# Identifying effects of Personality in Gender classification

Amit Prasad
17318406
prasada@tcd.ie

Samim Samsul Ekram
17314678
samsules@tcd.ie

Harshith Patel
17318631
patelh@tcd.ie

Nabanita Roy
17305618
royn@tcd.ie

**Abstract**

There have been various works that attempt to identify the gender of authors from texts by exploiting the relationship between different linguistic categories that gender can be related to. Most of these works consider corpora comprising of blogs, news articles or excerpts from books which have been thoroughly edited and revised to give away the story that it wants to convey in textual format. However, the corpus we consider for this research is transcripts of verbal conversations from a multimodal corpus which is devoid of major emotional features and bodily gestures that are helpful to identify the sentient of the participant. Our research aims to identify the effects of personality based features in gender classification using transcripts of conversations. We experimented with this corpus using different machine learning techniques. Our findings indicate that personality features have a positive effect in increasing the accuracy of gender classification although by a smaller margin because of the nature of the MULTISIMO dataset.

***Keywords*—** Gender Classification, Big Five Personality Traits, Multimodal, Weka, Author Profiling

## 1 Introduction

Thorough research has been conducted in the past to identify gender from text. These studies have identified different challenges in processing text for gender identification like short content, improper use of grammar, incomplete sentences, formal or casual text, etc. Additionally, works in the past have used different supporting features like age, region, user experience with the language, etc to improve the identification results. These researches are mainly conducted on non-conversational textual data, procured from written materials like blogs, supervised assignments and tweets which are most often thoughtfully constructed to give away useful information. On the other hand, key indicators of emotions in conversations are simultaneously conveyed using non-verbal expressions like intonations, gestures and facial expressions which are sparsely captured during transcription for textual analysis. In context of this fundamental difference in non-conversational and conversational discourse collections, our research work focuses on a corpus of transcribed conversations conducted specifically to capture all the modalities involved in conversations. While it is true that past research has proved the feasibility of predicting author characteristics using only textual information from non-conversational perspective, our work will focus on exploiting information on personality traits and identifying its effect on gender classification task using the transcripts of group conversations.

In rest of the paper, we first discuss related work in identification of gender from text in different dimensions. In section 4 we describe our approach of applying personality traits to the corpus under consideration. Next, in section 5 we discuss our analysis and the effects of personality on the classification accuracy.

## 2 Related Background

With growing complexity of the Internet and its abuse, it has become increasingly important to analyze text and identify its author. Such tasks are termed as Author Profiling. Two fundamental steps towards author profiling are gender and personality identification. Since our research is aligned towards personality differences based on gender, we have reviewed papers which addressed gender and personality identification separately as well as those which combined them.

## 2.1 Gender Classification

The task of gender identification has been thoroughly researched in the past using different techniques. With the advent of powerful machines, new machine learning algorithms are being applied for this task. In an article, Cheng et al. (2011) discusses an interesting approach using ML techniques for gender identification. The authors have identified and classified five sets of gender-related features - character based, word based, syntactic, structure based and function words. These types of features can further include several other features. Character based features include 29 stylometric features that can be adopted from Linguistic Inquiry And Word Count (LIWC) (Pennebaker et al., 2015). Extracted features can then be classified using algorithms like - Bayesian based logistic regression, decision trees and SVM, similar to work by Cheng et al. (2011). Several works have indicated that SVM delivers higher accuracy compared to other Models, although the training time required is much higher.

Newman et al. (2008) conducted a comprehensive study of gender differences in language use which involved analysis of 14000 text samples. The two methodologies used were LIWC and extended version of LIWC with large archival of text samples. Final results yielded a multivariate effect highlighting the gender differences (Newman et al., 2008). It was observed that women preferred words related to psychological and social process whereas men preferred object properties and impersonal topics. Besides, the outcomes of this research are essential for upcoming projects since it was experimented on a large corpus unlike other researches with significantly small number of samples (Newman et al., 2008).

Another interesting research work was carried out by Zhang and Zhang (2010). Their data set includes blog posts from many blog hosting sites and blog search engines, e.g., blogger.com, technorati.com, etc. Their dataset includes topics that are truly diverse and general. They used the tokenizer that comes with the part of speech(POS) tagger for the words and punctuations that people generally uses. A similar research was carried out by Isbister et al. (2017) where they collected a set of blogs. Blogs texts include several features like irregular punctuation and grammatical errors. The blogs touch various topics, such as, politics, sports, personal writings, traveling and so forth. The authors perform classification using Support Vector Machine(SVM) algorithm and k fold cross validation. The accuracy when using language independent features are well over 73% for all different languages using data independent features. (Isbister et al., 2017).

