MEDIATOR
An Explainable Multi-Aspect
User Profiling Online Platform

by Anastasios Louvoulinas

A thesis submitted in partial fulfillment for the Postgraduate degree

in the Faculty of Sciences
School of Informatics

Supervising Professor: Dr. Athena Vakali

Thessaloniki, September 2021
Declaration of Authorship

I, Anastasios Louvoulinas, declare that this thesis titled, ‘Mediator: An explainable Multi-aspect user profiling online platform’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: __________________________________________

Date: __________________________________________

ARISTOTLE UNIVERSITY OF THESSALONIKI
Abstract
Faculty of Sciences School of Informatics
Undergraduate Degree

by Anastasios Louvoulinas - 38

Nowadays, social media have a daily high impact in the lives of individuals and their evolution is at unprecedented levels, creating new opportunities for extracting and collecting essential knowledge about users. There is no doubt that this new era of social media provides a wealth of data that can reveal interesting information about individuals extracting aspects of users' personalities and preferences after proper analysis. For instance, the number of an individual’s followers could possibly indicate whether a person is extroverted or introverted, while their comments on certain posts can reveal whether an individual is conservative or open to new ideas.

On the one hand, many people participate in social media generating plenty of data which can be processed directly to extract valuable information. For instance, users can explicitly include information about their selves in their social media profile such as age and gender. However, sometimes people do not include public-available information about themselves, such as personality and psycho-demographics aspects in their posts or in their profiles since they may want to preserve their privacy. On the other hand, in some cases, the explicit information provided by the individuals may be fake or incomplete. Furthermore, some Social Networks (SNs) are not directly support explicit personal data, for example twitter accounts do not contain information about gender. That causes many challenges in user profiling analysis that may lead to inaccurate results. Hence, there is a need to extract implicit information about the users based on their activity in social media networks which is not provided directly by the users.

These implicit data can be extracted after properly analyzing available information in the user’s account, also referred to as User Profiling. Building an automated user profile is the main challenge in developing an adaptive personalized application for a wide spectrum of purposes. Research in personalization is used in several applications such as recommendation, e-learning, and personalized information retrieval, etc. In particular, User Profiling is defined as the scientific field which extracts useful information explicit, implicit or in combination (hybrid) and it can be very useful in a wide variety of use cases such as advertising since users’ demographics affect their purchasing choices, employee hiring in order to match specific candidates traits with position’s requirements, marketing as consumer/customer profiling is the only way to gather the insights needed to identify, segment and define the target audience for a specific product etc.

This work is motivated by the necessity to analyze the psycho-demographic traits of social media users, in order to provide accurate analysis for users’ without their direct input of data. Towards, this direction, the purpose of this work is to develop an online, open-source, explainable user profiling platform. The literature review revealed a gap in the existing
User Profiling approaches. In particular, the existing user profiling methodologies focus only on a few demographics or personality aspects, and thus, they do not export a complete and accurate user profile creation. On top of that, they do not provide explanations for their models, rather than that their predictive models are often thought of as black boxes that are impossible to interpret. It is crucial to understand that, in the end, these models will be used by humans who need to trust them, understand the errors they make, and the reasoning behind their predictions, since the right to Explainability emerges as a high priority of personalized services.

The structure of this thesis is as follows: *Chapter 1* introduces the problem statements, gives introductory information about social media user profiling, discusses the challenges and motivation in social media user profiling, and finally highlights the contribution of the proposed work. *Chapter 2* provides a systematic literature review on user profiling methods and techniques in social networks, describes the preliminaries of the existing research in order to make our work more comprehensible to the reader, discusses data ethics issues in user profiling, and concludes highlighting the research gap on current user profiling methods in social networks. *Chapter 3* presents, in detail, the methodology and the phases of the proposed approach, as well as the workflow and the developed system architecture. On top of that, it describes in detail all data science pipeline steps followed until the final results (i.e., data preprocessing, feature extraction), as well as the machine learning models and architectures that we designed and developed to predict the psycho-demographic aspects about a Twitter user. In *Chapter 4*, we present the experimental procedure followed in order to evaluate the performance of the Mediator platform. Furthermore, it provides a comprehensive presentation of the extracted results for every psycho-demographic aspect, and the explainability results over the model predictions that were extracted from a Twitter account through our platform. *Chapter 5* discusses the technologies that we used to develop our platform. In particular, we discuss the use case scenario, the system requirements and architecture of the developed platform, and the platform implementation in detail, i.e., backend architecture, explainable component and user interface - frontend. Finally, in *Chapter 6*, we provide our main conclusions, accompanied by future work suggestions.
Εκτεταμένη Περίληψη

