Proposal of

Using Bangla Text Data to Analyze Emotions

By

ARINDAM DEY (18-38458-2)

UMMA KHADIZA (18-38311-2)

TAHMEED MAHBUB (18-38743-3)

Introduction:

Sentiment analysis is the technique of identifying the emotional biases hidden behind a succession of words, which is used to gain a better understanding of the attitudes, opinions, and feelings represented in an online mention. Sentiment analysis, in general, tries to determine a speaker's, or other subject's attitude toward a topic, as well as the overall contextual polarity or emotional reaction to a document [1]. Sentiments are, by their very nature, subjective. Opinion and sentimental mining are key study areas since gathering people's opinions is difficult owing to the large amount of daily posts on social media [2]. The attitude of a text can be interpreted differently by various people. Bangla has the second-largest number of speakers and ranks sixth among the world's most widely spoken languages [3]. Nearly 200 million people speak Bangla as their primary language, 160 million of them are Bangladeshis [4]. People have been expressing their ideas on many topics through social networking websites such as Facebook, Twitter, and others, frequently in their own native language, during the last decade as the use of social media has grown [5]. About 90 percent of today’s data has been provided during the last two years and getting insight into this large scale data is not trivial [6]. Due to little or no human-to-human connection in such firms, it is becoming increasingly difficult for them to monitor and evaluate market trends, particularly when doing so by monitoring consumer reactions to their products or services. [7]. Affect analysis is a natural language processing (NLP) technique for recognizing the emotive aspect of text. The same textual content can be presented with different emotional slants [8]. The main goal of this sentiment analysis technique is to determine if the conveyed mood of a text document in Bangla is happy/sad/angry/tender/excited/scared when it comes to training data. Because each of these emotion classes may reflect a variety of feelings, we utilize them as our emotion category. For instance, the words "joy," "smile," "optimistic," "laugh," and "pleasure" are all examples [9]. Sentiment analysis can be applied at different level of scope.

Motivation:

Sentiment analysis is an essential component of natural language processing. In general, sentiment classification refers to the study of a speaker's expression to identify whether he or her has a favorable or negative view on a certain topic. With the fast expansion of e-commerce, sentiment analysis has the potential to have a significant impact on everyone's daily lives. For example, online product reviews have become a significant source of information for people making purchasing decisions. Because there are frequently too many reviews for consumers to read, figuring out how to automatically classify and infer emotion from them has become a major research topic. We offer a Sentiment Analyzer in this proposal that recognizes Bangla sentiment or opinion on a subject from Bangla text. We create certain phrase patterns and determine the sentiment orientation of those patterns. Because individuals spend hours everyday on social media and express their opinions, social networks are the primary sources of information on people's opinions and sentiments on many issues.The Research is more important for

1. The presentation of a carefully annotated Bangla emotion corpus that captures the diversity of finegrained emotion expressions seen in social-media content. For employing classical machinelearning approaches that typically perform well in classifying the six aforementioned emotion types.
2. To determine the best-performing model for emotion classification by comparing machine-learning classier performance to a baseline.
3. Pre-processing the data in a way so that it is readily usable by researchers.
4. Application of deep recurrent models on a Bangla and Romanized Bangla text corpus.

RESEARCH QUESTIONS:

• What sentiment analysis to analyze and who are involve?

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

• How can we identify customers' sentiment (positive or negative) from their opinions?

We will use some methods which are described in the methodology section.

• How accurately can we capture that sentiment?

The Accuracy will be like average 90% and we gave a clear statement on methodology section.

Objective:

Our project idea, as we all know, is based on emotion analysis from Bangla text data. Instead of separating the corpus only on the basis of positive and negative feelings, Emotion Analysis is employed. We identify the sentiment of a phrase or paragraph from Bangla text using the valiancy of a word in this proposal. We attempted to think up more specific emotion descriptors like sorrow, pleasure, contempt, surprise, fear, and rage, which are all based on human feeling. So, there are six fundamental emotion types in all. We use a model, formulae, and algorithms to determine a word's previous valence. There are six kinds of values that correlate to each word's sense: sorrow, happiness, disgust, surprise, fear, and rage. The senses change depending on the components of speech in the related phrase.