## 2.2 Personality - The Big Five Personality Model

Our work is not directly related to personality identification, however, we reviewed papers which attempted to model personality types to consider the inferences derived from them. This section is important to us because we also have data related to personality of each of the participants which we intend to use in our research work.

Personality has behavioral, temporal, emotional and mental attributes that define an individual. A standard model of classifying personality traits is the ”Big 5” or the Five factor Model(BFI) where the personality of an individual can be classified into five dimensions - Extraversion vs Introversion (e.g., outgoing, talkative, active), Emotional stability vs Neuroticism (e.g., anxious, depressive, touchy), Agreeable vs Disagreeable (e.g., trusting, kind, generous), Conscientiousness vs Unconscientiousness (e.g., self-controlled, responsible, thorough), and Openness (e.g., intellectual, artistic, insightful) (McCrae and John, 1992). Personality identification has interesting use-cases like identifying leaders, terrorists, profile matching on dating websites and tutoring systems.

Several works in the past have explored the task of personality recognition to relate properties of texts. Mairesse et al. (2007) explores the task of recognition of all Big Five personality traits in texts as well as conversations. The authors of this work have conducted an extensive survey of methods for identifying emotion, deception, speaker charisma, dominance in meetings, point of view of subjectivity, and extracting sentiment from text and conversations. Interesting markers have been composed by the authors. One such finding identifies that Extraverts talk more, talk louder, repeat words, use less formal language and more positive emotion words. Qualitative analysis helped to discover relationships and identify interesting patterns like, conscientious people use fewer swear words, content related to sexuality.

A large scale research on personality analysis was conducted by Yarkoni (2010) using 100,000 words collected from blogs. The aim of the research was to establish a pervasive correlation between personalities to be identified by a broad range of lexical verbs.Nearly 5,000 bloggers were invited to participate and 20% of them agreed to take the Big Five Inventory[1](John et al., 1991) Test. The advantage of using blog texts was to ensure that the content is not research-conscious as well as they cover a broad spectrum of topics. This is slightly different from the LIWC approach. Their approach is based on LIWC 2001 but keeps the data unstemmed as previous studies revealed that same stem words has different pattern of correlation with personality(Yarkoni, 2010). Category based analysis revealed strong correlation between big five traits and bloggers word frequency. (Yarkoni, 2010)

[1]https://www.ocf.berkeley.edu/ johnlab/bfi.htm

### 2.3 The Relationship between Gender and Personality

Like Yarkoni (2010) another wide-ranging research was carried out by Schwartz et al. (2013) 15.4 million Facebook messages were collected from 75 thousand volunteers, who also took standard personality tests for thorough analysis. The authors compared a popular and traditional method of closed vocabulary approach using LIWC (Pennebaker et al., 2015) against open vocabulary approach to analyze the relationship between language usage, personality and gender. The research did not only bolster the findings of prior research in this domain but also opened door to new relationship discoveries. This predictive model also indicates that open-vocabulary models are more insightful than traditional closed-vocabulary analysis. [2].

A cross-cultural research was conducted by Schmitt et al. (2008) across 55 nations. They used the standard BFI to examine personality traits and relate them cumulatively to gender categories. This multinational research was conducted across all continents to avoid cultural bias. Their objective was to investigate the possible reasons for the widening gap between male and female personality traits in egalitarian countries. A recent research focused on Russian high school students conducted by Ismatullina and Voronin (2017) demonstrates high BFI in females for all five traits. It is to be noted that this study is limited to high school students in Russia which signifies an effect of age and cultural bias as contributing factors. They further delve into the structure of the relationships among the Big Five personality traits in girls and boys and conclude that, for females, neuroticism showed negative correlation with extroversion while for males, it showed negative correlation with agreeableness.

## 3 The Corpus

For our research we use a fresh set of data that was acquired recently at Trinity College Dublin called the MULTISIMO Multimodal Corpus. It was designed to examine specifically the collaborative aspects in task-based group interactions by considering a speaker's multimodal signals and psychological variables(Koutsombogera and Vogel, 2017). This corpus aligns with our research goal as it encompasses informal conversational transcripts as well as BFI indexes of the speakers. In our research, we shall overlook the multimodal aspect and investigate the transcripts and the BFI collected from each of the participants in order to examine the gender-personality relation from conversation-based corpus.

