Developing Lexicon and Classifier for Personality Identification in Texts

[1]Kumar Gourav Das **[2]Dipankar Das**
[1]Department of Computer Science and Engineering
Future Institute of Engineering & Management, Kolkata, India
[2]Department of Computer Science and Engineering
Jadavpur University, Kolkata, India
[1]kumargouravdas18@gmail.com, [2]dipankar.dipnil2005@gmail.com

Abstract

Personality, an essential foundation of human behavior is difficult to identify and classify from texts because of the scarcity of explicit textual clues. Several works were attempted for personality identification by employing well-known lexicons like WordNet, SentiWordNet, SenticNet etc. However, a lexicon solely devoted for identifying different types of personality is rare. Thus, in the present article, we have discussed the methodologies to develop a personality lexicon from the Essay dataset, a personality corpus based on Big Five model. We have used a frequency based N-gram approach to extract the unique words as well as phrases with respect to each of the Big Five personality classes. In addition to the words, we have added another feature, corpus based probability of occurrence into the lexicon. Finally, we have evaluated our lexicon on a small Youtube personality dataset and found satisfactory coverage. In addition, we have developed a LIWC based classification framework by employing several machine learning algorithms followed by feature selection using information gain and correlation techniques. SVM and Logistic Regression achieved the maximum accuracies of 78.52% and 62.26% with a reduced set of feature size 15 and 10 selected by information gain and correlation attribute evaluation, respectively.

1 Introduction

Personality refers to the individual differences in characteristic patterns of thinking, feeling and behaving. Personality is considered as the most difficult human attribute to understand. Personality traits are traditionally measured through the use of questionnaires such as the Big Five Inventory (BFI) (Tett and Rothstein ,1991). However, an alternative approach is to analyze an individual's linguistic differences. Personality of a person is reflected in his behavior and speech which indirectly affects the job performance, one's effectiveness in work. Not only in jobs, there are so many other applications where we can use the advantage of personality identification including social network analysis e.g., Twitter (Pratama and Sarno, 2015), Facebook (Golbeck and Turner ,2011) (Alam Firoj and Ricardi, 2013) (Iacobelli Culotta, 2013), recommendation systems (Golbeck and Turner, 2013), sentiment analysis/opinion mining, Author Profiling (Rangel Pardo and Daelemans,2015), construction of emotion lexicon (B.G. Patra et al,2013)and many others. Personality is correlated with many other aspects of our daily life such as job success (R.P. Tett et al,1991), marital happiness (E.L. Kelly et al, 1987) too. Now, the recent trend is automatic identification of personality from some text or audio or may be video also. We can identify personality from various single modes (audio, video, texts etc.) as well as in multimodal way.

However, due to scarcity of proper audio dataset based on personality, we have started our experiment only on text dataset. We have used two standard text dataset, Essay (Pennebaker,et al. 1999) and Youtube (J.I. Biel et al., 2013).

Personality research is being nurtured as a developing field and only few works have been done till date. There are lexicons like SentiWordNet 3.0 (S. Baccianella et al, 2010), LIWC (Y.R.Tausczik et al, 2010) (F. Mairesse et al,

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 362–372

2007), Senticnet 3.0 (Cambria et al, 2012) etc. which help in identifying personality. However, to the best of our knowledge, there is no open source lexicon that contains words/phrases of a particular type of personality. Thus, one of our prime motivations is to develop lexicons for each of the Big Five personality type, separately.

In the present work, we have classified personality obtained from the written text based on the model of Big Five personality classes. The Big Five personality model is considered as a standard model for personality traits. This Big Five personality model has been used in many personality detection research works as they help in developing several applications.

