emotion analysis for social media content for English, Finnish, Swedish, Spanish and Portuguese languages using lexicon translation approach. Multilingual emotion lexicon is used for emotion analysis. Parallel corpora of movie subtitles are used as proxy for colloquial language. The main aims of the study are to determine at what extent the emotions are preserved in translation and reliability of pure lexicon approach for emotion analysis across languages. Better cross language agreement is achieved for languages having less cultural differences. Result shows that the lexicon approach gives good inter-language agreement but result of manual evaluation conducted suggests the need of further study to prepare better emotion classifier to overcome the limitation of lexicon approach.

The study in [6] has developed WordNet Affect lexicon resource and corpora for emotion analysis in three Indian languages namely Hindi, Bengali and Telugu using translation and automatic expansion approach. Satisfactory results have been reported for baseline system and morphology driven systems in comparison with English language.

The study in [7] has done emotion analysis of multilingual tweets using supervised machine learning algorithms namely naive bayes and svm.

Different groups of features are generated using lexicon resources namely WordNet-Affect [8], Hindi WordNet-Affect (HWNA) [9] and SentiwordNet [10]. Fine-grained emotion classification is done for six emotion categories of Ekman emotion model [11].

Tweets are classified into one emotion category which is best expressed.

In our study, multilingual tweets are classified in one or more emotion categories of Plutchik's wheel of emotions using supervised learning and hybrid approach.

III. MOTIVATION

Emotion analysis provides better understanding of users' feeling expressed in text compared to classifying text for polarity labels such as positive and negative. English language is heavily explored for Sentiment analysis at coarse-level and majority of non-English languages also have been studied for the same. Fine-level sentiment analysis as Emotion analysis which classifies text into discrete emotions is research field where English language is much investigated compare to non-English languages and specifically for Indian languages work reported for the same is very less.

Social Media platform provides opportunity to their users to use regional languages. Users also like to express their feelings in their native language. Thus, social media is rich source of multilingual data. India is multilingual country. English language is being used in higher education and Hindi is widely spoken language in India. Gujarati is official language of Gujarat state of India. There are large Gujarati immigrant communities in India and world too. In 2018, the number of social media users in India was 326.1 million and expected to be about 448 million in 2023 [12]. This motivate us to perform emotion analysis of multilingual social media data.

IV. DATASET COLLECTION AND ANNOTATION

To classify sentiment or emotion from text using supervised learning, annotated dataset is needed. Decent amount of annotated sentiment datasets for English text is developed while annotated emotion datasets in English text is comparatively less. In case of Indian languages, annotated datasets are few for sentiment and notably less for emotion.

We have developed annotated corpus for English and two Indian languages namely Gujarati and Hindi. Political tweets are collected and annotated for eight discrete emotions of Plutchik's wheel of emotions namely anger, anticipation, disgust, fear, joy, sadness, surprise, trust. Each tweet may express more than one emotions so each tweet is annotated for one or more emotion labels.

Tweets are collected using Twitter search API for the India's general election 2019. Specific keywords set and language filters are applied to twitter search API. Statistics of collected tweets are given in Table I.

Table I : Dataset Statistics

Language	No. of Tweets
Gujarati	1822
Hindi	5833
English	5000
Total No. of Tweets	**12655**

Tweets were manually annotated by three annotators. Tweets were annotated for multiple labels namely joy, trust, fear, surprise, sadness, anticipation, anger, disgust. Tweets were discarded during annotation process which were either not expressing any emotions or not related to political domain. The statistics of annotated tweets are presented in Table II.

Table II : Annotated Dataset Statistics

Language	No. of Tweets
Gujarati	1557
Hindi	3584
English	3915
Total No. of Tweets	**9056**

Kappa statistics for more than two categories are not giving promising result [13]. We have prepared two datasets of annotated tweets considering majority votes and all votes. If any two or more annotators annotate tweet for a given emotion, it is considered as majority vote. If minimum one annotator annotates tweet for a given emotion, it is considered as all vote, as vote of each annotator is taken into consideration. We call majority vote corpus as MV dataset and all vote corpus as AV dataset. Statistics of annotated datasets as per eight discrete emotions are listed in Table III.

V. EXPERIMENT

Supervised learning approach and Hybrid approach are used for multilabel emotion classification to classify tweets of Gujarati, Hindi and English languages in discrete eight emotions.

We have explored hybrid approach for emotion classification as problem with the lexicon method is low recall while problem with machine learning method is domain independence [14].