### 3.1 Data Curation Process

For this corpus, two randomly grouped participants were required to collaborate with each other to find the three most popular answers to each of three questions, guided by a facilitator. This conversation, based on a game-like task, was audio and video recorded from different angles to capture a complete set of gestures and behavioral aspects of the speakers. Motions and gestures of participants were captured using Kinect 2 sensors[3]. Prior to the task, the participants also completed answering the BFI and after the task was completed, they took an experience assessment survey as well. Additionally, the collected data also specifies if the participants had known each other and tracks the gender distributions of each session.

### 3.2 Data Summary

The data collection for this corpus consisted of 23 sessions of a total duration of approximately 4 hours (Koutsombogera and Vogel, 2017). Some vital information about the corpus and the participants are summarized as below:

- average session duration - 10 minutes
- Average age - 30 years
- Gender distribution - 25 female and 24 male
- Number of Nationalities - 18
- Number of native English speakers - 3
- Language used for conversation - English

[2]https://en.wikipedia.org/wiki/Thumb_tribe
[3]https://en.wikipedia.org/wiki/Kinect

# 4 Research Methods

## 4.1 Data Pre-processing

### 4.1.1 Keywords Extraction

The transcripts of the corpus were first used to extract keywords and their frequencies using ELAN [4]. Next, they were merged and sorted corresponding to each participant across all sessions. The frequencies are preserved with repetition of the words in the processed dataset.

### 4.1.2 Encoding BFI scores

The original BFI scores is continuous, ranging from 0 to 50. They were binary encoded to 1 when the score is more than 25 and 0 when less than or equal to 25. However, we preserved the original data in order to compare the classification accuracy of continuous versus categorical data.

### 4.1.3 Linguistic Features Extraction

The combined transcripts for each of the participants were used to track the occurrences of linguistic categories, namely, nouns, verbs, personal pronouns, wh-words, interjections and conjunctions. We cleaned the dataset and applied the Stanford POS Tagger [5] to accomplish this task. Additionally, we collected a cumulative sentiment score for each participant using Text Blob [6].

### 4.1.4 Complete Dataset Accumulation

The data was further made noise-free and combined with other information such as gender and age to craft a fully pre-processed data. The final dataset looks like fig 1.

### 4.1.5 Classification

On the preprocessed data we used weka Eibe et al. (2016) to classify the MULTISIMO dataset. We used Random Forest, Random Tree, Naïve Bayes, Naïve Bayes Multinomial Updateable algorithms to classify the data into male and female categories. The scoring metrics used are Accuracy, Root mean squared error and Kappa statistics. We used the algorithms on three subsets of our final dataset.

**Dataset 1:** All columns except BFI data (encoded as well as continuous)

**Dataset 2:** All columns except BFI encoded data

**Dataset 3:** All columns except BFI continuous data

### 4.1.6 Word Cloud Generation

To visualize and compare the words used by each of the participants we constructed word-clouds. We grouped the keywords based on the binary encoded BFI scores and gender and consequently for each BF personality trait four word-clouds are generated as:

1. BF personality trait score: 0 and gender: male
2. BF personality trait score: 0 and gender: female
3. BF personality trait score: 1 and gender: male
4. BF personality trait score: 1 and gender: female

[4] https://tla.mpi.nl/tools/tla-tools/elan/

[5] https://nlp.stanford.edu/software/tagger.html

[6] http://textblob.readthedocs.io/en/dev/

## 4.2 Final Dataset

In the final dataset we have ID- is the participant ID, Gender describes Participants Gender, Sentiment describes the Sentiment Score ranging from -1 to 1 , Nouns describes the total Nouns spoken by a participant,Verbs describes the total Verbs spoken by a participant,Personal Pronouns describes the total number of Pronouns which points to a person spoken by a participant,wh-words describe the wh-questions spoken by a participant,Interjections describes the total Interjections spoken by a participant,Conjunctions describes the total Conjunctions spoken by a participant,Gender_en encodes the Gender in 1 for Female,EXT describes the Extraversion aspect of an participant,AGR describes the Agreeableness aspect of an participant,CONSC describes the Conscientiousness aspect of an participant,NEURO describes the Neuroticism aspect of an participant,OPN describes the Openness aspect of an participant,Keywords are the words spoken by the participant,EXT_en describes the Binary Score of Extraversion aspect of an participant,AGR_en describes the Binary Score of Agreeableness aspect of an participant,CONSC_en describes the Binary Score of Conscientiousness aspect of an participant,NEURO_en describes the Binary Score of Neuroticism aspect of an participant,OPN_en describes the Binary Score of Openness aspect of an participant

