MBTI personality classification using Apache Spark

Conference Paper · November 2021

DOI: 10.1109/ICECCO53203.2021.9663858

5 authors, including:

Kamila Orynbekova
SDU University
11 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Assem Talasbek
Suleyman Demirel University
11 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Abylay Omar
Suleyman Demirel University
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Andrey Bogdanchikov
Suleyman Demirel University
16 PUBLICATIONS **99** CITATIONS
SEE PROFILE

MBTI personality classification using Apache Spark

1st Kamila Orynbekova
Department of Computer Sciences
Suleyman Demirel University
Kaskelen, Kazakhstan
kamila.orynbekova@sdu.edu.kz

2nd Assem Talasbek
Department of Computer Sciences
Suleyman Demirel University
Kaskelen, Kazakhstan
assem.talasbek@sdu.edu.kz

3rd Abylay Omar
Department of Computer Sciences
Suleyman Demirel University
Kaskelen, Kazakhstan
abylay.omar@sdu.edu.kz

4th Andrey Bogdanchikov
Department of Computer Sciences
Suleyman Demirel University
Kaskelen, Kazakhstan
andrey.bogdanchikov@sdu.edu.kz

5th Shirali Kadyrov
Department of Mathematics and Natural Sciences
Suleyman Demirel University
Kaskelen, Kazakhstan
shirali.kadyrov@sdu.edu.kz

***Abstract*—Personality determines how person make decisions, speak or react on different situations. In this paper explained shortly the specifics of Myers-Briggs Type Indicator personality classification, then details of preparation of the experiment to run on Apache Spark platform. In experiment three different classification algorithms (Logistic Regression, Naive Bayes, Support Vector Machine) are used to train and predict MBTI personality pairs on a Kaggle dataset consisting of 8675 users tweets. In the end explained the data preprocessing and algorithm training, testing, validation details and results. The models with different vector combinations have been compared, and results have been described.**

***Index Terms*—MBTI, Apache Spark, Logistic Regression, Naive Bayes, SVM**

I. Introduction

A personality type is a set of personal qualities and psychological characteristics to describe similarities between a certain category of people. It is used by psychologists to describe differences and similarities between people, simplify the understanding of people's characters, predict behavior and highlight common patterns and connections in studies of different types of people [1]. There are 16 personality types according to Myers-Briggs Type Indicator(MBTI). The personality type is formed by a combination of four traits (or preferences in MBTI terminology).

1) Scale E-I. Extraversion - Introversion: indicates where consciousness is directed.

2) Scale S-N. Sensing-Intuiting: indicates how information from the outside world is perceived.

3) Scale T-F. Thinking-Feelings: indicates how decisions are made.

4) Scale J-P. Judgment-Perception: indicates what methods of preparation for action will be used by particular person [2]. Personality classification using text data such as posts in different social network accounts, essays based on MBTI with high level of accuracy is one of the modern tasks in Computer Science. The stem of current experiment was the work of the scientists from PES University (India), where three classification methods (Naive Bayes, SVM, Neural Networks) were used to classify the users' MBTI personality types according to their twits, the Kaggle datasaet was used. [1] The aim of this paper is to improve accuracy of classification and do it on distributed system using Pyspark.

PySpark is an open source API of Apache Spark on Python used for distributive processing of big data. It doesn't have it's own file system but supports popular systems like Hadoop Distributed File System (HDFS), Hbase, Cassandra, Amazon S3, Amazon Redshift, Couchbase etc. PySpark works on a MapReduce paradigm, which were developed for processing large amount of data distributively using key/value format. To run pyspark on a local machine, Java and some other software needed. Therefore, instead of a complex setup procedure, Google Colaboratory was chosen, which perfectly suits hardware requirements, and also comes with a wide range of libraries for data analysis and machine learning. PySpark and Py4J packages were installed. Py4J allows Python programs running in a Python interpreter to dynamically access Java objects from a presented Java machine.