Supervised learning uses only primary features which are generated using tf-idf vector. Hybrid approach uses primary and secondary features. Primary features are generated using tf-idf weighting while secondary features are generated from SenticNet [15] resource. SenticNet is a lexicon resource for concept-level sentiment and emotion analysis. It is an affective commonsense knowledge base inspired by Plutchik's studies on human emotions. SenticNet provides sentic vector for a given concept which is four-dimensional vector. Sentic vector synthesizes the emotion exposure of concept in terms of Pleasantness, Attention, Sensitivity, and Aptitude. This sentic vector is used for generating secondary features.

A. Proposed Feature Generation Algorithms

Hybrid approach uses primary and secondary features. Primary features are generated using tf-idf weighting. Both algorithms use sentic vector to generate secondary features. One algorithm use only sentic vector of SenticNet resource. We call this algorithm SN Algorithm. Another algorithm calculates cosine similarities among sentic vectors. We call this algorithm CS-SN Algorithm.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- **SN Algorithm**

This algorithm takes sentic vector of each primary features and calculate Euclidean norm of each sentic vector to generate secondary features. In case of Gujarati language, primary features are translated in Hindi language as SentiNet resource is not available for Gujarati language.

Algorithm 1: Pseudocode for feature generation using Sentic vector

```
Input: Preprocessed Tweets
Output: Feature Vector
1. Generate primary features from tweets using TF-IDF weighting
   PF = { $pf_1, pf_2, \ldots pf_n$ }
2. If language is Gujarati, translate each $pf_i$ in Hindi.
3. Generate sentic vector for each primary feature $pf_i$
   PSV = { $pfs_1, pfs_2, \ldots pfs_n$ } where $pfs_i$ is sentic vector of $pf_i$
4. Take Euclidean norm of each $pfs_i$ to generate secondary features.
   SF = { $sf_1, sf_2, \ldots sfn$ }
```

- **CN-SN Algorithm**

It calculates sentic vector for discrete eight emotions. Secondary features are generated by taking cosine similarity between sentic vector of primary feature and eight emotion sentic vectors. Thus, it generates eight features from each primary features.

Algorithm 2: Pseudocode for feature generation using cosine similarity between sentic vector of primary feature and sentic vector of each eight emotions

```
Input: Preprocessed Tweets
Output: Feature Vector
1. Generate primary features from tweets using TF-IDF weighting
   PF = { $pf_1, pf_2, \ldots pf_n$ }
2. If language is Gujarati, translate each $pf_i$ in Hindi.
3. Generate sentic vector for each primary feature $pf_i$
   PSV = { $pfs_1, pfs_2, \ldots pfs_n$ } where $pfs_i$ is sentic vector of $pf_i$
4. Generate sentic vector for discrete eight emotions of Plutchik's wheel of emotions
   ESV = { $e_1, e_2, \ldots e_8$ }
5. Generate secondary features by taking cosine similarity between sentic vector of primary feature $pfs_i$ and sentic vector of each emotion vector $e_i$.
   SF = { $sf_{11}, sf_{12}, sf_{13}, \ldots sf_{18}$,
          $sf_{21}, sf_{22}, sf_{23}, \ldots sf_{28}$,
          …
          $sf_{n1}, sf_{n2}, sf_{n3}, \ldots sf_{n8}$ }
```

B. Proposed Architecture

Architectural block diagram is presented in Fig. 1. Tweets are preprocessed by performing stopwords removal, cleaning and tokenization. Cleaning is performed to remove urls and special symbols like punctuation marks.

Table III: Annotated Datasets Statistics as per emotion categories

Language	Emotions	Number of Annotated Tweets in Majority Vote (MV)	Number of Annotated Tweets in All Vote (AV)
Gujarati	Anger	225	586
	Anticipation	1090	1404
	Disgust	152	397
	Fear	29	269
	Joy	345	714
	Sadness	132	499
	Surprise	327	1040
	Trust	225	740
Hindi	Anger	271	1234
	Anticipation	2521	3448
	Disgust	516	1808
	Fear	104	1195
	Joy	362	1780
	Sadness	420	1792
	Surprise	1162	2773
	Trust	265	1142
English	Anger	155	1014
	Anticipation	2642	3714
	Disgust	59	1012
	Fear	80	1323
	Joy	332	1757
	Sadness	363	2028
	Surprise	753	2521
	Trust	166	1744

