2020 23rd International Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020.

Personality Traits Detection in Bangla: A Benchmark Dataset with Comparative Performance Analysis of State-of-the-Art Methods

Utpal Rudra, Abu Nowshed Chy, and Md. Hanif Seddiqui
Department of Computer Science & Engineering
University of Chittagong, Chattogram-4331, Bangladesh
utpal.cse@std.cu.ac.bd, nowshed@cu.ac.bd, and hanif@cu.ac.bd

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392722

***Abstract*—Nowadays, people are interested in using various online platforms to share their opinions and thoughts on various topics and issues. The informal user-generated contents in these platforms make it an important source for studying and modeling the personality of a person. Detecting and analyzing user personality plays an important role to design an effective recommendation system, Q/A system for customer care, employee assessment, and product promotions. Prior works on personality detection from user-generated text mostly conducted on the English language. However, there is no previous work and dataset available for automatic detection of user personality from Bangla text. In this paper, we bridge this research gap and present a benchmark Bangla personality traits detection dataset that consists of 3000 Bangla informal text collected from various online platforms. Besides, we present various baseline systems by exploiting state-of-the-art supervised classification methods and perform a comparative performance analysis that provides an important insight about this task. We believe this dataset might be beneficial for others for developing more effective models and we publicly release the dataset for future research purposes at the following link: https://git.io/JkW6V or use the expanded URL[1].**

***Index Terms*—personality traits detection, Bangla, benchmark dataset, supervised classification, deep learning**

I. Introduction

The increasing number of online platforms including social media, image sharing sites, video sharing sites, blogs, and news portals make it easier for people to raise their voices and express their opinion, emotion, and thoughts towards various topics. The topics are ranging from various issues, entities, individual's matter, political agenda, Govt. declarations, and other shareable matters through online media. Since the user opinions are greatly impacted by the user's inherent personality, therefore the opinions that are shared to the online platforms might be an important source to determine the personality of a user.

Personality is one of the most complex attributes possessed by a human being that exposes the uniqueness of a person. Usually, personality traits are defined as descriptions of people in terms of relatively stable patterns of behaviors, thoughts, and emotions [1]. The personality of a person varies individual to individual. Different things including a successful position on the job, having something special in life, and gaining positive or negative reviews in the socialistic works obviously represent someone's personality in society. Knowledge about an individual's personality allows us to make a decision about his/her preferences across various contexts and environments. Distilling such information is beneficial to enhance the performance of personalized recommendation systems, automatic Q/A system design, and product promotions [2]. Besides in computer-assisted tutoring systems, ubiquitous computing and forensics, employee recruitment, and career counseling, personality can be used as an important assessment [3], [4].

The personality of a person can be determined via a self-descriptive report, interview, or observation that is conducted by a psychologist. But these traditional ways are less practical, less scalable, and time-consuming. Besides some websites try to determine the user's personality to understand their preferences through a mandatory questionnaire test [5], where the person has to answer various questions. The limitations of these traditional approaches incur the researchers to design automatic personality traits detection system from the informal user-generated text in online platforms.

Most of the prior methods leveraged various lexicons, linguistic, psycholinguistic, and emotional features in a supervised learning framework to determine the personality of a user from the text. The used learning models are ranging from traditional SVM, KNN, and naive Bayes to the trendy deep learning based approaches including CNN and LSTM. The methods are mostly developed for the English language and addressed the problems in the view of English linguistics. To the best of our knowledge, there is no previous work and dataset available in Bangla that addressed the challenges of personality detection from Bangla text. Since Bangla is a low resource language and highly inflectional [6], therefore it is inevitable to bridge this research gap and create resources for future researches on this task.

The main contributions of this paper are two-fold: (1) We introduce a dataset for personality traits detection in Bangla language. To the best of our knowledge, this is the first dataset in Bangla for this problem domain and we release the dataset publicly for research purposes. (2) We present several baseline systems that build upon several statistical and deep learning based models. We also present the comparative performance analysis of these systems, which provides promising directions for future research.

