International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Personality Recognition using Multi-Label Classification

Anisha Yata[1], Prasanna Kante[2], T Sravani[3], B Malathi[4]

[1,2,3] *Student, Dept. Of CSE, Aurora's Technological and Research Institute, JNTUH, Hyderabad, Telangana, India.*
[4]*Assoc Prof, Dept. Of CSE, Aurora's Technological and Research Institute, JNTUH, Hyderabad, Telangana, India.*

Abstract - *To predict the personality of an individual, one needs to take a personality test. In this paper, we aim to automate the personality prediction of the users by implementing multi-label classifiers on textual data. The textual data collected are the views and opinions of the users posted by them on the social networking sites. It uses classification methods like Naive-Bayes, K- Nearest Neighbors and Support Vector Machine along with multi-label classifiers like Binary Relevance, Classifier Chains and Random k-Label sets for prediction of personality. We aim at comparing the accuracy obtained from these models which have been used and implement the best model for predicting the personality of new users.*

Key Words: **Personality classification, Naive-Bayes, SVM, KNN, Multi-Label Classification**

1. INTRODUCTION

"Personality refers to the individual differences in characteristic patterns of thinking, feeling and behaving. One is understanding individual differences in particular personality characteristics, such as sociability or irritability. The other is understanding how the various parts of a person come together as a whole" [1].

The automated classification of personality consists of comparing a user's personality against the standard personality tests taken. The Big5 is the most popular and standardized personality test. It uses five factors to describe the human psychology and personality [2]. These five factors are:

1. Openness
   (inventive/curious vs. consistent/cautious)
2. Conscientiousness
   (efficient/organized vs. easy-going/careless)
3. Extraversion
   (outgoing/energetic vs. solitary/reserved)
4. Assertiveness
   (friendly/compassionate vs. challenging/detached)
5. Neuroticism
   (sensitive/nervous vs. secure/confident)

The previous works (by Argamon et al 2005, Oberlander & Nowson 2006 and Mairesse et al. 2007), aimed at classifying the personalities of users based on long texts. Although, the current researchers aim at classifying personality based on data extracted from social networking sites and from long texts written in languages which aren't English [3]. It is to be noted that the case of personality classification is multi-labeled. This can be portrayed by the fact that a person can be extroverted, open as well as assertive, i.e, an individual can be assigned more than one class label. In this paper, we aim at classifying the personality of an individual by implementing multi-label algorithms along with base classifiers like Naive- Bayes, Support Vector Machine and K-Nearest Neighbours.

Section 2 talks about the papers referred to conduct this experiment and the inferences drawn from them. Section 3 depicts the flow graph or the research overview, which gives us the sequence of actions performed. Section 4 describes the multi-label and base classifiers used in this experiment along with the dataset and tools used. Section 5 gives us the result and conclusions drawn from this experiment.

2. LITERATURE SURVEY

A. Vinciarelli and G. Mohammadi [4] gave us insights, as to how can personality be determined with the help of computing technologies. The Automated Personality Recognition approach considers a spectrum of distal cues which includes written texts, non-verbal communication, social media, data collected from mobile devices and online computer games.

A. Kartelj, V. Filipovi, and V. Milutinovi[5], aimed at investigating possible improvements to the existing solutions of APC. This was done by analyzing different combinations of the available APC corpora, the psychological trait measures and learning algorithms. The data was obtained from social networks.

J. Oberlander[6] verified the accuracy that can be achieved when the authors of weblogs are classified into four major personality traits. This classification is done by using features of n-grams on binary and multi class classification.

B. Y. Pratama, R. Sarno[7] classified personality of users on twitter based on the tweets and the retweets posted by them on it. They are classified into the big five personality traits and their primary and secondary characteristics are specified. This paper uses algorithms like Naive-Bayes, SVM,

and KNN and gives a comparative conclusion of which classifier performs better.

3. ARCHITECTURE

The process involves Data Preprocessing, Attribute Selection, Classification Process and Model Generation. Firstly, we collect the raw data which is collected from the users, in the form of text. Now, this data must be preprocessed. This is done by implementing stemming, tokenizing, TF-IDF weighting, stop words. Next attribute selection is done. The classification model is generated using learning algorithms like Naive-Bayes, SVM, KNN.

The data is preprocessed by implementing the StringToWordVector filter in MEKA, where the IDFTransform, TFTransform is kept as true. The stemmer used is SnowBallStemmer, stopwordsHandler used is Rainbow and the tokenizer used is NGramTokenizer. The multi-label classifiers used are BR, CC, RaKEL with Naive-Bayes, SVM and KNN as base classifiers for each of them.