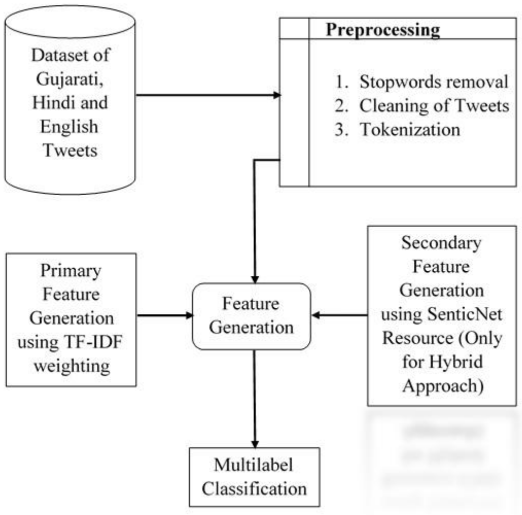
Dataset of Gujarati, Hindi and English Tweets
Preprocessing
1. Stopwords removal
2. Cleaning of Tweets
3. Tokenization
Primary Feature Generation using TF-IDF weighting
Feature Generation
Secondary Feature Generation using SenticNet Resource (Only for Hybrid Approach)
Multilabel Classification

Fig. 1 Architectural block diagram of emotion multilingual multilabel classification

Tf-idf weighing and sentic vector of SenticNet resource is used for feature generation. Primary features are generated using tf-idf weighting. For supervised learning, only primary features are generated. In case of hybrid approach, secondary features are generated using sentic vector of SenticNet resource. Two different algorithms mentioned in section V.A are used for the same. In case of Gujarati language, primary features are translated in Hindi. Multilabel classification is performed using supervised machine learning algorithms. Binary relevance method is used for multilabel classification.

VI. EXPERIMENT

Supervise learning and hybrid approach are two approaches used for Gujarati, Hindi and English tweets to perform multilabel classification of tweets in eight emotion categories. Two datasets are prepared namely Majority Vote (MV) and All Vote (AV). Three machine learning algorithms namely Logistic Regression, Multinomial Naive Bayes and SVM are used for classification.

Hybrid approach uses feature generation algorithm as mentioned in section V.A followed by machine learning algorithms for classification.

Table IV: Average F-measure of aggregated emotions

Language	Voting Types for Annotated Dataset	Experiment Type	Machine Learning Algorithms		
			Logistic Regression	Multinomial Naive Bayes	LinearSVC
English	All Votes (AV)	ML	**0.64**	0.6	**0.64**
		SN-ML	0.62	0.61	**0.63**
		CS-SN-ML	***0.65***	0.62	0.64
	Majority Votes (MV)	ML	0.81	0.78	**0.82**
		SN-ML	0.75	0.78	**0.82**
		CS-SN-ML	0.79	0.8	***0.84***
Hindi	All Votes	ML	***0.64***	0.62	***0.64***
		SN-ML	0.6	0.61	**0.62**
		CS-SN-ML	**0.63**	0.62	**0.63**
	Majority Votes	ML	***0.78***	0.71	***0.78***
		SN-ML	0.71	0.71	**0.76**
		CS-SN-ML	0.75	0.73	***0.78***
Gujarati	All Votes	ML	***0.75***	0.73	***0.75***
		SN-ML	0.68	0.68	**0.69**
		CS-SN-ML	**0.71**	0.68	0.7
	Majority Votes	ML	0.82	0.78	***0.84***
		SN-ML	0.68	0.72	**0.76**
		CS-SN-ML	0.74	0.74	**0.78**

For each datasets, three experiments namely supervised learning algorithm (ML), hybrid approach with SN-algorithm (SN-ML) and hybrid approach with CS-SN algorithm (CS-SN-ML) are performed. This six experiments are performed for three languages. Thus, total eighteen experiments are performed.

VII RESULT AND DISCUSSION

To evaluate the performance of multilabel classification methods, several metrics have been proposed in literature. In this study, average F-measure metric is used for evaluation purpose. Average of F-measure of eight emotion labels is calculated for each experiments mentioned in section VI. The same has been presented in

LinearSVC performs better than Logistic Regression and Multinomial Naive Bayes. Majority vote dataset gives better result compared to All vote as shown in Fig. 2.

For English language, CS-SN-ML outperforms for majority votes dataset. Thus, hybrid approach with CS-SN feature generation algorithm improves the LinearSVC classifier performance in case of English language. For Hindi language, machine learning approach and hybrid approach with CS-SN feature generation algorithm give better performance for Majority vote dataset. For Gujarati language, machine learning approach gives better performance than hybrid approach. Result degradation of hybrid approach for regional languages may be due to cultural influence which determines the expressed emotions in tweets. Other reason may be colloquial tweets. Words for certain emotions exist in one language may not exist in another language or even if they exist may not express the emotions with same intensity. These may be the reasons for degraded performance of hybrid approach for Gujarati language.

