SENTIMENT ANALYSIS ON BANGLA YOUTUBE COMMENTS USING MACHINE LEARNING TECHNIQUES

[1] VEERANKI LAKSHMI DURGA, [2] A. MARY SOWJANYA
[1]M.Tech, [2] Assistant Professor
Dept of CSSE, Andhra University College Of Engineering (A),
Visakhapatnam, AP, India.

Abstract: Sentiment Analysis (SA) is an opinion mining study analysing people's opinions, sentiments, evaluations and appraisals towards Societal entities such as products, services, individuals, organizations, events, etc. Of late, most of the research works on SA in natural language processing (NLP) are focused on English language. However, it is noted that Bangla language does not have a proper dataset that is both large and standard. As a result, recent research works with Bangla language in SA have fallen short to produce results that can be both comparable to works done by others in other languages and reusable for further prospective research. In this work, a substantial textual dataset of both Bangla and Romanized Bangla texts have been provided which is first of this kind and post-processed, multiple validated, and ready for SA implementation and experiments. Further, in this project scraping video information from YouTube and validate the data samples into one of three categories: positive (1), negative (0) and neutral. In this work used real-time analytics, simply means that data is analyzed right after data becomes available. Real-Time Analytics can produce insights without any delay.

Keywords— Web-scraping; Bangla language; Romanized Bangla; Sentiment Analysis; Text blob

I.INTRODUCTION

Bangla is spoken as the first language by almost 200 million people worldwide, 160 million of whom are Bangladeshi [1]. Bangladeshi people are found to get increasingly involved in online activities such as - getting connected to friends and families through social media, expressing their opinions and thoughts on popular micro-blogging and social networking sites, sharing opinions and thoughts by means of comments on online news portals, doing online shopping through online marketplaces and other such applications. However, it is becoming increasingly harder for such businesses to monitor and analyze market trends, especially when it is done by analyzing the reaction of the customers on their products or services, due to less or no human-to-human interaction in such businesses. Moreover, the task of going through comments and reviews from each individual customers and figuring out the sentiments within is tedious and in some cases simply intractable, especially considering that - usually very high volume of data is generated very quickly in this day and age of digital connectivity. Therefore, application of automated Sentiment analysis (SA)

Sentiment Analysis can play a vital role here for enhancing efficiency and productivity.SA is widely employed as a machine learning application in many areas, and is known by many other terms e.g. opinion extraction, sentiment mining, opinion mining, subjectivity analysis, emotion analysis, review mining, etc. Most of the research works found on SA are based on the English language, while Bangla SA is still at a formative stage. An interesting work by Das and Bandyopadhyay [2] on subjectivity detection included Bangla but it is not self-sufficient, as English is also needed. However, none of the works truly considered Bangladesh's perspective. We need to consider not just standardized Bangla, but Banglish (Bangla words mixed with English words) and Romanized Bangla. These three major types can again be loosely categorized in - good, standard, bad, wrong, totally wrong, particular to specific location (almost arcane), etc., depending on the level of clarity, grammatical correctness, meaningfulness, personal idiosyncrasies, impact of localization etc. Moreover, for the Romanized Bangla the added complexity is due to the variation in transliteration between people who know English well and those who do not [3]. The reason, that no clear standard is followed when 160 million Bangladeshi people write in any of the mentioned types, makes it all the more complicated and challenging to work with.

In the recent past, Deep Learning methods, specifically recurrent model-based deep learning models have enjoyed a lot of success in Natural Language Processing (NLP), compared to more traditional machine learning methods [4]. While there are other approaches to SA, in this research we will concentrate exclusively on deep learning based techniques. Our key contributions cover –

- A Web-scraping of YouTube Bangla and Romanized Bangla text samples, where each sample was annotated by two adult Bangla speakers.
- Pre-processing the data in a way so that it is readily usable by researchers.
- Application of deep recurrent models on the Bangla and Romanized Bangla text corpus.
- Pre-train dataset of one label for another (and vice versa) to see if it gives better results.

The paper is organized as follows. In section 2, we discussed the background of our work and the works of others in the same field that inspired and helped us in a way. In section 3, we discussed in details about the dataset that we used for our experiments. Section 4 discusses the methodology and also includes the experimental setup for the deep recurrent models. Section 5 has all the discussion about various results found from our experimentation, and lastly the article concludes with section 6.