We can take Bangla words from different social media also. The social media can be Newspaper, Facebook, Twitter and so on. Our main focus will be,

1. For our project, we've collected and presented Bangla datasets, which we've made public.
2. We performed statistical linguistic analysis on the datasets.
3. For the gathered datasets, we used state-of-the-art machine learning techniques and found reasonable accuracies.

Scope: This proposal proposes a process of sentiment analysis of text data which is written in Bangla language. This process can automate the analysis of user’s reaction towards a specific emotion like movie or TV show any social media. With more and more people expressing their opinions openly in the social networking sites, analyzing the sentiment of comments made about a specific text, movie and advertisement indicate how they feel. There are a lot of scope of the project proposal.

1. In this project we propose the detection of the six basic emotions namely anger, disgust, fear, joy, sadness and surprise for Bangla text.
2. Supervised and unsupervised learning methodology can easily be applied on Bangla text for emotion analysis or opinion mining for specific domain of text.Using Support Vector Machine algorithm, different models and algorithms, those contents will be cover partly.

Design of Case Study:

Some of our work is inspired by past work in these disciplines, while others are for the purpose of expanding our understanding. Our concept was largely motivated by numerous articles that attempted to categorize various groups, such as aggression and emotions. And it's one of the important works that we've been following. They are attempting to extract emotions from SMS texts. For highly resourced languages like English, Arabic, and other European languages, sentiment analysis or detection from text is a well-studied research subject [10]. Some of them employed TF-IDF to improve classification accuracy and Support Vector Machine to classify [11]. Their method is very similar to the one we used in our initial study [12]. They did, however, adopt the Vector Space Paradigm (VSM) as a document representation model. We were introduced to a huge amount of digital data (mostly opinionated texts e.g. statuses, comments, arguments, etc.) like never before with the advent of social media on the internet e.g. Facebook, Twitter, forum discussions, reviews, and its rapid growth, and to deal with this huge data the SA field experienced a similar growth. Sentiment analysis has been one of the most active study fields in NLP since the early 2000s [13]. As far as we know, some work has been done utilizing valiancy analysis to detect sentiment or mine opinions from Bangla text. There are some works on valiancy analysis of Bangla verbs [14] [15]. There are, nevertheless, several works that deal with sentiment detection in Bangla writings. [16]proposed an automatic sentiment recognition technique for Bangla text using a machine learning model. They also uncover some significant obstacles in the processing of Bangla language. [17]used Santi Word Net [18] and WordNet Affect [19] to propose an emotion tracking system based on topic or event by utilizing sense basis affect scoring algorithms for annotated news items and blog corpora in Bangla. From their dataset, they obtained the maximum emotion accuracy of 94.5 percent. The most recent study [19] for Bengali language using social communications data to discover the polarity of a Bengali text if it is whatever emotion has recently been published. They used the Bangla Post-Tagger Package for POS Tagging, as well as the Support Vector Machine and Maximum Entropy algorithms, to compare the performance of these two methods by testing with different sets of characteristics. A similar collection was carried out in [18], in which 1400 Bangla Tweets were automatically collected. However, their dataset is not publicly available, and it is quite tiny in size. Speech recognition, handwriting recognition, natural language processing, and other applications have all benefited from the usage of recurrent neural networks, or RNNs. [10]. Taher et al. found that SVM had the best precision and recall of 0.86 in their testing. N-gram Based Sentiment Mining for Bangla Text Using Support Vector Machine [13] Negativity Separation, which isolates a word's negative postfix from the actual word, putting more emphasis on the fact that the overall sentence contains negativity, was an innovative strategy utilized in this work. Another significant research [17] used opinion mining and mood extraction, in which the polarity of text was evaluated as positive, negative, or neutral. [11] has conducted a survey on sentiment analysis and opinion mining by evaluating text classification. In addition, [14] conducted a survey of sentiment analysis text data. [13] looked at the emotions expressed in English and Bengali texts and tracked them. We studied Bengali Language Processing through these literatures.