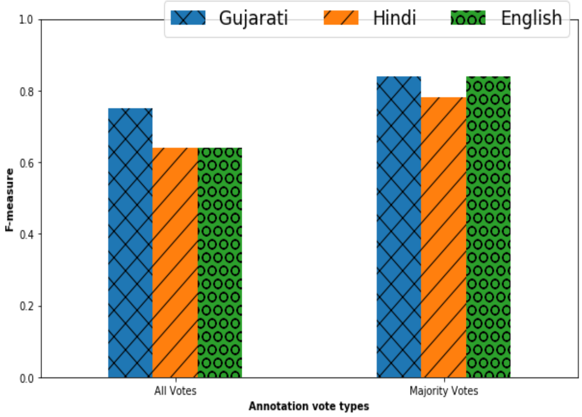

Fig. 2 Datasets performance

VIII CONCLUSION AND FUTURE WORK

The main aim of this study is to propose new methodology for classification of multilingual text into multiple emotion categories. Datasets are prepared by collecting and annotating tweets of Gujarati, Hindi and English languages for discrete eight emotions specified in Plutchik's wheel of emotions. Tweets are annotated manually by three annotators. By considering all votes and majority votes of annotation, two datasets are prepared for each languages. Machine learning and hybrid approach are used for classification of tweets. Two feature generation algorithms namely SN and CS-SN which make use of SenticNet resource are proposed for hybrid approach.

For English language, hybrid approach with CS-SN feature generation algorithm gives improved performance compared to machine learning approach and hybrid approach with SN feature generation method. For Hindi and Gujarati languages, performance of hybrid approach degraded due to cultural influence and colloquial tweets.

For Gujarati language, translation also plays role in degrading performance of hybrid approach. In future, hybrid approach performance for Gujarati language can be improved by developing SenticNet resource for Gujarati language.

REFERENCES

1. R. Plutchik, “Emotion: A psychoevolutionary synthesis: Harpercollins College Division,” 1980.
2. “Digital 2019: Global Internet Use Accelerates - We Are Social.” [Online]. Available: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates. [Accessed: 31-Oct-2019].
3. E. Öhman, T. Honkela, and J. Tiedemann, “The challenges of multi-dimensional sentiment analysis across languages,” in *Proceedings of the Workshop on Computational Modeling of People�s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2016, pp. 138–142.
4. W. Wolny, “Emotion analysis of Twitter data that use emoticons and emoji ideograms,” 2016.
5. H. Wang and J. A. Castanon, “Sentiment expression via emoticons on social media,” in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2404–2408.
6. D. Das, S. Poria, C. M. Dasari, and S. Bandyopadhyay, “Building resources for multilingual affect analysis--a case study on hindi, bengali and telugu,” in *Workshop Programme*, 2012, p. 54.
7. V. K. Jain, S. Kumar, and S. L. Fernandes, “Extraction of emotions from multilingual text using intelligent text processing and computational linguistics,” *J. Comput. Sci.*, vol. 21, pp. 316–326, 2017.
8. C. Strapparava, A. Valitutti, and others, “Wordnet affect: an affective extension of wordnet.,” in *Lrec*, 2004, vol. 4, no. 1083–1086, p. 40.
9. “Hindi WordNet, IIT Mumbai.” [Online]. Available: http://www.cfilt.iitb.ac.in/∼wordnet/ wn.old/.
10. A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining.,” in *LREC*, 2006, vol. 6, pp. 417–422.
11. P. Ekman, “An argument for basic emotions,” *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
12. “India - number of social network users 2023 | Statista.” [Online]. Available: https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/. [Accessed: 01-Aug-2019].
13. H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger, “Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus,” in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 13–23.
14. S. Goyal, N. Tiwari, and R. B. India, “EMOTION RECOGNITION: A LITERATURE SURVEY.”
15. E. Cambria, S. Poria, D. Hazarika, and K. Kwok, “SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

AUTHORS PROFILE

Lata Gohil is working as Assistant Professor in Computer Science and Engineering Department, Institute of Technology, Nirma University. She has received MCA degree from Gujarat Vidyapith, Ahmedabad. She has qualified GSET and GATE. Her research area is Information Retrieval and Text Mining. She is pursing PhD from CHARUSAT.

Dr Dhrmendra Patel received his Master of Computer Application degree from North Gujarat University. He received his PhD degree in computer science from Kadi Sarva Viswavidyalaya. His area of research is Web Mining, Fog Computing, Image Processing, Internet of Things etc. He has published 20 papers in national/international journal of repute. Currently he is working as an associate professor at CHARUSAT, Changa.