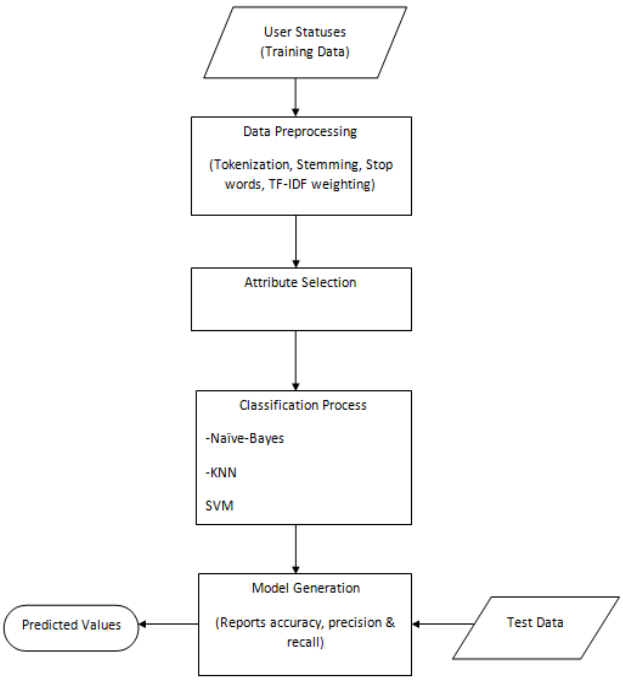

Fig -1: Architecture Diagram

4. IMPLEMENTATION

4.1 Dataset Used

The corpus used in this project is the MyPersonality dataset [3]. MyPersonality is a dataset which consists of user's personality scores and their Facebook statuses in the raw text along with several social network measures. The dataset consists of statuses from 71 users. These instances are partitioned into two disjoint sets of equal size, one will act as training set and the other as the test set.

4.2 Base Classifiers Used

The base classifiers used are Naive-Bayes(NB), K- Nearest Neighbours(KNN) and Support Vector Machine(SVM).

1)Naive Bayes Classifier:

It is a machine learning algorithm which is well known for it simplicity. Bayesian classification is also called as 'naive' as the computations involved in it are simple.

The class conditional probability is calculated by assuming conditional independence within the classes.

$$P(X|Y) = \frac{P(Y|X)P(X)P(Y)}{P(Y)}$$

1.1)Multinomial Naive Bayes on Text:

It is a modified version of Naive Bayes specially designed for text documents. NaiveBayesMultinomialText is a classifier which combines StringToWordVector filter capabilities with NaiveBayesMultinomial.It's advantage is that it works directly on string attributes.

2)k-NearestNeighbor

The role of nearest neighbour classifier is to represent the examples as data points, where d represents the number of attributes. It is the common classification based on the use of distance measures.

In KNN method, the training data becomes the model. When a classification is to be made to a new item, it's distance to each item in a training data set must be determined. The K-closest points in the training data are considered and then the new points are placed in new class which has maximum closest points.

3)Support Vector Machine:

Support Vector Machine (SVM) is a classifier which is defined by creating a separating hyperplane. It is used for linearly separable binary set. The main goal of svm is to classify the training vectors into classes and to design a hyperplane. It is also used for multi-dimensional datasets and the data points are represented as vectors.

The hyperplane which gives maximum margin is considered as the best hyperplane. Minimising the above term will maximize the separability and this gives the biggest margin here. SVM is very effective in high dimensional spaces, and we've different kernel functions for various decision functions.

The important point to be noted here is, SVM gives poor performance when #Features > #Samples.

4.3 Multi-Label Classifiers Used

Multi-Label classification classifies one instance into a set of labels. A multi-label classification can be solved by two approaches:

i. Problem Transformation

ii. Algorithm Adaptation Method

This paper aims at solving multi-label classification by implementing Problem Transformation.

Problem Transformation transforms a multi-label classification problem into one or more multiple single label problems.

The multi-labeled classifiers proposed to used in this experiment are Binary Relevance(BR), Random k-Labelsets(RAkEL) and Classifier Chains(CC).

Of the 16 evaluation metrics, 8 are considered for comparison here. These 8 metrics are: Hamming loss, Subset accuracy, Example based F1, Micro F1, Macro F1, Average Precision, Rank Loss, One error.

4.4 Tools Used

1)WEKA:

WEKA stands for Waikato Environment for Knowledge Analysis. It is a machine learning software written in Java, which was developed at the University of Waikato, New Zealand. It is free software and is used for Data analysis and predictive modelling.This tool consists of algorithms and tools which are used for Data mining.

2)MEKA:

A Multi-label/Multi-target extension to WEKA. MEKA is used when one needs to implement multi-label classification on a single instance.

5. RESULTS AND CONCLUSION

We compare the values of the evaluation metrics and draw conclusions as to which classifier would work best to predict the values of new users.