[1]https://github.com/nowshedcu/Personality-Traits-Detection-in-Bangla

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

The rest of the paper is organized as follows: Recent works that are relevant to personality traits detection described in Section II. We introduce our Bangla personality traits detection dataset in Section III. In Section IV, we present the framework of several state-of-the-art baseline systems, whereas the comparative performance analysis of these systems presented in Section V. Finally, concluding remarks with some future directions of our work are presented in Section VI.

II. Related Work

With the emerging trends of using various online platforms for communication and social networking, researchers are interested in utilizing these resources to determine the user personality. Most of the previous studies formulate the automatic personality detection problem based on the Big Five model, where human personality is defined based on five basic dimensions including agreeableness, conscientiousness, extroversion, neuroticism, and openness [7].

Early work on personality traits detection usually focused on the user-generated essay and explored various linguistic features and lexicons in a supervised learning framework [8], [9]. Pennebaker and King [8] collected written essays from a number of volunteers within a controlled environment and then the respective authors of the essay were asked to define their personality based on the Big Five model. Later, they used linguistic inquiry and word count (LIWC) features to determine the correlation between the author's personality and essay. Das et al. [9] presented an n-gram lexicon based on the probability of occurrences and explored the LIWC lexicon for personality traits identification from the essay dataset. Recently, Navonil et al. [10] used the convolutional neural network (CNN) for detecting personality traits from plain text. They have used 2467 essays that were tagged with author's personality traits and contained 1.9 million running words.

Besides, blogs are also become a rich repository for retrieving data and these are used to predict traits. Mohtasseb et al. [11] mined online data from different blogs to identify the bloggers' personality. Data from different websites where one can find their life partners [12] or human conversation [13] has also analyzed to get the human personality. Iacobelli et al. [14] used a large blog corpus and extracted various features based on textual, structured, and dictionary based to analyze the impact of feature selection.

Most recently, microblogging platforms e.g. Twitter gaining popularity among the online users due to its efficacy to satisfy real-time information needs. Researchers proposed various methods to identify personality traits from short texts e.g. tweets. Pratama et al. [3] explored various machine learning techniques for personality detection on their MyPersonality dataset which is comprised of a large set of user statuses and was perfectly labeled with Big Five dimensions. Liu et al. [15] explored the deep semantic features and atomic features for tweet representation to improve the performance of personality prediction. In this direction, contents of the social news website Reddit [16] and Q/A website StackOverflow [17] also explored to identify the personality of the user.

To address the personality detection challenges on other languages but English, various attempts are observed recently where researchers focused on developing datasets for their native language. In some of these works, Ramos et al. [4] presented a corpus for the Brazilian Portuguese language. Tighe and Cheng [18] address the problem based on the tweets of Filipino users and Liu et al. [19] utilized the Spanish and Italian tweets besides the English for their compositional model of personality detection.

However, in our target language Bangla, previous studies as well as resources are not available to address the challenges of automatic personality traits detection from informal user-generated texts. This observation motivates us to develop a benchmark dataset and analyze the performance of various baseline systems in this problem domain.

III. A Dataset for Bangla Personality Traits Detection Task

Now, we describe our data collection and annotation procedure for the personality traits detection task in Bangla.

A. Data Collection

Most of the previous studies on other native languages including English [15], Indonesian [3], Tagalog (national language of Philippine) [18], and Spanish and Italian [19] harnessed the popular microblogging platform twitter to collect the informal user-generated text. In contrast, Facebook is more popular to the Bangla language speaking people compared to twitter and also people are highly interested in sharing their opinion through the comment threads of online news sites and blogs. Therefore, considering these observations and maintaining the diversity of the data, we have collected a set of user generated informal Bangla text from various online platforms including Facebook[2] statuses and comments, Youtube[3] comments, and comments from Somewhereinblog[4], Roar.media[5], and shafaetsplanet[6] blogs. The texts are usually the publicly shared user statuses and comments. To collect the related text, we empirically used some keywords according to the Big Five model that best describe a user personality. We consider both the mono-sentential and multi-sentential text. Since, informal user-generated Bangla text may contain other languages including English, Hindi, and Banglish (Bengali written with English characters), we perform the data curation process using a language detector [20] and keep only those texts that contain only Bangla characters. After that, we distribute the dataset to the three volunteers to identify the important Bangla text and discard the noisy and unimportant text. In this phase, they are not trained with any concrete knowledge of personality traits, rather we asked them to decide whether a given text contains the user opinion or not and the intensity of the opinionated dimension. We just rely on their