Στις μέρες μας, τα μέσα κοινωνικής δικτύωσης έχουν υψηλό αντίκτυπο στη καθημερινή ζωή των σύγχρονων ατόμων και η εξέλιξη τους, σε προτοφανή επίπεδα, δημιουργεί νέες ευκαιρίες για την εξαγωγή και συλλογή βασικών γνώσεων για τους χρήστες. Δεν υπάρχει αμφιβολία ότι αυτή η νέα εποχή των μέσων κοινωνικής δικτύωσης παρέχει πληθώρα δεδομένων που μπορούν να αποκαλύψουν ενδιαφέρουσες πληροφορίες για ατόμων εξάγοντας δεδομένα για πτυχές της προσωπικότητάς τους και των προτιμήσεών τους μετά από σωστή ανάλυση. Για παράδειγμα, ο αριθμός δεδομένων που μπορεί να αναφέρει ένα άτομο μπορεί να υποδηλώνει αν είναι εξωστρεφές ή εσωστρεφές, ενώ τα σχόλια τους σε ορισμένες αναρτήσεις μπορούν να αποκαλύψουν εάν ένα άτομο είναι συντηρητικό ή ανοιχτό σε νέες ιδέες.

Από τη μία πλευρά, πολλοί άνθρωποι συμμετέχουν στα μέσα κοινωνικής δικτύωσης κοινοποιώντας πολλά προσωπικά δεδομένα τα οποία μπορούν να υποβληθούν σε άμεση επεξεργασία για την εξαγωγή πολύτιμων πληροφοριών. Για παράδειγμα, χρήστες μπορούν να κοινοποιήσουν πληροφορίες σχετικές με την ηλικία και το φύλο, ή να αναφέρονται στις ηλικίες και το φύλο σε ορισμένες αναρτήσεις. Ωστόσο, μερικές φορές οι χρήστες δεν περιλαμβάνουν δημόσιες πληροφορίες σχετικές με την προσωπικότητά τους και τις ψυχοδημογραφικές πτυχές τους στις αναρτήσεις τους και στα προφίλ τους, καθώς μπορεί να ταθέλουν να διατηρήσουν το απόρρητό τους. Από την άλλη πλευρά, οι πληροφορίες που παρέχονται από τα άτομα μπορεί να είναι ψεύτικες ή ελλιπείς, επιπλέον, ορισμένα κοινωνικά δίκτυα δεν υποστηρίζουν άμεσα ρητά προσωπικά δεδομένα, για παράδειγμα οι λογαριασμοί στο Twitter δεν περιέχουν πληροφορίες σχετικές με το φύλο. Αυτό προκαλεί πολλές προκλήσεις στην ανάλυση προφίλ χρήστη που μπορεί να οδηγήσει σε ανακριβή αποτελέσματα. Ως εκ τούτου, υπάρχει ανάγκη εξαγωγής πληροφοριών σχετικά με τους χρήστες με βάση τη δραστηριότητά τους στα δίκτυα κοινωνικών μέσων, η οποία δεν παρέχεται απευθείας από τους χρήστες.

Αυτά τα δεδομένα μπορούν να εξαχθούν μετά από σωστή ανάλυση των διαθέσιμων πληροφοριών στο λογαριασμό του χρήστη, που αναφέρεται ως User Profiling, το οποίο μεταφράζεται ως ‘προφίλ χρήστη’. Η δημιουργία ενός αυτοματοποιημένου προφίλ χρήστη είναι η κύρια πρόκληση στην ανάπτυξη μιας προσαρμοζόμενης εφαρμογής για ένα ευρύ φάσμα ακολουθιών, όπως η διαφήμιση, ηλεκτρονική μάθηση και εξατομικευμένη ανάκτηση πληροφοριών, κ.λπ. Ειδικότερα, το User Profiling ορίζεται ως το επιστημονικό πεδίο που εξάγει χρήσιμες πληροφορίες από το προφίλ του χρήστη σε κάποιο κοινωνικό δίκτυο και μπορεί να είναι πολύ χρήσιμο σε μια μεγάλη ποικιλία περιπτώσεων χρήσης, όπως η διαφήμιση, καθώς το δημογραφικό στοιχείο των χρηστών επηρεάζει τις αγοραστικές επιλογές τους, η πρόσληψη επαγγελμάτων και η διαφήμιση, καθώς η διαφήμιση, καθώς το μάρκετινγκ ως προφίλ καταναλωτών/πελατών είναι ο μόνος τρόπος για να συγκεντρωθούν οι απαραίτητες γνώσεις για τον προσδιορισμό, την τμηματοποίηση και τον ορισμό του κοινού -στόχου για ένα συγκεκριμένο προϊόν κ.λπ.

Η μεταπτυχιακή διπλωματική αυτή εργασία παρακινείται από την ανάγκη να αναλυθούν όχι μόνο τα χαρακτηριστικά της προσωπικότητας των χρηστών των κοινωνικών δικτύων, αλλά

Από την υπόλοιπη μέρη, παλαμιές και δημιουργίες κοινωνικών δικτύων.
και τα ψυχοδημογραφικά τους στοιχεία, προκειμένου να γίνει ακριβής ανάλυση για τις ψυχοδημογραφικές πτυχές των χρηστών χωρίς την άμεση εισαγωγή δεδομένων. Οι εξαγόμενες πληροφορίες μπορούν να χρησιμοποιηθούν σε μια μεγάλη ποικιλία περιπτώσεων χρήσης όπως αναφέρθηκε προηγουμένως. Προς αυτήν την κατεύθυνση, ο σκοπός αυτής της εργασίας είναι η ανάπτυξη μιας διαδικτυακής, επεξηγημένης και Ανοιχτού και Ελεύθερου κώδικα (EL/ΑΚ) πλατφόρμας προφίλ χρηστών. Η βιβλιογραφική ανασκόπηση αποκάλυψε ένα κενό στις υπάρχουσες προσεγγίσεις προφίλ χρηστών.