II. Background

A. Sentiment Analysis

A key point of our work is Sentiment Analysis, on Bangla (and Romanized Bangla) language. Although the term "Sentiment Analysis" may have appeared for the first time in Nasukaw and Yi [5] , research works on sentiment appeared as early as in 2000 [6-8]. With advent of social media on internet e.g. Facebook, Twitter, forum discussions, reviews, and its rapid growth, we were introduced to a huge amount of digital data (mostly opinionated texts e.g. statuses, comments, arguments etc.) like never before, and to deal with this huge data the SA field enjoyed a similar growth. Since early 2000, sentiment analysis has become one of the most active research areas in NLP.

However, most of the works are highly concentrated on English language, favoured by the presence of standard data sets. Standard datasets allow researchers to do their own experiments and compare their contributions with those of others. For the English language, an example of such a standard SA dataset is the IMDB Movie Review Data set, which contains 50,000 annotated (positive or negative movie review) movie reviews made by the viewers. This scraping was originally created by Maas, Daly [9] and since then has been used by a multitude of different studies.

A detailed survey paper [10] presented an overview on the recent updates in SA algorithms and applications, categorizing and summarizing total 54 articles that had been published till 2014. Godbole, Srinivasaiah [11] collected opinions from newspaper and blogs, and assigned scores indicating positive or negative opinion to each distinct entity in the text corpus to do SA. In [12], they proposed and investigated a paradigm to mine the sentiment from a popular real-time micro-blogging service like Twitter, and they fashioned a hybrid approach of using both corpus-based and dictionary-based methods in determining the semantic orientation of the tweets.

B. Sentiment Analysis for Bangla

It is quite unfortunate that there is no standard collection of data, such as - the IMDB dataset, Twitter corpus etc. for Bangla texts. One effort for standardization came from an automatic translation of positive and negative words of SentiWordNet [13]. However, no corpus was created from this work, thereby limiting its usage to word level determination of sentiment, rather than the more complex natural language processing methods. Additionally, such simplified techniques do not consider the variety of ways in which people usually write, e.g. spelling mistakes, using colloquial terms etc.

A small dataset of Bangla Tweets were collected along with Hindi and Tamil by Patra, Das [14], where the authors reported on the outcome of a shared Sentiment Analysis task of Indian languages. They used 999 Bangla tweets for training and 499 for testing. They did some post processing such as pruning of emoticons from the tweets and removal of duplicated posts. This data was annotated manually by native speakers. However, in terms of usability the dataset's small size is a limiting factor for modern deep learning techniques. Another similar collection was done in [15], where 1400 Bangla Tweets were collected automatically. However, their dataset is not publicly available, and the size of the dataset is rather small.

A slightly larger corpus was collected, automatically annotated and manually verified by Das and Bandyopadhyay [2], as their collection was almost 2500 Bangla text samples from news items and blog posts. The uniqueness of their collection over the ones collected by others [14, 15] was the average size of 288 words of their samples, which is quite a bit larger than the 144 character Tweet limit. With most of the other works proceeded in the similar way, the two biggest issues with the current state of affairs in Bangla SA research are - first and foremost, the absence of a standard and big enough dataset to compare against, which makes comparison of research work extremely difficult, and secondly, none of the Bangla SA research takes into account the very prominent practical aspect of the use of *Romanized Bangla* [3].

C. Deep learning

AI (Artificial Intelligence) has been traditionally done in two ways – i) Knowledge based, and ii) Representation learning based. Knowledge base approach to AI uses logical inference rules to reason about statements input by users. CYC was one of the most famous of such projects [16]. The failure of knowledge based approach was the driving force into finding a way to give AI the ability to gather its own knowledge by extracting patterns or learning from the data – popularly known as Machine Learning. This new algorithm was based on representation of data or feature. That is, the system is given a number of features about the task in hand on which it will give a decision. Clearly if any of the features were wrong, it would mean wrong representation of the data and the system would not perform well. To rectify this situation representation learning based [17] algorithm was used. This algorithm gave better results than the manually tailored representation of data, and allowed systems to adapt to new tasks with ease. However, using this algorithm it was required that high level abstract features from the raw data were extracted without any error caused by misinterpretation due to the factors of variation, as there can be such factors (e.g. an accent in speakers speech) which would cause false representation in absence of highly sophisticated (human like) understanding. However, deep learning performed better with this issue, as it provides with complex representations expressed in terms of a number of other simpler representations.