[2]https://www.facebook.com/
[3]https://www.youtube.com/
[4]https://www.somewhereinblog.net/
[5]https://roar.media/bangla
[6]http://www.shafaetsplanet.com/

Agreeableness	Conscientiousness	Extroversion	Neuroticism	Openness
স্যার, আপনার হাসিটা অসাধারণ। এত সুন্দর করে কেউ হাসে কি করে?	প্রত্যেক টা বাবা যদি জনক না হয়ে এমনি পিতা হত, তাহলে পৃথিবীটাই পরিবর্তন হয়ে যেত।	সবসময় সুখে দুখে আর্জেন্টিনার পাশে ছিলাম, আছি এবং থাকব। আর্জেন্টিনার জন্য সবার নিকট দোয়া প্রত্যাশী।	পরিকল্পনা মন্ত্রী ই সবচেয়ে বেশি অপরিকল্পিত কাজ করে।	প্রস্তুতি ম্যাচে জার্মানিকে গতবারের তুলনায় অনেকটাই দূর্বল লাগছে।ব্রাজিল আর স্পেনের মধ্যে কেউ চ্যাম্পিয়ন হবে।
মানুষ তার কর্মের মাঝে বেচে থাকে, উনিও থাকবেন জীবিত বাঙালির মনে। বিনম্র শ্রদ্ধা আপনাকে।	পরিবহন মালিকদের বিরুদ্ধে ব্যবস্থা নিতে হবে, তবেই সড়কের অনেক ভোগান্তি থেকে যাত্রীরা রক্ষা পাবে।	আপনার মতো আমারও কিছু ইচ্ছে আছে কিন্তু জগৎ-সংসার এর যাতাকলে পড়ে এই ইচ্ছেগুলো ইচ্ছেই রয়ে গেল।	মানবিক গুনাবলি বিহীন মানুষ। লজ্জা, আর ঘৃণা ছাড়া আর কিছুই তাদের জন্য নয়। জীবনেও এরা মানুষ হবে না।	জীবনটা সেই মানুষের সাথে কাটানো উচিৎ, যার চেহারার চেয়ে মনটা অনেক বেশি সুন্দর।

Fig. 1: Example of sample Bangla texts for various personality classes.

Category	#Total	#Train	#Test
Agreeableness	511	408	103
Conscientiousness	528	422	106
Extroversion	628	502	126
Neuroticism	646	516	130
Openness	687	549	138
Total Instances	3000	2397	603
Other Statistics			
-Total Number of Unique Words: 14427 -Average Number of Words Per Document: 16.35 -Average Number of Characters Per Document: 95.05			

TABLE I: Distribution of samples in the training and test set across different categories.

intuition to select the user opinionated text. After that, the collected data are sent to the annotation phase.

B. Human Annotation

After completing the data collection procedure, we initiate the data annotation procedure. In this regard, we consider the Big Five model [7], which is the most used taxonomy in prior studies [3], [15] of personality traits detection task. According to the Big Five model, the five user personality traits are defined as follows:

Agreeableness: A person who has such types of personality is kind, affectionate, sympathetic, and helping nature. They agree with people easily as they have a little menial nature.

Conscientiousness: This class label is reserved for self-disciplined, organized, efficient, and reliable people. They have a logical mindset with an aim for achieving what they want.

Extroversion: This category applies to the persons who are talkative, assertive, and attention-seeking. They always want to express themselves in front of people.

Neuroticism: This class contains the personality of people who are self-pitying, anxious, depressed, and have negative emotions in their activity.

Openness: This category represents the curious, artistic, imaginative person. Persons who possess this kind of personality are also appreciated by arts, emotions, and adventurous activity.