According to Big Five model, personality is assessed in five dimensions of OCEAN –

a. **O**penness **(***inventive, curious***)**
b. **C**onscientiousness **(***organized***,** *efficient, sincere***)**
c. **E**xtroversion **(***energetic***,** *sociable*)
d. **A**greeableness **(***friendly, trustable and compassionate***)**
e. **N**euroticism **(***apprehensive, sensible***)**

In the present work, we have developed a lexicon of words and phrases corresponding to each of the Big Five personality classes. However, we have restricted ourselves to find only those words that belong to only one particular class of personality and not in any other class. The approach used in this work is fully automated and no manual or human interaction has been carried out. The hypothesis considered is a two tier filtration strategy; first, we identified the distinct words of each personality class that do not belong to any other class by using the set disjoint operations. Using this approach, we have obtained four different sets of words and phrases corresponding to each of the Big Five personality classes. Thereafter, we have considered the intersection of four different set of words and phrases as previously obtained and formed a lexicon for each personality class. We have used a n-gram method where in case of unigrams, we have obtained unique set of words and in case of bigrams and tri-grams, we extracted a unique set of phrases. Finally, the probability of each word or phrase has been calculated in order to add the occurrence probability information into the lexicon. In addition, we have explored the LIWC tool and developed a classification module for identifying and classifying the instances of both **Essay** dataset and **Youtube** personality dataset with a reduced set of features identified using information gain and correlation based techniques. The rest of paper is organized as follows. In Section 2, we have discussed the related work ,in Section 3 we have discussed about the dataset and preprocessing . Section 4 describes the lexicon development whereas Section 5 describes the developmental phases of LIWC based classification module. Finally, Section 6 mentions the observa tions and comparisons followed by conclusions and future work.

2 Related Work

Correlation between linguistic clues and personality traits have been identified to discover the way for carrying research in the area of automatic personality classification. We mainly focused on classifying personality traits based on text due to scarcity of multimodal dataset. To the best of our knowledge, the field be in its infancy. Though several researchers have started their struggles in identifying personality from text by adopting various approaches, n-gram always has a huge impact in most of the cases (J.Oberlander et al, 2006).

We extracted linguistic features from essay dataset using a text analysis tool, Linguistic Inquiry and Word Count (LIWC), (F.Mairesse et al. 2007), (G.Sidorov2006). Several authors used the LIWC tool for identifying the impacts of different linguistic features on different personalities as discussed in (Yla R.Tausczik et al, 2009), (F. Mairesse et al, 2007). LIWC is a text analysis tool that counts and sorts words based on their psychological and linguistic category. NRC is another lexicon that contains more than 14000 distinct words annotated with 6 emotions like *anger, fear, sadness, joy, disgust* and *surprise* along with two types of sentiments like *positive* and *negative*. The NRC lexicon has been used in other related work on personality where the authors explored the features of NRC and LIWC both (Mohammad et al., 2013) (G. Farnadi et al, 2014). MRC is a psycholinguistic database that contains psychological and distributional information of more than 150,00 words annotated with 14 features like phonemes (*Nphon*), syllables (*Nsyl*)(Coltheart, 1981) .

On the other hand, rough set based machine learning techniques have been used for personality identification (Gupta et.al 2013) whereas Naïve Bayes, KNN and SVM were also employed for

personality identification on Twitter texts (B. Y. Pratama et al. , 2015). A few authors have also investigated the age and gender related information from formal texts (Burger, J.D, 2011).

In contrast to such previous attempts, in the present work, we aimed to develop a personality lexicon of five different Big Five classes where the words even phrases are categorized according to the Big Five personality model. It has to be mentioned that one of our strict criteria that has been followed here is that no word or phrase of a particular personality class should mingle with words of other personality class. The words are also associated with their probability scores which make the lexicon useful for classifying personality from texts. Moreover, we have used information gain and co-relation techniques to conduct the feature ablation study for developing a personality classifier also.

3 Dataset and Preprocessing

In order to start with our experiments, we have used two text datasets. For developmental purpose, we have used the Essay dataset (Pennebaker, J. W., 2007) and for testing the coverage and performance evaluation purpose, we have used the YouTube dataset (J.I. Biel, 2013). Huge number of researchers used these two datasets to develop and test various personality detection models. Thus, we have considered these two as our gold standard datasets.