| ID | GENDER | AGE | Sentiment | Nouns | Verbs | Personal Pronouns | wh-Words | Interjections | Conjunctions | Gender_en | EXT | AGR | CONSC | NEURO | OPN | Keywords | EXT_en | AGR_en | CONSC_en | NEURO_en | OPN_en |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P004 | F | 24 | 0.18125 | 338 | 139 | 3 | 11 | 10 | 1 | 1 | 23 | 37 | 26 | 22 | 37 | and or A Ah | 0 | 1 | 1 | 0 | 1 |
| P005 | M | 23 | 0.063392857 | 371 | 156 | 5 | 12 | 11 | 1 | 0 | 24 | 39 | 39 | 23 | 38 | A A Ah And A | 0 | 1 | 1 | 0 | 1 |
| P006 | F | 23 | 0.115595238 | 421 | 176 | 6 | 13 | 12 | 2 | 1 | 24 | 44 | 36 | 19 | 34 | A Ah Alright | 0 | 1 | 1 | 0 | 1 |
| P007 | F | 24 | 0.142036125 | 478 | 201 | 7 | 17 | 13 | 2 | 1 | 20 | 32 | 31 | 26 | 47 | A Are Basica | 0 | 1 | 1 | 1 | 1 |

Table 1: Final Dataset

# 5 Results

## 5.1 A Comparison of Gender Classification Algorithms on Sub-Datsets

Table 1 is a compilation of the results we obtained by using classification algorithms on the three sub-datasets enlisted in section 4.1.6.

| Algorithm Used | Scoring Metric | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|---|
| Random Forest | Accuracy (%) | 63.2653 | 67.3469 | 61.2245 |
| | RMSE | 0.5059 | 0.4757 | 0.5125 |
| | Kappa statistic | 0.265 | 0.3445 | 0.2209 |
| Random Tree | Accuracy (%) | 65.3061 | 63.2653 | 71.4286 |
| | RMSE | 0.589 | 0.6061 | 0.5345 |
| | Kappa statistic | 0.3018 | 0.2601 | 0.4255 |
| Naïve Bayes | Accuracy (%) | 65.3061 | 65.3061 | 61.2245 |
| | RMSE | 0.5444 | 0.5513 | 0.5685 |
| | Kappa statistic | 0.3041 | 0.3053 | 0.2235 |
| NaiveBayes Multinomial Updateable | Accuracy (%) | 57.1429 | 63.2653 | 63.2653 |
| | RMSE | 0.5937 | 0.5919 | 0.4793 |
| | Kappa statistic | 0.1418 | 0.2638 | 0.2674 |

Table 2: Compiled Table of Results

By comparing the typical classification algorithms such as Random Forest, Random Tree, Naïve Bayes, Naïve Bayes Multinomial we experimented with our sub-datasets to find the best fit. It was observed that adding personality based information in the form of Big Five indexes improved the classification accuracy of these algorithms. Random forest performed the best on Dataset 2 with an accuracy of 67.3469%. On the other hand, the best performance was exhibited by Random Tree on Dataset 3 but all other classifiers did not perform as well on this dataset. It is our opinion that this major increase in accuracy could be due to over-fitting. We used k-fold cross validation using the built-in functionality in Weka (Eibe et al., 2016) to avoid over-fitting. This issue of over-fitting observed in the results might be an error in their current framework.

## 5.2 Word-cloud Analysis

The words that appeared on the word-clouds were mostly interjections and conjunctions. Besides words like 'Airplane', 'Hospital', 'Hair', 'Guitar', 'Drum' and 'Bread' occurred frequently. The appearances could not be differentiated on the basis of gender or personality traits. The best possible reason for their uniform occurrences

is because they were probably answers of the questions asked. Apart from that, our analysis of Word cloud


Figure 1: Example of Word-clouds generated for males and females having high 'Agreeableness' respectively

suggests that Males as well as Females who exhibited a high score on 'Agreeableness' excessively used neutral vowel sounds and confirmatory words such as 'alright', 'absolutely'. Conscientious females have used words like 'Alright', 'Good' to show affirmation as well. It was also observed that Extravert males laughed a lot during the conversations. Males who scored high on neuroticism have also used negative words like 'No' and Gross' and have no occurrence of words like absolutely and good. In contrast to that, males who scored high on 'Openness' used positive words like 'alright','absolutely', 'good' and laughed a lot. Therefore, we see good correlation between the choice of words between participants who scored high on 'Openness' and 'Extraversion'.