D. Recurrent Neural Network

Recurrent Neural Network or RNN in short, has been widely used in speech recognition, handwriting recognition, natural language processing and others. Moreover, RNN is the precursor to NLP. While traditional neural networks failed to create a persistent model that would somewhat mimic the way our memory cells work for learning and remembering information, RNN – a class of ANN, has an interesting model design with a loop used as a feed-back connection which makes the information persistent [18, 19]. The loop enables the flow of information from one step to the next. It is like there are multiple copies of same network, where a successor gets information from all the predecessors, connected in architecture that excels at processing sequential data.

E. Web-Scraping

It is a process of automating the extraction of data in an efficient and fast way. With the help of web scraping, you can extract data from any website, no matter how large is the data, on your computer. When you extract web data with the help of a web scraping too, you would be able to save the data in a format such as CSV. You would then be able to retrieve, analyze and use the data the way you want. So web scraping simplifies the process of extracting data, speeds it up by automating it and creates easy access to the scrapped data by providing it in a CSV format. In simple terms, web scraping saves you the trouble of manually downloading or copying any data and automatically. There are numerous applications of web scraping under this segment which can be sub-divided into different categories as follows:

- Competitor Price Monitoring
- Monitoring MAP Compliance
- Fetching Images and Product Descriptions
- Monitoring consumer Sentiment
- Aggregated News Articles
- Market Data Aggregation
- Extracting Financial Statement
- Insurance

Web-Scraping Applications in Data Science:

- Real-Time Analytics
- Predictive Analysis
- Natural Language Processing
- Machine Learning Training Models

IV Scraping Details

Our dataset is called the BRBT dataset where BRBT stands for Bangla and Romanized Bangla Texts. This Bangla Sentiment Analysis (SA) scraping consists of thousands of YouTube URL. This is unique because not only this is larger compared to others, but it also encompasses the till-now-ignored Romanized Bangla. Romanized Bangla is the Bangla written in English alphabets. Inclusion of Romanized Bangla in the dataset is paramount, because the ease of writing Bangla using any standard QWERTY keyboard (without a Bangla keyboard e.g. Bijoy® keyboard) and the simplicity of using English as base language for the posts, have popularized Romanized Bangla not just in personal messages and micro-blogs but also in Govt. sanctioned mass messages/announcements. The dataset is currently kept private for safe keeping and further improvement. However, it may be available by personally contacting the owner/authors, and signing a consent form.

Post collection data processing

- *Removal of emoticons:-* emoticon, hash-tags were removed to give annotators an unbiased-text-only content to make a decision based on three criteria - positive, negative and ambiguous.
- *Removal of proper nouns:-* Proper nouns were replaced with tags to provide ambiguity. All text samples were collected from publicly available sources and did not reflect the opinion of the authors.
- *Manual validation (by native speakers):-* Collected data samples are manually annotated into one of three categories: positive (1), negative (0) and neutral

  (A).Each text sample was independently manually annotated by two different native Bangla speaking individuals for total two validations. Each annotator validated the data without knowing decisions made by other. This ensures that the validations are unbiased and personal.

TABLE I. DATASET VALIDATION SAMPLES

Text Sample	Translation	1st Annotator	2nd Annotator
অেনক ভাে◌লা হেয়েছ গান!	Very nice song!	Positive	Positive
মমাি◌ ক সড়ক দুঘটনায় ৩ জন ি◌নহত।	3 dead in a tragic road Accident.	Negative	Negative
Chotobelar modhur din gulo khub miss kori	Really miss the sweet childhood days	Positive	Negative
Symphony er set gula kemon?	How are Symphony mobile Sets?	Positive	Neutral
আেলা আেলা তিম কখেনা আমার হেবনা	Light, light, you'll never be Mine	Neutral	Negative

D. Double Validation Analysis

Table 2 gives shows the confusion matrix between the labels given by the two annotators. We can see that the annotators agreed on 75% of the texts samples, giving us a base-line of human level agreement for this data set. Not surprisingly, the greatest amount of disagreements arises on text samples which at least one of the annotators labelled as neutral.