3.1 Eassy Dataset

Essay dataset (Pennebaker, J. W., 2007) is a large dataset that consists of 2468 text documents labeled with personality classes. The labeled personalities are based on the classes of Big Five personality traits. The classes are *Openness* (O), *Conscientiousness*(C), *Extraversion* (E), *Agreeableness* (A) and *Neuroticism* (N).

3.2 Youtube Dataset

Youtube personality dataset (J.I. Biel, 2013) consists of a collection of speech transcriptions, and personality impression scores of 404 YouTube users. These files are also tagged with the Big Five personality classes. Their speeches were transcribed by professional annotators and the transcriptions contains approximately 10K unique words and 250K word tokens.

3.3. Preprocessing Text

3.3.1. Labeling

We have started our experiments by considering each and every personality class separately because we were trying to find out unique words or phrases with respect to each of the Big Five personality classes. For that very reason, at first, we considered only those files that belong to only one specific class. Each character of such a tuple of five represents each of the Big Five personality classes (e.g.**O**penness -Y, **C**onscientiousness -N, **E**xtroversion-N,**A**greeableness-Y,**N**euroticism-N) represents the instance belongs to *Openness* and *Agreeableness*). The basic steps of pre-processing are mentioned below.

3.3.2. Lower case conversion

Change the whole text into lower case so as to maintain consistency in our further approaches.

3.3.3. Tokenizing

Change each of the sentences into a collection of single words.

3.3.4. Filtering

We have eliminated the stop words and numbers because stop words are common words that have no meaning but are compulsory to maintain the grammatical structure of language (e.g., *is, am are*). At first, we find out the count, i.e. the number of texts that belong to only one specific personality class and then the total number of texts that belong to that specific class irrespective of whether the file belongs to other classes or not. Then, we count the total number of phrases and words for both these two types of classes and calculated the percentage of occurrence of phrases and words in one specific personality class.

4 Lexicon Developing Module

We assumed that the words people use in their daily life reveals important aspects of their social and psychological uniqueness. Our objective is to explore different methods to find out words that are commonly used by the people belonging to a particular personality class. Therefore, we designed the n-gram module to identify the words or phrases that distinguishingly classify an instance of that particular class.

4.1 N-gram Module

We developed a lexicon for different personality classes that contains not only unigrams but bigrams and trigrams also. The distinctions between linguistic styles and linguistic contents can be seen in how two people may make a simple request. E.g., "*Would it be possible for you to give me a glass of water?*" and "*Give me a glass of water*" both the sentences express the speaker's desire for water and direct the listener's action. However, the two utterances also reveal the speaker's personality. N-gram feature would help us to find the unique words of each and individual personality type. Thus, in order to find the unique words of each personality class, we carried out different levels of experiment. We try to find out those texts that belong to a specific personality class using Equation 1.

$$T_{c=\ \cap_{i=1}^{n}(\theta c - \theta_i^{\circ})\ i \in all\ class\ except\ C}^{w} \quad (1)$$

$$where\ T_w^c = Total\ number\ of\ unique\ words\ of\ a\ specific\ class$$

$$\theta c = words\ belong\ to\ a\ specific\ class$$

$$\theta_i^{\circ} = words\ belong\ to\ the\ reamining$$

Fig 1: Equation for unique word count

Consequently, the frequencies of those unique words have been estimated. It was observed that stop words do not help in detecting the personality. Thus, the stop words were removed for counting the unigrams only but, for bigrams and trigrams, the stop words were not removed as bigrams and trigrams were considered to be our potential repositories of personality phrases. Initially, we estimated top 300 n-grams for each class. Then, using equation 1, we calculated the n-grams that belong to only that class. Next, the same process is repeated for obtaining top 500 and 1000 n-grams. Similarly, the unique words of other personality classes were also calculated.

E.g., a few instances of the lexicon formed for each of the Big Five personality classes along with their frequencies are "*strange"* that occurs in **openness** class 18 times, "*suppose"* that occurs in **agreeableness** class 14 times. The bigram "*really don't"* occurs in **Neuroticism** class 32 times. The trigram "*I don't know* occurs" in **Extrovert** class 40 times etc.