METHODOLOGY

Methodology can help in the development of a unique blueprint for the proposed solution. Emotion analysis on Bangla text data was recommended in this study. We're aiming to evaluate our work based on computation speed and precision. To determine accuracy, a precision and recall factor will be utilized. Six distinct techniques were used in this study.

Methods

1) The ratio of parts of speech.
2) TF-IDF for cosine similarity.
3) Using a custom TF-IDF, compute cosine similarity.
4) Uni-gram and stammer Nave Bayes model
5) A Nave Bayes model with a Bi-gram stammer and a normalizer is used.

Experiments of the Methods

A. Parts of Speech Ratio

The positive and negative datasets are used as classifiers in this method. The positive and negative datasets are first labeled using a proprietary POS tagger. After that, both classifiers' nouns, adjectives, verbs, pronouns, and conjunctions are counted. For both positive and negative classifiers, the POS ratio is calculated. When a query is received, it is POS tagged and the POS ratio is calculated. After that, the hamming distance between the positive and negative classifiers and the query is determined. The classifier is then defined by the minimum distance.

Fig. 1. Parts of speech ratio model.

B. Cosine Similarity Using TF-IDF

The cosine similarity between documents is computed here. The positive dataset is labeled D0C1, the negative dataset is labeled D0C2, and the query dataset is labeled D0C3. The main goal is to discover the highest degree of similarity between D0C1 and D0C3, as well as D0C2 and D0C3. Three vectors are created to find the cosine similarity: D0C1 as PV (Positive Vector), D0C2 as NV (Negative Vector), and D0C3 as QV (Query Vector).

Term Frequency (TF): All of the papers' unique words are enumerated, and their frequencies are counted. The number of documents that contain the unique words (document frequency, or DF).

The vectors are calculated when the TF, DF, and IDF for both the documents and the query have been obtained.

C. Cosine similarity using custom TF-IDF

Because of the many parts of speech in Bangla, the same words might have multiple meanings. As an example, ‘তারা’(star) can be used as Noun. Again, ‘তারা’(they) can be used as Pronoun.

Here, the Bangla word ‘তারা’ has different meanings when it’s used in different sentences. Like,

“তারা ক্রিকেট খেলকে”

“আোকের তারা খে◌কেত অকে◌ সুন্দর”

A custom POS tagger is employed to solve this problem, which recognizes the portions of speech in a text.

The words “তারা” (star) (Noun) and “তারা” (they) (Pronoun) are treated differently while looking for the TF and DF. The technique is then repeated as in the preceding section. **Cost Estimate:**

Sl. No.	Name of Task	Quantity	Unit price (BDT)	Total price (BDT)	Remarks
	Data Collection	NA	NA	2000	Can be increase
	Experiments	5	1000	5000	
	Implimentaion	NA	1000	1000	
	Analysis	5	500	2500	
	Final Exicution	NA	2000	2000	

Work Schedule:

Task Name	20/12/20	30/12/20	9/1/21	19/1/21	28/1/20	9/2/21	19/2/20	28/2/20
Proposasl Submission	🟩	🟩						
Project objective		🟩						
Deliverable and due dates		🟩	🟩					
Rules and responsiblities		🟩	🟩					
Tracking and Exicution tools			🟩	🟩				
Simulation				🟩	🟩			
Coding					🟩	🟩		
Progress Report						🟩	🟩	
Impliment and analysis						🟩	🟩	
Writing and presentation							🟩	
Final Report								🟩

Conclusion:

The study's limitations are those aspects of the design or methodology that impacted or influenced the interpretation of the findings. Sentiment, like all views, is fundamentally subjective and might even be illogical. When attempting to quantify sentiment, it's important to use a broad — and relevant — sample of data. There is no such thing as a point of data that isn't relevant. It's the total that counts. One or more indirect reasons may impact an individual's emotion toward a brand or product; for example, someone may be having a terrible day and tweet a critical remark about something they otherwise had a fairly neutral view about.With a large enough sample, outliers are diluted in the aggregate. Also, since sentiment very likely changes over time according to a person’s mood, world events, and so forth, it’s usually important to look at data from the standpoint of time.”