The metrics in bold represent the best values. Here, HL, SA, EBF, mF1, Mf1, AP, RL, and OE represents Hamming Loss, Subset Accuracy, Example Base F1, Micro F1, Macro F1, Average Precision, Rank Loss and One Error respectively.

These parameters are defined as the follows[8]-

1) Hamming Loss:

It reports, on an average, how many times a class label was incorrectly predicted. Its value lies between 1 and 0. If the value of Hamming Loss is less, performance is said to increase.

2). Subset Accuracy:

Subset accuracy reports whether the set of labels which has been predicted for a sample matches exactly to the corresponding set of truth labels. It's value lies between 1 and 0. If the value of Subset Accuracy is more, performance is high.

3)Example Based F1:

It is the average of the harmonic mean of example-precision and example-recall for every example. It's value lies between 1 and 0. If the value of Example Based F1 is more, performance is said to increase.

4) Micro F1:

It is the harmonic mean of micro precision and micro recall. It's value lies between 1 and 0. If the value of Micro F1 is more, performance is said to increase.

5) Macro F1:

It is the average of the harmonic mean of precision and recall of all the labels. It's value lies between 1 and 0. If the value of Macro F1 is more, performance also increases.

6)Average Precision:

It is the average fraction of labels ranked above an actual label yi that actually are in yi. It's value lies between 1 and 0. If the value of Average Precision is high, the classifier is said to have a better performance.

7) Rank Loss:

It evaluates the average proportion of the class label pairs which have been incorrectly ordered for an instance. It's value lies between 1 and 0. If the value of Rank Loss is less, performance is said to increase.

8)One Error:

This is defined as the fraction of examples whose top-ranked label does not belong to the label set which is relevant. It's value lies between 1 and 0. If the value of One Error is less, performance is said to be more.

Table-1: Evaluation Metrics Comparison

	HL	SA	EBF	mF1	Mf1	AP	RL	OE
BR+MNB	0.403	0.445	0.54	0.604	0.405	0.81	0.234	0.407
BR+SVM	0.322	0.441	0.384	0.586	0.464	0.643	0.327	0.494
BR+KNN	0.012	0.988	0.852	0.986	0.981	0.999	0.001	0.143
CC+MNB	0.377	0.386	0.488	0.539	0.284	0.709	0.358	0.407
CC+SVM	0.331	0.382	0.305	0.567	0.473	0.6	0.356	0.564
CC+KNN	**0.001**	**0.998**	**0.857**	**0.998**	**0.998**	**0.999**	**0.001**	**0.143**
RAkELd+MNB	0.418	0.141	0	0	0	0.559	0.504	0.565
RAkELd+SVM	0.307	0.455	0.39	0.621	0.499	0.631	0.323	0.565
RAkELd+KNN	**0.001**	**0.998**	**0.857**	**0.998**	**0.998**	**0.999**	**0.001**	**0.143**

From the above table, we can draw the conclusion that KNN as base classifier gives the optimal results when used with the multi-label classifier Classifier Chains or Random K-labelsets.

ACKNOWLEDGEMENT

We would like to express our gratitude to Mr. J. SRIKANTH, Director, Aurora's Technological and Research Institute for providing us congenial atmosphere and encouragement. We express our sincere thanks to Head of the Department Ms. A. Durga Pavani, for giving us the support to us throughout this course. We convey thanks to our Project Guide Ms. B Malathi of Department of CSE & IT for providing encouragement, constant support and guidance which was of a great help to complete this work successfully.

REFERENCES

[1] Kazdin, A. E. (2000), "Encyclopedia of Psychology", Oxford Press.

[2] Goldberg, L.R. (1993). " The structure of phenotypic personality traits". American Psychologist. 48: 26–34.

[3] F. Celli, F. Pianesi, D. Stillwell, M. Kosinski, "Workshop on computational personality recognition (shared task)," In Proceedings of WCPR13, in conjunction with ICWSM-13, 2013.

[4] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," IEEE Trans. Affect. Comput., vol. 5, no. 3, pp. 273–291, Jul./Sep. 2014.

[5] A. Kartelj, V. Filipovi, and V. Milutinovi, "Novel approaches to automated personality classification: Ideas and their potentials," in Proc. of 35th international convention, MIPRO. 2012, pp. 1017-1022, IEEE.

[6] J. Oberlander, "Whose thumb is it anyway? Classifying author personality from weblog text," In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 627–634, 2006.

[7] B. Y. Pratama, R. Sarno, "Personality classification based on Twitter text using Naive Bayes KNN and SVM", 2015 International Conference on Data and Software Engineering (ICoDSE), pp. 170-174, 2015.

[8] Madjarov, G., Kocev, D., Gjorgjevikj, D., & Dzˇeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45, 3084–3104.