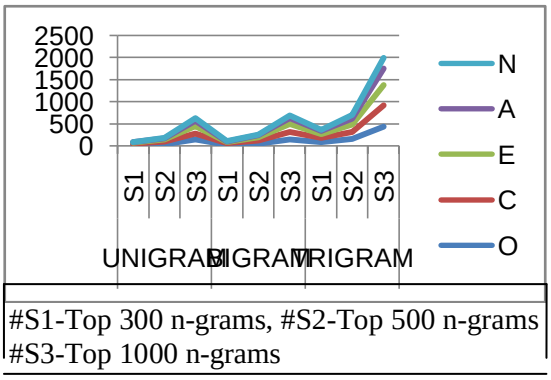


Fig 2: N-gram counts

4.2 Probability Calculation Module

The lexicon for each of the Big Five personality classes has been prepared in our previous step. Next, we need to find the occurrence in terms of probability of each word based on Equation 2.

$$P_w = \frac{T_u^c}{C_w} \quad (2)$$

$$where\ , P_w = Probability\ of\ occurence\ of\ a\ word\ in\ that\ class$$

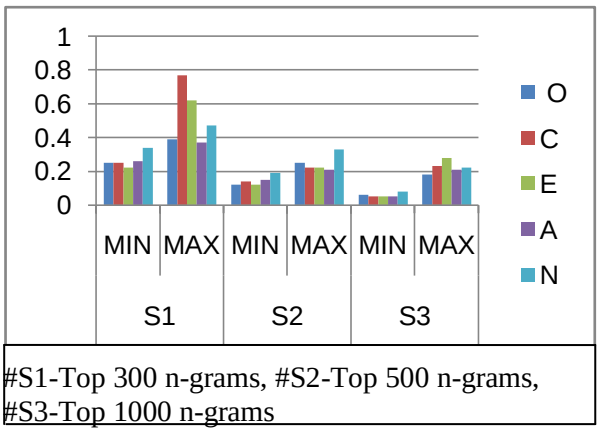
$$T_u^c = Total\ number\ of\ unique\ unigram\ of\ that\ class$$

$$C_w = Total\ occurence\ of\ that\ word$$

Fig 3: Equation for counting n-gram probability

The range of probability is identified by the lowest and the highest probability scores obtained for each personality class. For example, if the word *W1* occurs *X* times in a particular Big Five class say *Z* ,and the total number of unique unigram of Z class is Y, the occurrence probability of that word *W1* is *X/Y*. The occurrence probability is also calculated for both bigrams and trigrams. The probabilities of unigrams are shown in Figure 3.

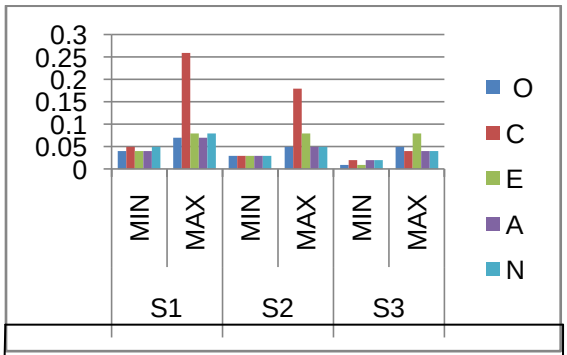
From our experiment, we observed that initially, we have started our experiment with top 300 n-grams and as we are interested in finding only those words that belong to only that specific class, so we apply two tier filtering. However, in order



#S1-Top 300 n-grams, #S2-Top 500 n-grams, #S3-Top 1000 n-grams

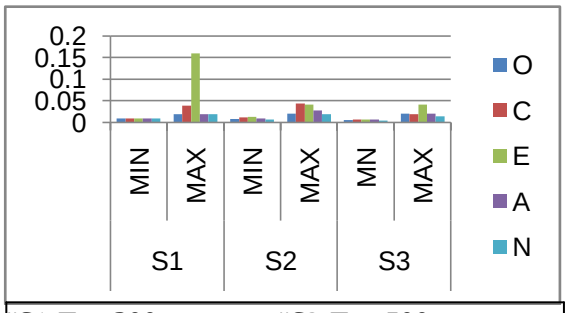
Fig 4: The occurrence probability graph of unigram

to follow this technique, we achieved very less number of words and phrases. Thus, we continue our experiments with top 500 and 1000 n-grams. While increasing the size, we observed that.