A quasi-experiment was undertaken to validate the applicability of emotion recognition in an e-learning setting. It was based on a standard online lesson for utilizing a software application that included user emotion recognition channels monitoring. Today's sentiment analysis is quite precise. However, subtleties like irony, comedy, or sarcasm are difficult to detect with a simple sentiment analysis.

References

[1] M. Rahman, "Identifying and Categorizing Opinions Expressed in Bangla Sentences using Deep Learning Technique," *International Journal of Computer Applications* , vol. 176 , no. 0975 – 8887, pp. 13-17, 17, April 2020.

[2] "Survey on Text-Based Sentiment Analysis of Bengali Language," *1st International Conference on Advances in Science, Engineering and Robotics Technology* , vol. 978, no. 7281, pp. 1-6, 2019.

[3] M. S. H. Rumman Rashid Chowdhury*, "Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques," *International Conference on Bangla Speech and Language Processing,* vol. 7281, no. 978, pp. 22-27, 2019.

[4] M. R. A. Asif Hassan, "Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models," *International Conference,* vol. 567, no. 192, pp. 12-17, 2014.

[5] M. H. R. Chowdhury, "“Bangla handwritten character recognition using convolutional neural network with data augmentation," *International Conference on Bangla Speech and Language Processing,* vol. 376, no. 1998, p. 04, 2019.

[6] "Sentiment Detection from Bangla Text using Contextual Valency Analysis," *K. M. Azharul Hasan,* vol. 668, no. 12, pp. 23-24, 2014.

[7] S. C. a. P. Bhattacharyya, "Valency Analyzer of Verb Arguments for Bangla” proceeding," *International Conference on Bangla Speech and Language Processing,* vol. 766, no. 345, pp. 13-17, 2011.

[8] C. Strapparava, "The affective weight of the lexicon," *International Conference on Language Resources,* vol. 4481, no. 474, pp. 34-38, 2006.

[9] k. khan, "Data analysis," *International Conference on Computer and Communication Systems,* vol. 998, no. 123, pp. 7982, 2019.

[10] H. V. L. T. P. Le, "Aspect analysis for opinion mining of vietnamese text," *international conference on advanced computing and applications (ACOMP),* vol. 1, no. 1109, pp. 118-123, 2015.

[11] J. D. S. a. P. S. Haddela, "A term weighting method for identifying emotions from text content," *4th International Conference on Electrical Information and Communication Technology (EICT),* vol. 978, no. 7281, p. 381–386, 2019.

[12] M. K. B. G. S. R. A. R. Mehra, "Sentimental analysis using fuzzy and naive bayes.," *Computing Methodologies and Communication (ICCMC), 2017 International Conference* , vol. 168, no. 991, p. 945–950, 2017.

[13] B. L. L. a. S. V. Pang, "sentiment classification using machine learning techniques.," *the ACL-02 conference on Empirical methods in natural,* vol. 10, no. 654, pp. 217-220, 2002.

[14] S. C. a. P. Bhattacharyya, "Valency Analyzer of Verb Arguments for Bangla," *International Conference on Bangla Speech and Language Processing(ICBSLP,* vol. 78, no. 4799, pp. 278-282, 2011.

[15] B. a. L. L. Pang, "Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval," *International Journal of Information Engineering and Electronic Business(IJIEEB),* Vols. 2(1-2), no. 135, p. 1–135, 2008.

[16] M. S. I. K. M. Azharul Hasan, "Sentiment Recognition from Bangla Text," *17th Int'l Conf. on Computer and Information Technology,* vol. 3, no. 6, pp. 316-327, 2013.

[17] A. &. B. S. Das, "Phrase level polarity identification for Bengali," *International Journal of Computational Linguistics and Applications*, vol. 1, no. 2, p. 169–181, 2010.

[18] A. a. S. Esuli, "SENTIWORDNET: A publicly available lexical resource for opinion mining," *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* vol. 06, no. 5, pp. 12-25, 2006.

[19] C. a. V. A. Strapparava, "Wordnetaffect: An affective extension of WordNet," *International Conference on Bangla Speech and Language Processing(ICBSLP),* vol. 2, no. 4, pp. 1083-1086, 2004.