II. Literature review

Sentiment analysis is one of the trending areas of research in Natural Language Processing related to text mining [3]. With the increase of social media usage, computational linguistics and text mining are becoming very crucial from many aspects [4]. As a part of research in this direction, the Personality Classification attempts to group individuals according to their psychological types which goes back to Carl Jung [5]. Following [6], some researchers regard personality as a five dimensional inventory and call it Big-five model, while the others prefer four dimensionality based on [5], known as Myers-Briggs Type Indicator (MBTI) [7]. For the task of Personality Classification from text, clearly social media provides a huge amount of resource where users hint their actual personality in a written manner [8]. In fact, as simple as the size of a person's social network already gives some insights on his personally [9]. Some of the early works, see e.g. [10], studied correlation between personality and publicly

978-1-6654-0945-2/21/$31.00

shared demographic data of the social network users, and others used linguistic features for the personality recognition [11].

One of the first attempts to classify personality based on user's Twitter profile is carried in [12]. In this work, fifty Twitter users were recruited to collect the data. To this end, 45-item Big-five model questionnaire survey was conducted to label the personalities. From each user, up to 2000 latest tweets were collected. The preprocessing steps contain aggregation, quantification, and 74 dimensional feature extraction including linguistic specifics. As a result of ZeroR and Gaussian Processes, both are machine learning algorithms, it was shown that each personality score can be predicted to within 11% - 18% of their true value.

Later in 2013, a Naïve Bayes algorithm with semi-supervised approach was utilizes for the personality classification task [13] using 4621 tweets. The Big-five model predictions were made based on descriptive features extracted groups of tweets where mean accuracy reached 84%.

MBTI personality prediction based on MOODLE course activities and Facebook course group activities was proposed in [14]. The data was collected from 240 Business College students with ten activity based descriptives and tested on 10 various classification models, namely, OneR, NaiveBayes, KNN, BayesNet, Kstar, Random forest, J48, JRIP, KNN /IBK, RandomTree, Decision Table. The best three performing architecture were found to be OneR, Random forest, and J48 with the respective accuracy of 97.40%, 93.23%, and 92.19%. While the accuracy of some of the algorithms are high, this approach is very specific to certain course activities, not end to end text analysis based approach.

Word count based personality classification approach is considered in [15] using both Big-five model and MBTI datasets collected from various social media sources. Their method uses singular value decomposition of matrices and experiments on different languages show that on MBTI personal classification the highest F1 score reaches 0.96 on Sensing/Intuition factor of French tweets.

Big-five personality trait based classification task is studied in [16] using 2467 stream of-consciousness essays using a convolutional neural networks (CNN). CNN filters were used to obtain n-gram feature vectors followed by the fully connected dense layer. Among various experiments, the best accuracy reaches 62.68%.

Social media based MBTI personality prediction problem was studied in [17] using 8660 tweets. The gold standard Linguistic Inquiry and Word Count methodology, a lexical resource EmoSenticNet, and open source semantic network ConceptNet were utilized in the processing stage. Together with TF-IDF 1850 dimensional feature vectors were generated. Three models, Naïve Bayes, SVM, and Neural Networks, were trained and tested for their accuracy. The SVM outperformed the others with maximum prediction accuracy reaching 88.4% on *Sensing vs Intuition* class.

A recent work [18] considers MBTI personality recognition on 643912 Turkish texts from 647 social media users. They report to achieve F1 score of 60% where TF ve TF-IDF feature vectors were used with Naive Bayes and FastText methodology.

While Naive Bayes is considered the classical approach, there are other comparable methods. In particular, a recent study [19] utilizes K-means clustering and Gradient Boosting techniques that outperforms Naive Bayes.

As the above review of literature suggests, this is relatively new area of research and more work is needed to improve the automated personality classification from social media posts [20].

III. Methodology

A. Dataset

In this work MBTI datased from open source platform Kaggle is used. Dataset contains information about 8675 users. Each row of the dataset has a users's MBTI personality trait label and their latest 50 tweets.

B. Preprocessing

Before using the tweets data is split to 50 different tweets and concatenated to one text from all of them. This is done to make it easier to process data in the next steps.