#S1-Top 300 n-grams, #S2-Top 500 n-grams, #S3-Top 1000 n-grams

Fig 5: The occurrence probability graph of bigram



#S1-Top 300 n-grams, #S2-Top 500 n-grams, #S3-Top 1000 n-grams

Fig 6: The occurrence probability graph of trigram

we have obtained more number of words and phrases But, we get some words and phrases whose individual frequency is very less and thus their occurrence probability is also very less in that class. Therefore, we can conclude that the words do not have any influence in classification.

4.3 Evaluation

Now, we want to test it against on another dataset. For testing, the dataset used is YouTube Dataset. A simple algorithm for testing is defined in Figure 7. We have adopted a strict evaluation scheme such that each of the test documents should belong to only one personality class. We do not get satisfactory result. Overall we get 35% accuracy.

5 LIWC based Classification Module

The advancements in the field of personality trait classification till could not answer the question that which features are most significant for personality

Step 1. *Split the text into sentences.*

Step 2. *Preprocess the dataset.*

Step 3. *Categorize each word of a sentence using Big Five personality lexicon*

Step 4. *A sentence is classified into one of the Big Five personality class based on the majority class of the words of that sentence*

Step 5. *A file is classified into one of the Big Five personality classes based on the majority class of the sentences of that file.*

Fig 7: Algorithm for testing Youtube dataset

classification. Thus, initially, we have started with some basic approach to build a lexicon for Big Five personality class. However, we could not achieve satisfactory results and thereafter we use LIWC, a widely used text analysis tool to improve our result.

5.1 LIWC

Linguistic Inquiry and Word Count, is a widely used text analysis tool to efficiently classify texts. We extracted 69 features of LIWC for each of the documents. Then, we tried to reduce the size of the feature set. The classifiers are then built on the re-

duced set of features and the performances are compared with respect to a complete set of features. This research aims to show that the usefulness of LIWC on personality identification and how different feature reduction techniques such as Information Gain, PCA can help in getting better result for classification.

5.2 Feature Extraction

LIWC was developed by Pennebaker et al., 2007. It is a text analysis tool that is employed to quantify features and allowed for text classification and prediction. LIWC is a dictionary that contains 80 categories. For each file, we consider each word and search through the dictionary. If the target word is found in the dictionary, the category count of that word is incremented. Though the dictionary contains 80 features, we initially started with our experiment on 69 features. We count the number of *anger, sad, pronoun, posemo* (positive emotion word), *negmo* (negative emotion word) etc. Based on the method, each file of essay dataset was fed into the LIWC. The output file contains 69 features and each feature has one of the Big Five personality traits as the classification label.

5.3 Classification

In order to validate the feature set, a number of experiments have been performed to evaluate how accurate they are in predicting Big Five personality traits. A 10-fold cross validation was performed on our feature set to assess the accuracy. We tested a number of popular classification algorithms like Support Vector Machine (libSvm), SMO, Multilayer Perceptron and Simple Logistic Regression.

5.4 Feature Selection

The feature set is reduced by selecting a subset of original features. The removed features are not used in classification anymore. One of the aims of feature selection methods is to determine a subset of features for which the accuracy is maximized.

5.4.1 Information Gain

As we want to determine which attribute in a given set of training feature vectors is most useful we use information gain. Information gain tells us how important a given attribute of the feature vector is thus helps in reducing the feature set size while keeping the accuracy same. One of the most important contributions of this research is to determine the most important features among the 69 LIWC features that can used for classifying the Big Five personality, keeping the accuracy same or making it better.