Following the definition of these personality traits class, we have prepared some curated samples of each category. Then, the three volunteers are trained with the definition of each category with examples. The volunteers are then asked to select the most representative category for each given text among the five personality traits classes including Agreeableness, Conscientiousness, Extroversion, Neuroticism, and Openness. After completing the annotation scheme, we select those texts where all three volunteers select the same label for the corresponding texts and discard the other texts. To ensure the quality of annotation, we randomly select some annotated sample asked the volunteer again to categorize them and asked them about the explanation of their selection. To illustrate the quality of our annotation as well as demonstrate the Bangla text sample for each of the five personality traits category, we present two samples for each category in Figure 1.

Finally, we randomly select the 3000 Bangla user-generated text with the personality traits label from the annotated corpus. The number of unique Bangla words available in our dataset is 14427. The average number of words and characters available in each document is $\approx$16 and $\approx$95, respectively. We randomly shuffle the data of each category and split the data into train and test part according to the 80:20 ratio. The details statistic of the dataset is shown in Table I.

C. Task Formulation

The major goal of the personality traits detection task is to analyze the content of the informal Bangla texts and categorize them into the most representative personality traits class. Therefore, we cast our personality traits detection as a multi-class classification problem and labels each Bangla text to the corresponding personality class. The task is formally defined as follows:

Task: Classifying Bangla text by the user's personality traits.

Given a user-generated Bangla text, a system needs to determine the most representative personality class for this text. Standard evaluation measures including macro averages recall, precision, F1 scores, and accuracy is used to estimate the overall system performances. We consider the F1 score, macro averages across all five personality classes as the primary evaluation measure for this task.

IV. Baseline Methods

In this section, we explore various state-of-the-art baseline methods on our Bangla personality traits detection dataset. The overview diagram of our baseline framework is depicted in Figure 2.

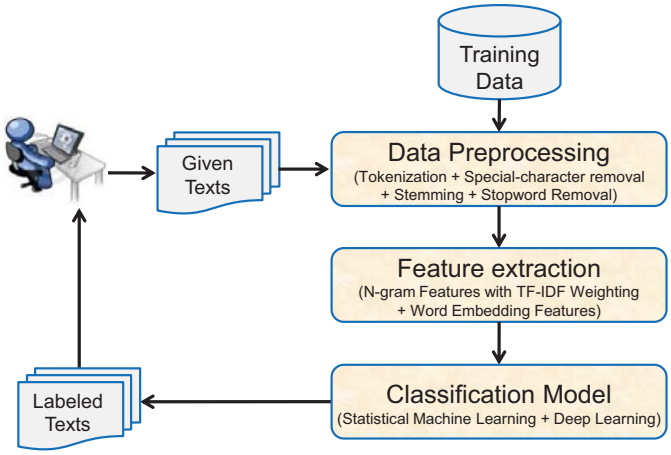

Fig. 2: Overview of our baseline methods.

At first, the baseline system fetches Bangla text from the user and initiates the data preprocessing stage. In the preprocessing stage, we employ various text processing techniques including tokenization, special-character removal, stemming, and stopword removal to reduce the noise and make it suitable for feature extraction. To transform the text documents into feature representation, we extract the n-gram features with the TF-IDF weighting scheme and word embedding features from pretrained embedding models. Based on the extracted features, a supervised classifier is employed to classify the texts into the corresponding personality traits label. Finally, the labeled documents are returned to the user.

A. Data Preprocessing

We initiate the preprocessing stage with tokenization. We use the bnlp_toolkit[7] to tokenize the Bangla text. We also use a publicly available Bangla stemmer[8] for stemming each token or word, so that we can get the canonical form of the word. Since stopwords have a negative impact on classification performance, we remove it from the text. For stopword removal, we use a publicly available Bangla stopword list[9]. Besides, we remove all the special characters and single letter words from the text.

B. Feature Extraction

Feature extraction is the process of modeling a document for fitting it to the learning algorithm. For feature extraction, we follow the bag-of-words (BoW) representation of the documents which describe a document based on the word occurrence statistics of that document. After BoW representation of the documents, we extract the n-gram features, where the composition of term frequency (TF) and inverse documents frequency (IDF) is used as the weighting scheme.