# 6 Conclusions

The dataset that we worked on required a considerable amount of pre-processing because of the transcription choices that were made beforeKoutsombogera and Vogel (2017) and also that the nature of the corpus. It is evident that corpus with conversations are hard to deal with because, we have already mentioned, much of the information is embedded in their gestures. Therefore, one potential reason for barely 60 to 70 % is because the dataset does not embed the informations that a carefully crafted blog post or an article might contain. The second reason for the same is the number of data-points that we have is quite small for any classification algorithm to produce optimal results. Therefore, even with multiple features, the average accuracy obtained is 63%. Besides, while SVMs are hailed as one of the best binary classifiers, it achieved less than 60% accuracy and, therefore, could not perform well on this dataset.

The word-clouds generated also failed to draw significant inferences on gender-personality relation. The main reason for this is the skewed usage of words as a result of recording responses to questions.Thus certain keywords appeared more for females as well as males when grouped by each of the traits of the Big Five model. The only conclusion that could be drawn was that males laughed more during the conversations and used the word 'Drum' while females laughed less and scarcely used the word 'Drum'. These two were the only two stark differences noticed on male-female comparison without considering personality variations. A minute look on the word-clouds generated based on gender and personality yielded the results enlisted in section 5.2.

Overall, we observed in the results summarized in section 5.1, the addition of BFI helped to improve the predictions except for in Random Tree which boosted the performance on adding the binary encoded BFI. This proves that personality traits have an impact on gender classification. However, it is hard to comment on the extent of impact personality traits has on gender classification because of the issues with the dataset summarized in the first paragraph.

To conclude, we would like to make some future work suggestions. Firstly, to improve the accuracy we should pre-process the data with better tools such as LIWC and isolate the keywords that have a greater impact in the classification of male and female genders. A weighted keyword analysis approach could also be incorporated further. Besides, since data related to each session are also available, we could look into each session and process them

individually for micro-analysis, which in our opinion fits more into the nature of the dataset. We generated word-clouds to classify individuals based on their personality traits and gender. In future, we propose to use this data in our experiments to improve the performance of the predictions. Since the dataset we have considered is small but has very crucial information that focuses more on personality indexes and interactions between participants, we would like to integrate similar data to build human-interaction models that would help to focus on interactive data of participants and classify better.

# 7 Acknowledgements

We would like to thank Dr. Maria Koutsombogera for providing us with the MULTISIMO multimodal corpus and necessary resources related to it. Additionally, we would also like to thank Prof. Dr. Carl Vogel and Carmen Klaussner for guiding us through the research throughout the course.

# References

Cheng, N., R. Chandramouli, and K. Subbalakshmi (2011). Author gender identification from text. *Digital Investigation 8*(1), 78–88.

Eibe, F., M. Hall, I. Witten, and J. Pal (2016). The weka workbench. *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques 4*.

Isbister, T., L. Kaati, and K. Cohen (2017). Gender classification with data independent features in multiple languages. In *Intelligence and Security Informatics Conference (EISIC), 2017 European*, pp. 54–60. IEEE.

Ismatullina, V. and I. Voronin (2017). Gender differences in the relationships between big five personality traits and intelligence. *Procedia-Social and Behavioral Sciences 237*, 638–642.

John, O. P., E. M. Donahue, and R. L. Kentle (1991). The big five inventoryversions 4a and 54.

Koutsombogera, M. and C. Vogel (2017). The multisimo multimodal corpus of collaborative interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 502–503. ACM.

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

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

Newman, M. L., C. J. Groom, L. D. Handelman, and J. W. Pennebaker (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes 45*(3), 211–236.

Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn (2015). The development and psychometric properties of liwc2015. Technical report.

Schmitt, D. P., A. Realo, M. Voracek, and J. Allik (2008). Why can't a man be more like a woman? sex differences in big five personality traits across 55 cultures. *Journal of personality and social psychology 94*(1), 168.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one 8*(9), e73791.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality 44*(3), 363–373.

Zhang, C. and P. Zhang (2010). Predicting gender from blog posts. *University of Massachussetts Amherst, USA*.