TABLE II: CONFUSION MATRIX OF MANUAL ANNOTATIONS

Validation		Second Validation		
		Positive	Negative	Neutral
First	**Positive**	2817	538	392
	Negative	178	3864	404
	Neutral	27	95	1022

MODEL IMPLEMENTATION

Our Video consists of three categories of comments–

- Positive,
- Negative, and
- Neutral.

Depending on the Video used and number of categories classified, we used three types of fully connected neural networks layer, which mainly differ by the number of nodes in the output layer (Fig. 1). One and two output nodes were used for categorizing between positive and negative sentiments, and three output nodes were used when ambiguous labels were also included.

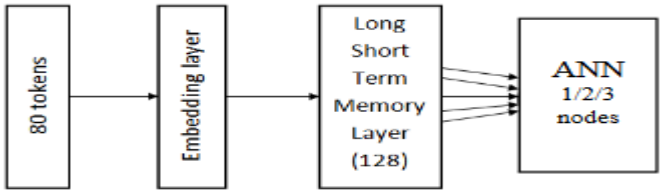

Fig. 1. Dense layer schematic

We used data for one validation set as pre -training for another validation set. More specifically, first we fit data from 1st validation in the model to pre-train for 2nd validation data which is fit in the same model afterwards. Likewise, we fit data from 2nd validation to pre-train for 1st validation data. This sort of pre-training was to check whether it can be useful to pre-train on an independently sentiment analysis data even if the labels did not match.

V EXPERIMENTS

A. Text blob

Text Blob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Text Blob stands on the giant shoulders of NLTK and pattern, and plays nicely with both.

Features:

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies

Tags used for different types of Bangla datasets –

Dataset Type	Tag Used in Experimental labels
Bangla and Romanized bangla(Total)	BRBT
Bangla (only)	Bangla
Romanized Bangla (only)	RB

Tags used based on loss function –

Loss function used	Tag Used in Experimental labels
Binary cross entropy	Bin
Categorical cross entropy	Cat

Tags used based on Annotation data modification-

Annotation data modification	Tag used in experimental labels
Annotation value of 'A' removed (label along with data removed)	Ra
Annotations value of 'A' converted to 2	A to 2

VII. RESULTS AND DISCUSSION

Highest accuracy was attained by Bangla YouTube Comment Scraping with googletrans API and YouTube API, Ambiguous removed and non-fixed max_features, with 85% of accuracy – which is 15% more than chance for two scraping. However, this experiment on BRBT dataset with categorical loss, URL Scraping, ambiguous converted to 2, has a low accuracy score of 15% but for a three category it scores 85% more than chance (90%). Therefore, it is clear that most of experiment done by real-time analytics. However, none of the experiments with fixed max_features (vocabulary size for Embedding layer) scored well compared to the non-fixed variants.

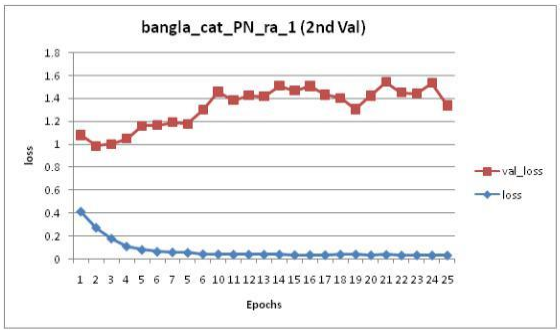

Fig. 2. loss-val_loss graph for bangla_cat_PN_ra_1 (2nd validation)

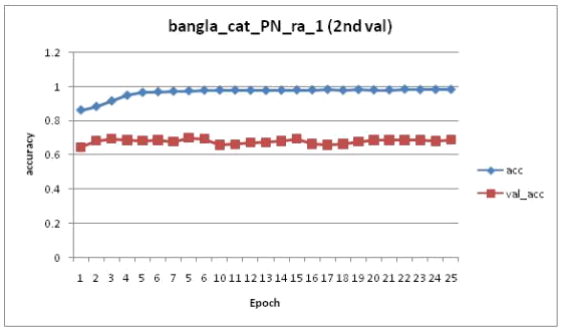

Fig. 3. acc-val_acc graph for bangla_cat_PN_ra_1 (2nd validation)

Fig 4: Scrapping YouTube video Comments.

