PERSONALITY RECOGNITION

CONTENTS

1.Problem Statment
2.Abstract
3.Dataset used
4.Our System
5.Evaluation Measures (with values)
6.References

1.Problem Statment:

"Personality Recognition" includes automatic classification of authors' personality traits, that can be compared against gold standard annotation obtained by means of the big5 personality test.

2.Abstract:

Here we are analysing the personality traits of authors, of given facebooks status data set(of 250 users) and our model is based on identification of:

- Style based features
- Sentimental Analysis
- Total no.of posts of author
- Avg no.of concepts author talking about in all his statuses(Concept Extraction)
- Social Networking Features
- Trigram based approach

(a)Introduction

For the natural and social interaction it is necessary to understand human behavior. Behavior involves an interaction between a person's underlying personality traits and situational variables. The situation, that a person finds himself or herself in, plays a major role on his or her reaction. However, in most of the cases, people respond with respect to their underlying personality traits , and gaining this insight of a web user's personality is very valuable for applications that rely on personalisation such as:

- Recommender Systems
- Personalized Advertising
- Online Marketing
- Sentiment Analysis/Opinion Mining
- Deception Detection
- Social Network Analysis
- and many others...

(b)Background information:

There are several theories for personality traits in the literature but the most widely used
personality traits model is the Big-5,.It describes the human personality as a vector of five values corresponding to bipolar traits. This is a popular model among the language and computer science researchers and it has been used as a framework for both personality traits identification and simulations.
The five big5 personality traits includes:
1. Extraversion: (x/e)(sociable vs shy)This trait includes characteristics such as excitability, sociability, talkativeness, assertiveness and high amounts of emotional expressiveness.
2. Neuroticism: (n)(neurotic vs calm)Individuals high in this trait tend to experienceemotional instability, anxiety, moodiness, irritability, and sadness.
3. Agreeableness:(a)(freindly vs uncooperative) This personality dimension includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors.
4. Conscientiousness:(c)(organized vs careless) Common features of this dimension include high levels of thoughtfulness, with good impulse control and goal-directed behaviors.
Those high in conscientiousness tend to be organized and mindful of details.
5. Openness: (o)(insightful vs unimaginative) This trait features characteristics such as imagination and insight, and those high in this trait also tend to have a broad range of interests.

Making use of the linguistic features associated with those classes, we generated different classifier for each class respectively.

3.Data Description:

corpus for Personality Recognition includes:

mypersonality.csv :

➢ includes authors, Facebook statuses in raw text, gold standard labels (both classes and scores) and several social network measures like network size, betweenness, density, brokerage etc...

➢ Texts have been originally collected by David Stillwell and Michal Kosinski, and anonymized by Fabio Celli.

➢ Each proper name of person has been replaced with a *PROPNAME* string. Famous names, such as "Chopin" and "Mozart", and locations, such as "New York" and "Mexico", have not been replaced.

Some more Statistics of Facebook Data(mypersonality.csv):

➢ The data was collected from 250 different users and the number of statuses per user ranges from 1 to 223.

➢ From the corpus analysis, it is observed that besides words, it contains tokens such as internet-slang (e.g. WTF-what the F***), emoticons (e.g., :-D), acronyms (e.g., BRB-be right back) and various shorthand notations that people use in their status.

➢ with splitting of 66%(train data) and 34%(test data) the statistics of mypersonality.csv are as follows:

• In total there are 6,545 train and 3,372 test instances after the split.
• The maximum number of tokens per user status message is 89,
• minimum 1 and
• the average is 14.

4.System Explanation:

we extracted the following features from given facebook dataset to identify the personality:

- ➢ Style based features
- ➢ Sentimental Analysis
- ➢ Total no.of posts of author
- ➢ Avg no.of concepts author talking about in all his statuses(Concept Extraction)

Style Based Features:

With the aim of modeling the style of writing we considered readability features as well as the use of emoticons. All these features are topic-independent. The complete set is described below. Each item is a list of individual features represented by frequencies and combined into a vector space model.