Term frequency (TF) denotes the number of occurrences of a word in a document. Since words that are highly occurred in a specific document usually represent the context of the document, the high TF score indicates the high importance of that word for this document [21]. However, some words occurred highly in every documents due to the sentence structure of Bangla. To reduce the effect of these words in feature weighting, we use composition of TF and IDF i.e. TF-IDF as a feature weight.

The inverse document frequency (IDF) estimate the generic importance of a word across a set of documents or corpus. It is the logarithmic ratio of the total number of documents to the number of documents containing the word [21]. Formally, it is defined as follows:

$$IDF_i = \log \frac{|D|}{1 + |W : w_i \in W|}$$

where $|D|$ is the total number of documents in the corpus and $|W : w_i \in W|$ is the number of documents where the word appears.

Besides, for the deep learning based models we use the distributed representation of words (known as word embedding or word2vec) as features. The representations are learned from the word co-occurrence statistics of a large corpora and help the learning models to obtain better performance.

C. Classification Model

Once completing the feature extraction stage, we train a set of supervised learning models broadly grouped into two types including statistical machine learning based models and deep learning based models.

1) Statistical Machine Learning Models: Statistical machine learning models usually exploit a set of mathematical functions to learn the nature of the data towards a target class. In our framework, we use the naive Bayes (NB) [22] probabilistic classifier which uses the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of this model is the assumption of word independence that makes it far more efficient than the other approaches. Besides, we use the stochastic gradient descent (SGD) and support vector machine (SVM) [22] models that can deal with both the linear and non linear characteristics of the data with a great efficiency. We also explore two tree based classifiers including decision tree (DT) and random forest (RF) for personality traits identification [22].

2) Deep Learning Models: Deep learning models are composed of a series of algorithms that mimic the learning process of human brain. These models can find out the underlying relationship in a dataset and handle non-linear characteristics of the data very well. To achieve this, we exploit the fully-connected multiple hidden layers based model, multilayer perceptron. We also explore the two major deep learning architectures including FastText [23] and convolutional LSTM (C-LSTM) [24] model.

[7]https://pypi.org/project/bnlp-toolkit/
[8]https://pypi.org/project/bangla-stemmer/
[9]https://www.ranks.nl/stopwords/bengali

TABLE II: Performance of various statistical and deep learning based baseline methods on Bangla personality traits detection.

Method	Overall (Macro Average)				Catergory-wise F1 Score				
	Recall	*Precision*	*F1 Score*	*Accuracy*	Agreeableness (AGR)	Conscientiousness (CNS)	Extroversion (EXT)	Neuroticism (NRT)	Openness (OPN)
Statistical Machine Learning Methods									
Decision Tree	0.23	0.27	0.22	0.24	0.19	0.20	0.23	0.20	0.30
SGD	0.27	0.26	0.26	0.26	0.30	0.24	0.33	0.26	0.18
Random Forest	0.29	0.29	0.29	0.29	0.33	0.19	0.36	0.30	0.25
MultinomialNB	0.30	**0.36**	0.30	0.31	0.31	0.20	0.41	0.31	0.29
SVC	**0.32**	0.31	**0.31**	**0.32**	0.36	0.28	0.42	0.28	0.21
Deep Learning based Methods									
MLP	0.27	0.27	0.27	0.27	0.31	0.25	0.37	0.25	0.16
FastText [23]	**0.36**	**0.37**	0.36	**0.37**	0.35	0.29	**0.45**	**0.38**	0.34
C-LSTM [24]	**0.36**	**0.37**	**0.37**	**0.37**	**0.41**	**0.31**	0.40	0.35	**0.36**

V. Experiments and Evaluation

In this section, we evaluate the performance of various state-of-the-art baseline methods on our Bangla personality traits detection dataset.