VIII. Conclusion

To meet the goals, a BRBT (Bangla and Romanized Bangla Text), Scrape the YouTube video URLs and scroll down then select the YouTube video for scraping. The video text is stored in the Text Blob. The main advantage in the project is it accepts a multilingual language for scraping the YouTube videos using YouTube API and googletrans API. There are total 32 different experiments based on the same model with only differences in dataset used, loss function applied, modification done (or not) on data (proper noun replaced with <PN> tags, duplication removal etc.) etc. While most of the experiments scored accuracy higher than chance in percentage, Bangla dataset with categorical cross entropy as loss function and non-fixed max_features for the embedding layer with "Ambiguous removed" scored highest with 85% in accuracy for 3 categories (positive, Negative, Neutral comments of video), and Bangla and Romanized Bangla dataset (modified text set) with categorical cross entropy loss, non -fixed max_features, and "Ambiguous converted to 2" scored highest with 85% in accuracy for 3 category. The implementation of pre-training dataset of one label for another has showed that, even if the labels do not match it is useful to pre-train on an independently annotated SA data.

References

[1]A. Das and S. Bandyopadhyay, "Senti word net for bangla," Knowledge Sharing Event-4: Task, vol. 2, 2010.

[2]S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in bangla micro blog posts," in Informatics, Electronics &Vision (ICIEV), 2014 International Conference on. IEEE, 2014, pp. 1–6.

[3]M. S. Islam, M. A. Islam, M. A. Hossain, and J. J. Dye, "Supervised approach of sentimentality extraction from Bengali face book status," in Computer and Information Technology(ICCIT), 2016 19th International Conference on. IEEE, 2016,pp. 383–387.

[4]A. K. Paul and P. C. Shill, "Sentiment mining from bangla data using mutual information," in Electrical, Computer & Telecommunication Engineering (ICECTE), International Conference on. IEEE, 2016, pp. 1–4.

[5]R. Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM, vol. 56, no. 4, pp. 82–89,2013.

[6]T.Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space,"arXiv preprintarXiv: 1301.3781, 2013.

[7]Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv: 1408.5882, 2014.
[8]C. dos Santos and M. Gatti, "Deep convolutional neural net-works for sentiment analysis of short texts," in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69–78.
[9]Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolution neural networks for sentence classification," arXiv preprint arXiv: 1510.03820, 2015.
[10]H.Shirani-Mehr,"Applications of deep learning to sentiment analysis of movie reviews," in Technical Report, Stanford University, 2014.
[11]P. Ekman, "An argument for basic emotions,"Cognition &emotion, vol. 6, no. 3-4, pp. 169–200, 1992.
[12]P. K. Bhowmick, "Reader perspective emotion analysis in text through ensemble based multi-label classification framework," Computer and Information Science, vol. 2, no. 4, p. 64, 2009.
[13]V. K. Singh, "Sentiment analysis research on Bengali language texts," International Journal of Advanced Scientific Research Development (IJASRD), vol. 02, pp. 122–127, 2015.
[14]M. Al-Amin, M. S. Islam, and S. D. Uzzal, "Sentiment analysis of Bengali comments with word2vec and sentiment information of words," in Electrical, Computer and Communication Engineering (ECCE), International Conference on. IEEE, 2017,pp. 186–190.
[15]A. Hassan, M. R. Amin, N. Mohammed, and A. Azad, "Senti-ment analysis on bangla and Romanized bangla text (brbt) using deep recurrent models,"arXiv preprint arXiv:1610.00369, 2016.
[16]D. Das and S. Bandyopadhyay, "Developing Bengali wordnet affect for analyzing emotion," in International Conference on the Computer Processing of Oriental Languages, 2010, pp. 35–40.
[17]R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," Current biology, vol. 24, no. 2, pp. 187–192, 2014.
[18]T. Pranckevičius and V. Marcinkevičius, "Comparison of naïve Bayes, random forest, decision tree, support vector machines ,and logistic regression classifiers for text reviews classification," Baltic Journal of Modern Computing, vol. 5, no. 2, p. 221, 20