1.Frequency of Part-of-speech(all 36)

	Representation	Expansion
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential there
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	to
26	UH	Interjection
27	VB	Verb, base form

28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

2.Frequency of Special Symbols
3.Frequency of :
1PS – First Person Singular
1PP – First Person Plural
2P – Second Person
3PS – Third Person Singular
3PP – Third Person Plural
4.Frequency of Emoticons of:

- anger
- disgust
- fear
- happy
- sad
- surprise

5. Avg length of status
Punctuations count
unique words/total words
ratio of upper case words
ratio of upper case letters
6.Frequency of Health Related words

<u>Sentiment Analysis</u>
Extracted positive,negative and neutral percentage of emotions of each status update from "http://text-processing.com/api/sentiment/" and get average value of it for each user.
Basic Definitions:

Sentiment:

A thought, view, or attitude, especially one based mainly on emotion instead of reason

Sentiment Analysis:

aka opinion mining.Sentiment Analysis is the use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text

<u>Concept Extraction:</u>
From the linguistic aspect, we usually say that the main "building blocks" of a sentence are Noun Phrases (NP) and Verb Phrases (VP). The Noun Phrases are usually the topics or objects in the sentence, or in simple words – this is what the sentence is talking about, while Verb Phrases describe some action between the objects in the sentence. Take this example

"Facebook acquired Instagram"

About Who/What? – Facebook and Instagram > Noun Phrases
What happened? – acquired (=acquisition) > Verb Phrase
Here we extract only the Noun Phrases from the sentence and get average value of concepts that person talking about in his statuses. And for Identifying concept we define some simple patterns which describe the structure of a Noun Phrase, for example:
NN = content
JJ+NN = visual content
NN+NN = content marketing

<u>Social Networking Features:</u>

Given in facebook dataset and those include the following:

Network_size	Network size is the total number of people in the egocentric network including ego
betweenness	Ego betweenness centrality of an ego can be defined as the extent to which an ego lies between alters within the network (Freeman, 1979). Ego betweenness is high when alters are not well interconnected, and thus many of the shortest paths run trough ego.

n_betweenness	As ego betweenness is related to the size of the network, it should be normalized in order to allow for comparisons between egocentric networks of different size. Normalization used here involves dividing betweenness by number of all possible pairs between alters (this method is also employed in UCInet package) (add graph showing the relation between betweenness and size and normalized betweenness and size)
Density	Density indicates how many connections (edges) are there between alters as compared to the maximum possible number of edges. For an undirected egocentric graph it is calculated by dividing total number of (edges) by maximum possible number of edges. Density score here can be slightly different from one provided by UCInet as it is being calculated for the whole ego network including ego (as opposed to calculating density in the egocentric network with ego removed as it is being done in UCInet).
brokerage	Is the number of alters' pairs that are not directly connected
nbrokerage	As brokerage also depend on the size of the network, it is being normalized by dividing it by the number of all possible pairs between alters

After Extraction of all the above featueres ,we are going to represent the given user as a vector of above features and trained 5 different classifiers for different personality traits using Gaussian Naive Bayesian classification.(since its giving better results for given data set)

Classification

We use Gaussian Naive Bayesian model for the classification of the feature. For classification we divided facebook dataset in 80% for training and 20% for testing Since Bayesian classification doesn't remove any of the features while classifying (like SVM) by default, and as it works fine only with limited number of features, we try to reduce no.of dimensions of feature vector where we selected personality-trait related features using correlation coefficient clustering in removing similar/redundant features from the concept proposed in "**Feature Selection via Correlation Coefficient Clustering**" by Hui-Huang Hsu and Cheng-Wei Hsieh, where the concpet includes:

Feature Selection via Correlation Coefficient Clustering

For hundreds or even thousands of collected features, there must be features that are very similar to each other(where similarity is identified by the absolute value of correlation coefficient), and we can take these features as the same kind of features. We certainly do not need to use all features of the same kind for classification. After clustering analysis identifies all different kinds of features, we can remove a great number of redundant features. The classification performance in both the computational speed and the classification accuracy can be improved with the removal of these redundant features.And uses k-means algorithm with no.of clusters as 25 to cluster the features based on absolute value of correlation coefficent.

Correlation Coefficients of Extracted Features for different personality Traits:

PROPERTY	copn	ccon	cext	cagr	cneu
1PP	0.0765676333	0.0024962763	0.0865821817	0.0006958661	-0.06287539
1PS	-0.042727399	-0.0869117342	0.081004474	0.0114844889	-0.0373909172
2P	0.0864474833	-0.0667284689	0.0864049775	0.0264301706	-0.053730298
3PP	0.1187452534	-0.0771869153	0.0273905605	-0.0369374626	-0.0594025718
3PS	0.053688794	-0.12353863	0.0933200799	-0.0231670055	-0.0304637066
ANGER	0.030612265	-0.0487132439	-0.0019370147	0.071154881	0.0276102305
b/w ness	0.0419107427	0.1060812691	0.2532316047	0.0523779456	-0.1303435695
brokerage	0.041323686	0.1061971466	0.2542016247	0.0526302144	-0.1312660709
CC	0.0520406278	-0.1063682522	0.087256752	-0.0171750737	-0.0424592572
CD	-0.0180495594	-0.0994707081	0.0267306686	-0.0501898649	-0.0318629642
density	0.0483368352	-0.1400907146	-0.2359418364	-0.0812060088	0.0973538693
DISGUST	-0.0631428024	-0.0754652639	0.0192966298	-0.0517420296	0.0361415575
DT	0.0358769956	-0.1027211821	0.0728337	-0.0156466903	-0.0376174364
EX	0.0920291371	0.0503846238	0.1793638211	0.036673016	-0.0938964184
FEAR	0.0568323765	-0.0594097155	0.0352601884	-0.0277744302	0.0481020638
FW	0.0780611564	-0.0289157466	-0.0455446438	-0.0811098031	-0.0482464612
HAPPY	0.0228294719	-0.0820108655	0.0979441475	0.0482225307	-0.0912724888
HEALTH RELATED WORDS	0.0052773324	-0.0385351113	0.0713213034	-0.0075784158	-0.0811138305
IN	0.033432651	-0.0994488855	0.0942636228	-0.0102462329	-0.0521070006
JJ	0.0125711713	-0.0953005722	0.078718008	-0.005598396	-0.0431182757
JJR	0.0662112036	-0.1054451098	0.0428438818	0.0340062584	-0.0526460189
JJS	0.049157931	-0.1072203976	0.1386626164	0.0068421865	-0.0641642487
LENGTH	0.0417281795	-0.1128537683	0.0773080384	-0.0109471207	-0.0380238963
LS	-0.0090131528	0.0323535009	-0.0114809682	0.0288759032	0.0609931219
MD	0.0623142618	-0.1255910144	0.0489459307	-0.024604486	-0.0300459966
n/w size correlation	0.0167175972	0.1430358526	0.3124164713	0.0668664732	-0.1814658695
nb/wness	-0.0635901023	0.1203162255	0.2192584526	0.11158367	-0.0277160927
nbrokerage	-0.0136994709	0.0815277197	0.2280214796	0.0851394232	-0.0808022954
NN	0.0295063803	-0.1049940771	0.0812454626	-0.0218709744	-0.0372674639
NNP	0.0648428465	-0.1221394683	0.0246568624	-0.0195819924	-0.0116240011
NNPS	-0.0851717677	-0.1445800255	0.0905542284	-0.1137286702	0.0162682601
NNS	0.0638013373	-0.1018903102	0.0682996289	0.0267852758	-0.0497831625
PDT	0.0463588936	-0.0751189347	0.0992417086	-0.0570677134	-0.0406067403
POS	0.0160507205	-0.1081780939	0.0143192611	-0.0296024987	-0.0168358537
PRP	0.0246743427	-0.1176812082	0.0672158691	-0.0219780398	-0.0189912678
PRP$	0.0123323017	-0.0948793462	0.1079985388	0.0169833658	-0.0569759841
PUNCTUATIONS COUNT	0.0431773217	-0.1156986388	0.0904914726	0.0180306002	-0.0260750183
RATIO OF UPPER CASE WORDS	0.0445968019	-0.1144975113	-0.0519566376	0.0027958297	0.0891012572
RATION OF UPPER CASE LETTERS	-0.1257350474	0.07152381	-0.0113567738	0.0001339159	-0.0384564171
RB	0.0191876147	-0.1212586065	0.0551541039	-0.0197630852	-0.023707629
RBR	0.0119279266	-0.0698468523	0.0151938117	-0.059877445	0.0021206636
RBS	0.0824731726	-0.0422697862	0.0609386437	0.0369431122	-0.0788413851
RP	0.0241148588	-0.1399032029	0.0720970943	0.0238152906	-0.0376601794
SAD	-0.0917657456	-0.0656020125	0.118139837	-0.0423073	-0.0373819225
SPECIAL SYMBOLS	0.0764736723	-0.0726343181	0.0999895602	0.0532948863	-0.0332772799
SURPRISE	0.0879544689	-0.0729496187	0.0893655863	0.0466268269	-0.0928742299
TO	0.0255353435	-0.0771826442	0.1141719832	0.0065890174	-0.0761447342
transitivity	-0.0552321776	-0.0246108415	-0.2742833967	-0.1496594978	0.1407751977
UNIQUE WORDS/TOTAL WORDS	-0.0407354116	0.0171933275	-0.0783610534	-0.0420070475	0.0882929676
VB	0.0479296462	-0.0980190478	0.0948748304	-0.003506303	-0.0427368582
VBD	0.0088015404	-0.1439262317	0.0543322472	-0.0252929412	-0.0037719406
VBG	-0.0003154681	-0.109934354	0.0887792818	-0.0070106007	-0.0340588391
VBN	0.0615171588	-0.0973376049	0.0823843075	-0.0103006463	-0.0070024262
VBP	0.0458828552	-0.1285487244	0.0418489107	-0.0098428825	-0.011634284
VBZ	0.086544498	-0.1104677139	0.0626385932	-0.0101093146	-0.0267640737
WB	-0.0160720444	-0.0716491813	0.1001339083	0.0292381945	-0.0659605174
WDT	0.0338934546	-0.0951735194	0.0807800802	-0.0522309194	-0.0026902543
WP	0.1282110027	-0.08055813	0.025212129	0.0606404692	-0.0812167759
WP$	-0.0700775152	0.0460351798	0.1448749774	-0.0637267325	-0.0735329236