By considering the top 10 LIWC features of Information Gain, the obtained result was not satisfactory. Then, on increasing the size of the LIWC feature set with 15 more features, the result was not improved. Finally with a LIWC feature set of size 20, the result is nearly the same when compared to the result that is obtained with a LIWC feature set of size 69. Thus, in future, our aim will be to strengthen the feature set by extracting features from other lexicons like MRC, NRC and other optimization techniques like PCA.

5.4.2 Correlation Attribute Evaluation

After extracting features using LIWC, we wanted to reduce the feature set size and that's why we apply Information Gain. Now, we use another feature reduction technique, Correlation Attribute that evaluates the worth of an attribute by measuring the correlation between it and the class.

5.5 Result Analysis

Initially, the experiment has been performed on 69 features. We achieved better result on Libsvm on *radial basis function* kernel compared to Libsvm on *polynomial* kernel. Next, we tried to reduce our feature set. We test our result using two feature reduction techniques, one is Information Gain and another is Correlation Attribute evaluation. We test our results in two dataset, one is Essay dataset and another one is Youtube dataset. In essay dataset, we obtained very good result (accuracy of 78%) and in Youtube dataset we achieved 56% accuracy. In Table 1 and Table 2, we give the details of result.

5.5.1 Feature Level Analysis

In this experiment, we have observed the influence of different LIWC features on classification. We have done our experiment with different variations of features and tried to analysis the Precession (P),

Recall (R) and F-measure (F) to identify the importance of different features.
From LIWC, we started our experiment with 69 categories of words. Using Information Gain, when we ranked the attributes, we obtained top 10 features like **home, we, job, inhib (**inhibition**), excl (**exclusive**)** etc. which are very important importance of different features.

From LIWC, we started our experiment with 69 categories of words. Using Information Gain, when we ranked the attributes, we obtained top 10 features like **home, we, job, inhib (**inhibition**), excl (**exclusive**)** etc. which are very important categories for classification. Then, when we increase the size, the categories like **occup (**occupation**), leisure, anger** are added. Finally, when we considered top 20 features, we achieved the best classification result and some important features like **sad, negmo (negative emotions)** which were added further. Thus, we can say that among 69 features, these features have more importance than other features.

Using correlation attributes and when we rank the attribute under top 10 features, we get features like **smile, you, home, posfeel (**positive feeling**)** etc. Then, increasing size, we obtained features like **friends, time, school, eating** etc. Finally, while considering top 20 features, we achieved the best result on some features like **we, past, family, achieve, see** etc. as mentioned in Table 3 and Table 4.

6 Observation and Conclusions

A Personality Lexicon for Big Five Personality classes have been developed. The main objective is to find out some unique words that are mostly used by a particular type of personality. According to the design module, a lexicon with top 300, 500 and 1000 n-grams has been obtained. Our observation says when we continue our experiment with top 300 n-grams, the size of our lexicon is small and as we increase it to 500 n-grams and 1000 n-grams our lexicon size increases but it also contains words whose frequency in the text are very less.

On the other hand, in case of calculating occurrence probability of individual word belonging to a particular personality class, we observed some issues. When we take top 300 word, the occurrence probability is very high and as we take top 500 and 1000 n-grams, the occurrence probability decreases. As a result they do not help us much in classification. Thus, we can conclude that When we take top 300 n-grams, we get best result.

We developed our lexicon based on Essay dataset and tested it on YouTube dataset. As there is no topic related restriction on both dataset, the datasets contains diverse topics and that makes our job more difficult for personality identification and thus to develop a proper lexicon of a personality class becomes more difficult.

For development of lexicon, we already have discussed that we used two levels of filtering to eliminate all the words except a few which belongs to a particular class only. In order to maintain this process, we eliminate many words that may be important for us in classification. For example, the frequency of word **"Strange"** occurs in **Openness** class is **239** times and in **Extrovert** class is **20** times as because we are interested to find only those words that belong to **Openness** class. We eliminate the word **"Strange"** from the lexicon of Openness**.** As frequency of the words "strange" is so high in open class, so it may be an important unigram for the Openness class. So, for better classification result, we have to apply some threshold value which can be a future prospective of thiswork.