1) Remove twitter handlers and URLs: All user handlers, which starts from '@' sign are removed. This sign is used for mentioning another user in the tweet. Some of the tweets contains only the URLs to external resources. Regular expressions are used to remove them.

2) Remove stop words: All words that do not bring specific meaning are removed from each tweet. Python has natural language toolkit nltk, which has the corpus of stop words in English language. All of the stop words from each tweet that in this nltk's corpus are removed.

3) Stemming: Stemming is the process of transforming the word to their base form. This is the one of the basic techniques that used for data preprocessing. So each word in tweets go through this transformation.

4) Tokenize: Finally all tweets are tokenized. Tokenization is the process of splitting the text into smaller parts called tokens suitable for the Machine Learning classification algorithms. After tokenization we got a sequence of tokens.

C. Vector generation and training

To convert preprocessed data into vectors both CountVectorizer and TfidfVectorizer has been used. Count vectorizer generates vector from the given text based on the frequency of each word that occurs in the entire text. TF-IDF (term frequency - inverse document frequency) is measure that shows relevant a word in document. To calculate this metrics we need to parameters. First one is how many times a word appears in a document, and the second one is inverse document frequency of the word across a set of documents. In this experiment both vectors are passed to the classification algorithms separately. There are Logistic Regression, Naive Bayes, SVM classifiers. SVC kernel is used for SVM and multinomial Naive Bayes algorithm.

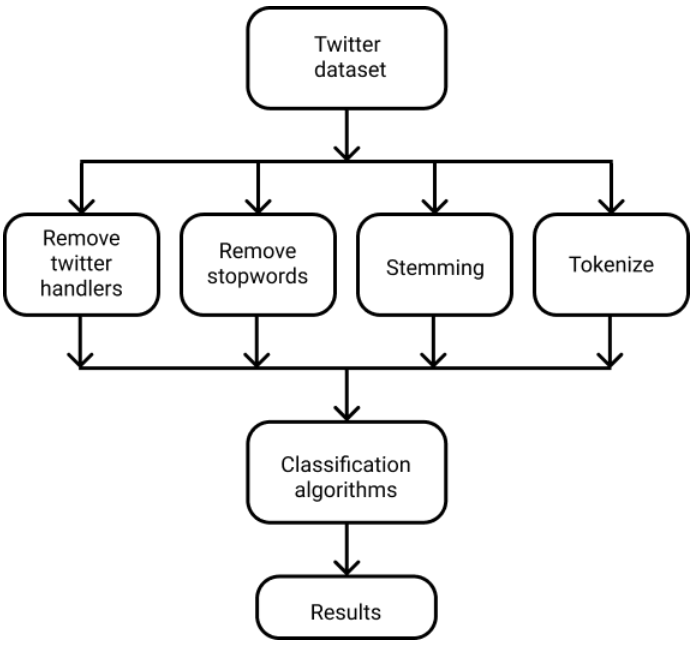

Fig. 1. Workflow

IV. Results of the Experiment

In the experiment PySpark library is used. Before the experiment, dataset divided into train, test and validation parts. In these proportions 0.80, 0.19, 0.01 respectively. After preprocessing our data is converted into vectors by using CountVectorizer and TfidfVectorizer. Then vectors passed to the classification algorithms. Each label is constructed from 4 pairs of psychological types. Each pair is guided through each model. The model is evaluated by using PySpark's binary classification evaluator. The results of the classification are given in "TABLE I" and "TABLE II".

TABLE I
Accuracy of algorithms with TF-IDF

	Logistic Regression	*Naive Bayes*	*SVM*
IE	79.67	79.37	79.43
NS	86.46	85.27	85.69
FT	77.28	77.28	79.90
JP	64.94	66.91	67.50

"TABLE I" displays the results of classification algorithms applied on pairs of psychological types with term frequency feature extraction. For IE label logistic regression gives 79.67 accuracy, Naive Bayes algorithm gives 79.37 and Support Vector Machines gives 79.43 accuracy. The best accuracy score for IE label with TF-IDF vectorizer give Logistic Regression classifier. For FT label the performance shows logistic regression with 86.46 accuracy score. For NS pair the highest accuracy has SVM classifier with 79.90 score. For JP label highest accuracy is 67.50 that gives SVM classifier.