A. Experimental Setup

We develop our proposed baseline models using the Google Colaboratory [25] platform. We utilized the Scikit-learn [22] implementation of the SVM (SVC), multinomial naive Bayes (MultinomialNB), SGD, decision tree, and random forest classifiers. We performed the 5 fold stratified cross validation on the training set to select the optimal hyperparameters. In our experiments, we consider the word n-grams (1- and 2-grams) features with the TF-IDF weighting scheme. The n-gram features are used to train all the classifiers but FastText and C-LSTM. For these two models, we used the word embedding features extracted from 300-dimensional fastText embedding model pre-trained on common crawl and Wikipedia [26]. We used the default MLP settings (single hidden layer with 100 neurons) from the scikit-learn. The C-LSTM network is composed of a single layer CNN module on top of the single layer LSTM module according to the settings of [24]. The CNN kernel size is set to 2, batch size is set to 32, and number of epochs is set to 20. We use the softmax activation and adam optimizer. We trained the FastText classifier with 2-gram based features from pretrained fastText embeddings [26]. We used the default settings for the other parameters.

B. Results and Analysis

We now evaluate the comparative performance of our proposed baseline methods. According to the benchmark defined in Section III, we consider the macro average F1 score as the primary evaluation measure to compare the system performances. We present the comparative performance of various baseline methods on Bangla personality traits detection dataset in Table II.

The results in Table II show that SVC obtained the best performance among the statistical machine learning methods and C-LSTM obtained the best performance among the deep learning methods in terms of F1 score. It also shows that deep learning methods that used the features from pretrained embedding models including FastText and C-LSTM, outperform the traditional BoW based approach by a large margin ($\approx 6\%$). This is because our dataset is based on the informal user generated contents which contains the rare and noisy words incessantly. Therefore, BoW based approaches suffer the severe vocabulary mismatch problem and cannot address the polysemy problems. Besides, BoW considers all words of a document are independent, therefore failed to capture the word order information. We also report the category-wise performance of each methods. It showed that C-LSTM obtained the best performance for the *Agreeableness*, *Conscientiousness*, and *Openness* category, where FastText obtained the best performance for the other two categories.

To investigate the performance of the best performing baseline methods in both types including SVC and C-LSTM, we conduct the confusion matrix analysis of these methods. The main diagonal cells of the table represent the numbers true positives and the other cells represent the incorrect classifications with their related confusion probabilities. From the table, we see that C-LSTM obtained the similar kind of performance for all the categories but Conscientiousness. This deduces the effectiveness of the deep learning based approaches for the personality traits detection in Bangla.

VI. Conclusion and Future Direction

In this paper, we formulate the personality traits detection as a multiclass classification problem and proposed a first publicly available benchmark dataset on Bangla. We also present various state-of-the-art baseline methods and analyze their performance on this dataset. It showed that deep learning based approaches that utilized the features from pretrained embedding models performed well compared to the other approaches. We believe that the provided dataset and reported findings will be beneficial for other researchers to design more effective features and models for personality detection in Bangla.

In the future, we will explore the effective deep learning technologies for improving the performance. We also have a a plan to exploit the improved method in a distant supervision framework and create a larger version of this dataset.