By using only n-gram approach we didn't get satisfactory result .Then we use LIWC for classification and we get very good result. Then we try to reduce the feature set by reducing the size of the feature set while keeping the accuracy same. We then use information gain optimization technique and reduce the size of the feature set from 69 to 20 while keeping the accuracy same.

Classifier	Accuracy (in %)							
	(Size = 69)		(Size = 10)		(Size = 15)		(Size = 20)	
	#IG	#CRA	#IG	#CRA	#IG	#CRA	#IG	#CRA
Libsvm(#1)	78.52	78.52	**78.18**	73.48	**78.52**	77.51	**78.52**	78.52
Libsvm(#2)	68.45	68.45	67.11	66.44	67.78	66.77	65.77	65.10
Multilayer Perceptron	63.75	63.75	33.55	35.23	42.61	41.27	47.31	56.71
Simple Logistic	41.94	41.94	30.20	31.87	28.18	32.88	30.20	32.88
SMO	38.92	38.92	25.50	30.53	26.84	31.20	29.86	35.23
Libsvm(#1):libsvm with on radial kernel. Libsvm(#2):libsvm with on polynomial kernel.#IG: Information Gain. #CRA: Correlation Attribute.								

Table 1: Result Analysis on different size feature set and on different classifier on *Essay* dataset

Classifier	Accuracy (in %)							
	(Size = 69)		(Size = 10)		(Size = 15)		(Size = 20)	
	#IG	#CRA	#IG	#CRA	#IG	#CRA	#IG	#CRA
Libsvm(#1)	56.60	56.60	56.60	56.60	56.60	56.60	56.60	56.60
Libsvm(#2)	45.28	45.28	45.28	47.16	37.73	30.18	41.50	39.62
Multilayer Perceptron	49.05	49.05	56.60	45.23	52.83	47.16	43.39	50.94
Simple Logistic	49.05	49.05	52.83	**62.26**	58.49	**62.26**	52.83	**60.37**
SMO	56.60	56.60	56.60	56.60	54.71	56.60	54.71	56.60
Libsvm(#1):libsvm with on radial kernel. Libsvm(#2):libsvm with on polynomial kernel.#IG: Information Gain. #CRA: Correlation Attribute.								

Table 2: Result Analysis on different size feature set and on different classifier on *Youtube* dataset

FEATURE	#NOF	#P	#R	#F
Eating, Home, we……home, job	10	0.86	0.78	0.79
We, Insight, occup……Other, Excl, Anger	15	0.88	0.78	0.80
See, prep, sad, motion…….anger, we, job, home	20	0.88	0.78	0.80

#NOF=number of file, **#P=**precision, **#R=**Recall, **#F=**F-measure.

Table 3: Feature selection using Information Gain and analysis with respect to precession, Recall and F-measure

FEATURE	#NOF	#P	#R	#F
Smile, you, home, sports…….senses.	10	0.82	0.73	0.74
Friends, time, school…….smile, leisure	15	0.87	0.77	0.79
we, past, family……..home ,achieve ,school	20	0.88	0.78	0.80

#NOF=number of file, **#P=**precision, **#R=**Recall, **#F=**F-measure.

Table 4: Feature selection using Correlation Attribute and analysis with respect to precession, Recall and F-measure

References

Alam Firoj, Evgeny A. Stepanov, and Giuseppe Riccardi. "Personality traits recognition on social network-facebook." *WCPR (ICWSM-13), Cambridge, MA, USA* (2013).

Burger, John D., John Henderson, George Kim, and Guido Zarrella. "Discriminating gender on Twitter." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301-1309. Association for Computational Linguistics, 2011.

Biel, Joan-Isaac, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. "Hi youtube!: Personality impressions and verbal content in social video." In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 119-126. ACM, 2013.