TABLE II
Accuracy of algorithms with CountVectorizer

	Logistic Regression	*Naive Bayes*	*SVM*
IE	79.96	78.18	84.29
NS	86.11	85.03	88.57
FT	77.64	77.64	93.50
JP	65.68	67.68	75.87

V. Conclusion

In conclusion the experiment gives a new angle to observe the results of application of machine learning classification algorithms for MBTI personality classification problem. Three different algorithms give various results for each pairs of personality parameter.

The strong advantage of this research work is that all experiments done on the PySpark which uses Apache Spark platform. So usage of these algorithms can be extended by applying it on really large corpus of data and running on suitably large cluster. So the classification process can be performed on real time social network server or on any proxy server in real time where all these data are accumulated, which can be a future research work.

References

[1] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1076-1082, doi: 10.1109/ICACCI.2018.8554828.

[2] S. Ismail, B. Basharirad and S. Ismail, "Significant of MBTI personality model on decision making in university program selection," 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017, pp. 62-67, doi: 10.1109/ICITISEE.2017.8285560.

[3] W.Medhat, H. Ahmed, and K. Hoda. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5, no. 4, pp. 1093–1113, 2014.

[4] K. Chakraborty, B. Siddhartha, and B. Rajib. "A survey of sentiment analysis from social media data." IEEE Transactions on Computational Social Systems 7, no. 2, pp 450–464, 2020.

[5] C.G. Jung, "Personality types,". The portable Jung, pp. 178–272, 1971.

[6] R.R. McCrae and O.P. John. "An introduction to the five-factor model and its applications," Journal of personality, 60(2), pp. 175–215, 1992.

[7] K.P.Kumar and M.L. Gavrilova, "Personality traits classification on twitter," In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-8). IEEE. September 2019.

[8] M.D. Back, J.M.Stopfer, S. Vazire, S. Gaddis, S.C. Schmukle, B. Egloff, and S.D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," Psychological science, 21(3), pp. 372–374, 2010.

[9] A.S. Acar and M. Polonsky. "Online social networks and insights into marketing communications," Journal of Internet Commerce, 6(4), pp. 55–72, 2007.

[10] P.A. Rosen and D.H. Kluemper. "The impact of the big five personality traits on the acceptance of social networking website," AMCIS 2008 proceedings, p.274.

[11] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore. "Using linguistic cues for the automatic recognition of personality in conversation and text," Journal of artificial intelligence research, 30, pp. 457–500, 2007.

[12] J. Golbeck, C. Robles, and K. Turner. "Predicting personality with social media," In CHI'11 extended abstracts on human factors in computing systems, pp. 253–262, 2011.

[13] A.C. Lima and L.N. de Castro. “Multi-label semi-supervised classification applied to personality prediction in Tweets,” In 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (pp. 195-203). IEEE. September 2013.

[14] M.S. Halawa, M.E. Shehab, and E.M.R Hamed, “Predicting student personality based on a data-driven model from student behavior on LMS and social networks,” In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 294-299). IEEE. October 2015.

[15] N. Alsadhan and D. Skillicorn. “Estimating personality from social media posts,” In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 350-356). IEEE. November 2017.

[16] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria “Deep learning-based document modeling for personality detection from text,” IEEE Intelligent Systems, 32(2), pp.74–79, 2017.

[17] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, R. “Persona traits identification based on myers-briggs type indicator (MBTI)-a text classification approach.” In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1076-1082). IEEE. September 2018.

[18] K. Sönmezöz, O. Uğur, and B. Diri, “MBTI personality prediction With machine learning.” In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1–4). IEEE. October 2020.

[19] Z. Mushtaq, S. Ashraf, and N Sabahat, “Predicting MBTI Personality type with K-means Clustering and Gradient Boosting.” In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1–5). IEEE. November 2020.

[20] H. Ahmad, M.Z. Asghar, A.S. Khan, and A. Habib, “A systematic literature review of personality trait classification from textual content,” Open Computer Science, 10(1), pp. 175–193, 2020.