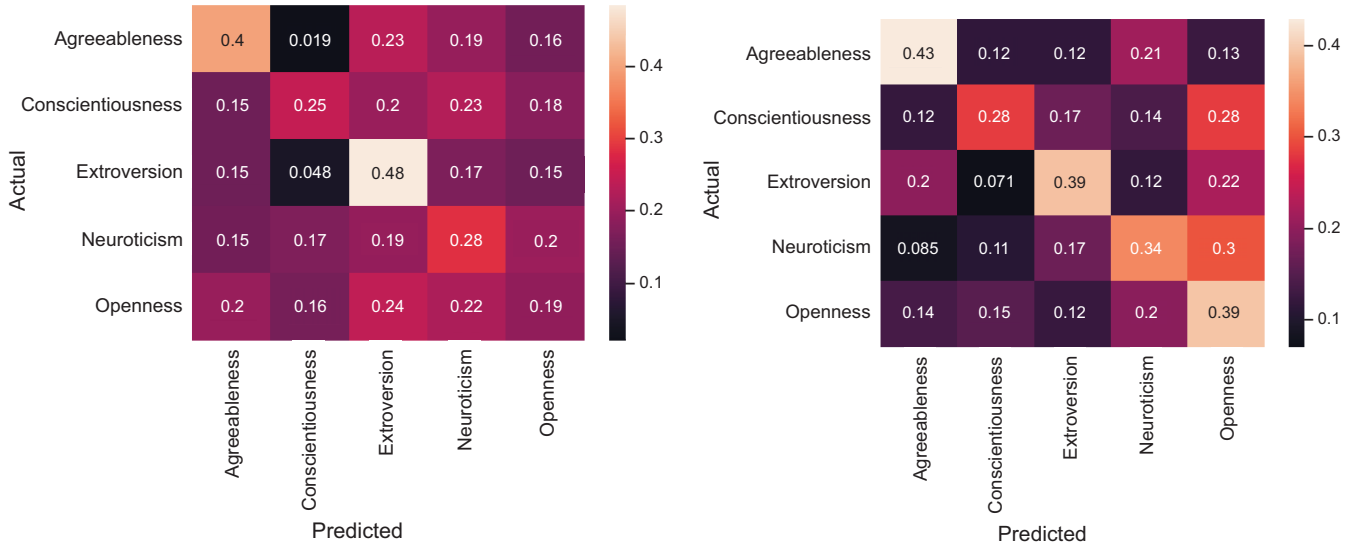

Fig. 3: Confusion matrix of SVC (left) and C-LSTM (right).

ACKNOWLEDGMENT

We would like to thank Mohammad Morshed Rana, Sourav Rudra, Hridoy Shill, and Jony Rudra for providing their valuable time during data annotation process.

REFERENCES

[1] J. Allik, "Personality dimensions across cultures," *Journal of personality disorders*, vol. 19, no. 3, pp. 212–232, 2005.

[2] R. Lambiotte and M. Kosinski, "Tracking the digital footprints of personality," *Proceedings of the IEEE*, vol. 102, no. 12, pp. 1934–1939, 2014.

[3] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2015, pp. 170–174.

[4] R. Ramos, G. Neto, B. Silva, D. Monteiro, I. Paraboni, and R. Dias, "Building a corpus for personality-dependent natural language understanding and generation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[5] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in personality*, vol. 40, no. 1, pp. 84–96, 2006.

[6] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive bayes classifier," in *16th Int'l Conf. Computer and Information Technology*. IEEE, 2014, pp. 366–371.

[7] L. R. Goldberg, "The development of markers for the big-five factor structure." *Psychological assessment*, vol. 4, no. 1, p. 26, 1992.

[8] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[9] K. G. Das and D. Das, "Developing lexicon and classifier for personality identification in texts," in *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 2017, pp. 362–372.

[10] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[11] H. Mohtasseb, A. Ahmed *et al.*, "Mining online diaries for blogger identification," 2009.

[12] M. B. Donnellan, R. D. Conger, and C. M. Bryant, "The big five and enduring marriages," *Journal of Research in Personality*, vol. 38, no. 5, pp. 481–504, 2004.

[13] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[14] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large scale personality classification of bloggers," in *international conference on affective computing and intelligent interaction*. Springer, 2011, pp. 568–577.

[15] F. Liu, J. Perez, and S. Nowson, "A recurrent and compositional model for personality trait recognition from short texts," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2016, pp. 20–29.

[16] M. Gjurković and J. Šnajder, "Reddit: A gold mine for personality prediction," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2018, pp. 87–97.

[17] B. Bazelli, A. Hindle, and E. Stroulia, "On the personality traits of stackoverflow users," in *2013 IEEE International Conference on Software Maintenance*. IEEE, 2013, pp. 460–463.

[18] E. Tighe and C. Cheng, "Modeling personality traits of filipino twitter users," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2018, pp. 112–122.

[19] F. Liu, J. Perez, and S. Nowson, "A language-independent and compositional model for personality trait recognition from short texts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 754–764.

[20] N. Shuyo, "Language detection library for java," 2010. [Online]. Available: http://code.google.com/p/language-detection/

[21] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2014.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[24] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A c-lstm neural network for text classification," *CoRR*, vol. abs/1511.08630, 2015.

[25] E. Bisong, "Google colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 2019, pp. 59–64.

[26] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.