Trigram based approach:

This approach includes the generation of two features (say F1 and F2)for each user status,Where "F1" represents normalized frequency of trigrams w.r.t to current personality trait and "F2" represents normalized frequency of trigrams w.r.t remaining classes

Proceducre:

Step-1:

Identify all possible trigrams w.r.t to individual personality trait

step-2:

Iterate through individual status , identify all trigrams of respective status and compare it with the trigram set of "respective personality trait" and "trigrams sets of all the remaining personality traits" and increament the count F1 and F2 accordingly.

Step-3:

Normalize F1 and F2 (by dividing it with the count of trigrams in respective status)

Step-4:

Represent the given dataset in feature vector form (F1,F2) and train each personality trait to different classifier using SVM to produce 5 different classifiers.

5.Evaluation measures:

In the shared task guidelines it is suggeste to use precision,recall,FI as evaluation metrics.

Explanation:

To calculate precision,recall and F-Score requires confusion matrix which is explained as follows:

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

		Predicted	
		Negative	Positive
Actual	Negative	**a**	**b**
	positive	**c**	**d**

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect of predictions that an instance negative, and
- d is the number of correct predictions that an instance is positive.

Several standard terms have been defined for the 2 class matrix:

- The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

  AC=(a+d)/(a+b+c+d)

- The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

  True Positive Rate(TP) or recall=d/c+d

- The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated usingthe equation:

  False Positive Rate(FP)= b/a+b

- The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

  True Negative Rate(TN)= a/a+b

- The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

False Negative Rate=c/c+d

- Finally, precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

Precision=d/b+d

- **F-score=2 (precision*recall/(precision+recall))**

Obtained Values:

Personality Trait	a	b	c	d
Extraversion	31	0	13	5
Openness	0	15	0	34
Neuroticism	15	8	11	15
Agreeableness	9	16	4	20
Conscientiousness	2	22	0	25

Calculate Measures

Personality Trait	Accuracy	True Positive Rat(TP) Recall	False positive Rate(FP)	True Negative Rate(TN)	False Negative Rate	Precision	F-score	Trigram Accuracy
Extroversion	**74%**	0.28	0	1	0.72	1	0.44	41.17%
Openness	70	1	1	0	0	0.695	0.82	70.58%
Neuroticism	**62%**	0.577	0.348	0.652	0.407	0.652	0.613	43.13%
Agreeableness	**60%**	0.833	0.64	0.36	0.166	0.5555	0.667	58.82%
Conscientiousness	**56%**	1	0.9166	0.0833	0	0.532	0.695	50.98%

6.References:

1. http://clic.cimec.unitn.it/fabio/wcpr13/celli_wcpr13.pdf
2. http://clic.cimec.unitn.it/fabio/wcpr13/verhoeven_wcpr13.pdf
3. http://clic.cimec.unitn.it/fabio/wcpr13/farnadi_wcpr13.pdf
4. http://clic.cimec.unitn.it/fabio/wcpr13/tomlinson_wcpr13.pdf
5. http://clic.cimec.unitn.it/fabio/wcpr13/markovikj_wcpr13.pdf6.
6. http://clic.cimec.unitn.it/fabio/wcpr13/alam_wcpr13.pdf
7. http://clic.cimec.unitn.it/fabio/wcpr13/mohammad_wcpr13.pdf
8. http://clic.cimec.unitn.it/fabio/wcpr13/appling_wcpr13.pdf
9. http://clic.cimec.unitn.it/fabio/wcpr13/iacobelli_wcpr13.pdf
10. http://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CDkQFjAC&url=http%3A%2F

%2Fojs.academypublisher.com%2Findex.php%2Fjsw%2Farticle%2Fdownload%2F051213711377%2F2428&ei=vLUwU4HtOoSWrAecoICQDw&usg=AFQjCNFWkR3tZQqEHaJ5_xhsaGZtfago7w&sig2=7kAEW2TGxGpidRxXtzDu2w

11. http://ceur-ws.org/Vol-1096/paper3.pdf

Team Members:

201001118 - DHWANIT GUPTA
201101020 - ARPIT SHARMA
201305518 - YADAVALLI SINDHUSHA
201307687 – CHARUDATT PACHORKAR