Biel, Joan-Isaac, and Daniel Gatica-Perez. "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs." *IEEE Transactions on Multimedia* 15, no. 1 (2013): 41-55.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *LREC*, vol. 10, pp. 2200-2204. 2010.

Cambria, Erik, Catherine Havasi, and Amir Hussain. "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis." In *FLAIRS conference*, pp. 202-207. 2012.

Farnadi, Golnoosh, Shanu Sushmita, Geetha Sitaraman, Nhat Ton, Martine De Cock, and Sergio Davalos. "A multivariate regression approach to personality impression recognition of vloggers." In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pp. 1-6. ACM, 2014.

Gupta, Umang, and Niladri Chatterjee. "Personality traits identification using rough sets based machine learning." In *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*, pp. 182-185. IEEE, 2013.

Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." In *CHI'11 extended abstracts on human factors in computing systems*, pp. 253-262. ACM, 2011.

Golbeck, Jennifer, and Eric Norris. "Personality, Movie preferences, and recommendations." In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp.1414-1415. ACM, 2013.

Hu, Rong, and Pearl Pu. "Enhancing collaborative filtering systems with personality information." In *Proceedings of the fifth ACM conference on Recommender systems*, pp. 197-204. ACM, 2011.

Iacobelli, Francisco, and Aron Culotta. "Too neurotic, not too friendly: structured personality classification on textual data." In *Proc of Workshopon Computational Personality Recognition, AAAI Press, Melon Park, CA*, pp. 19-22. 2013.

Kelly, E. Lowell, and James J. Conley. "Personalityand compatibility: a prospective analysis of marital stability and marital satisfaction." *Journal of personality and social psychology* 52, no. 1 (1987): 27

Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. "Using linguistic cues for the automatic recognition of personality in conversation and text." *Journal of artificial intelligence research* 30 (2007): 457-500..

.

Mohammad, Saif M., and Svetlana Kiritchenko. "Using nuances of emotion to identify personality.

" *Proceedings of ICWSM* (2013).

Max. Coltheart. 1981. . "the mrc psycholinguistic database."". *The Quarterly Journal of Experimental Psychology* ,, 33.

Oberlander, Jon, and Scott Nowson. "Whose thumb is it anyway?: classifying author personality from weblog text." In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 627-634. Association for Computational Linguistics, 2006.

Pratama, Bayu Yudha, and Riyanarto Sarno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM." In *Data and Software Engineering (ICoDSE), 2015 International Conference on*, pp. 170-174. IEEE, 2015.

Pervaz, Ifrah, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. "Identification of Author Personality Traits using Stylistic Features: Notebook for PAN at CLEF 2015." In *CLEF (Working Notes)*. 2015.

Pennebaker, James W., and Laura A. King. "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology* 77, no. 6 (1999): 1296.

Patra, Braja Gopal , Hiroya Takamura, Dipankar Das, Manabu Okumura and Sivaji Bandyopadhay.”Construction of Emotional Lexicon Using Potts Model.”In IJCNLP ,pp.674-679.2013.

Rangel F., Celli F., Rosso P., Potthast M., Stein B., Daelemans W. Overview of the 3rd Author Profiling Task at PAN 2015 . In: Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391.

Sidorov, Grigori, and Noé Alejandro Castro-Sánchez. "Automatic emotional personality description using linguistic data." *Research in computing science*20 (2006): 89-94.

Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29, no. 1 (2010): 24-54.

Torii, Yoshimitsu, Dipankar Das, Sivaji Bandyopadhyay, and Manabu Okumura. "Developing japanese wordnet affect for analyzing emotions." In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 80-86. Association for Computational Linguistics, 2011.

Tett, Robert P., Douglas N. Jackson, and Mitchell Rothstein."Personality measures as predictors of Job performance: a meta-analytic review." *Personnel psychology* 44, no. 4 (1991):703-742.

Wilson,Michael, “MRC psycholinguistic database: Machine usable dictionary,version 2.00” .Behavior Research Methods 20 .no. 1(1988): 6-10.