Συγκεκριμένα, οι υπάρχουσες μεθοδολογίες προφίλ χρηστών επικεντρώνονται μόνο σε μερικές πτυχές προσωπικότητας και επομένως, δεν εξάγουν μια πλήρη και ακριβή δημιουργία προφίλ χρήστη. Επιπλέον, δεν παρέχουν εξηγήσεις για τα μοντέλα τους, αντιθέτως τα μοντέλα μηχανικής μάθησης που χρησιμοποιούν συχνά θεωρούνται ως μαύρα κουτιά που είναι αδύνατο να ερμηνευτούν. Είναι κρίσιμο να καταλάβουμε ότι, τελικά, αυτά τα μοντέλα θα χρησιμοποιηθούν από ανθρώπους που πρέπει να τα εμπιστεύτονται, να κατανοήσουν τα λάθη που κάνουν και τη λογική πίσω από τις προβλέψεις τους.

Συνεπώς, το κίνητρο αυτής της εργασίας είναι αποτέλεσμα των προαναφερθέντων κενών και προκλήσεων προς ένα αποτελεσματικό προφίλ χρηστών. Πιο συγκεκριμένα, οι κυριότερες προκλήσεις που εντοπίστηκαν είναι οι εξής:

- Πολύτιμες πληροφορίες για άτομα, μπορούν να εξαχθούν μέσω της ανάλυσης κοινωνικών μέσων, εξυπηρετώντας διάφορους σκοπούς και εφαρμογές, όπως η διαφήμιση και το μάρκετινγκ. Ωστόσο, τα δεδομένα που χρησιμοποιούνται πρέπει να επεξεργάζονται κατάλληλα, για να είναι εφικτό να δημιουργήθονται ακριβή μοντέλα πρόβλεψης.
- Οι πληροφορίες που παρέχονται από τους χρήστες ενδέχεται να μην είναι επαρκείς ή ενδέχεται να είναι ελλιπείς ή ψευδείς, οδηγώντας σε ανακριβή αποτελέσματα, δηλαδή προφίλ χρηστών. Αυτό συμβαίνει γιατί μπορεί κάποιος να μην επιθυμεί να δημοσιοποιεί προσωπικές του πληροφορίες. Έτσι, δημιουργείται η ανάγκη για προηγμένες τεχνικές που θα μπορούσαν να εξαγάγουν αυτόματα τις πληροφορίες που χρειάζονται, χρησιμοποιώντας τις διαθέσιμες πληροφορίες.
- Η δημιουργία προφίλ χρηστών είναι μια κοινώς χρησιμοποιούμενη μέθοδος για την αποτελεσματική δημιουργία μιας ρητής ψηφιακής αναπαράστασης της ταυτότητας ενός ατόμου χρησιμοποιώντας μηχανική μάθηση. Ωστόσο, πρέπει να διεξαχθεί έρευνα σχετικά με κατάλληλα εργαλεία, προσαρμογές και βελτιστοποιήσεις που πρέπει να γίνουν για να είναι αποτελεσματικά.
- Τα υπάρχοντα εργαλεία επικεντρώνονται σε περιορισμένες πτυχές των προφίλ χρηστών. Για παράδειγμα, τα πιο κοινά είναι τα βασικά δημογραφικά χαρακτηριστικά όπως το φύλο, η ηλικία και η εθνικότητα. Ωστόσο, για την παροχή μιας πλήρης και πλούσιας εικόνας των χρηστών, θα πρέπει να περιλαμβάνονται περισσότερα ψυχοδημογραφικά στοιχεία που εξυπηρετούν κάθε είδους ανάγκες των χρηστών.
- Τέλος, οι ψυχοδημογραφικές πτυχές πρέπει να παρουσιάζονται στο τελικό χρήστη μαζί με τον/την λόγο/ους τους οποίους προέκυψε αυτό το αποτέλεσμα/πρόβλεψη. Δηλαδή, ο χρήστης πρέπει να γνωρίζει γιατί τους αποδίδεται μια συγκεκριμένη ‘ετικέτα’, προκειμένου να ληφθεί υπόψη τόσο η διαφάνεια των αποτελεσμάτων όσο και η ευρωστία. Για το λόγο αυτό, χρειαζόμαστε επεξηγήσεις για τα συγκεκριμένα αποτελέσματα που εξάγονται.

Προς αυτήν την κατεύθυνση, η συμβολή αυτής της εργασίας για μια αποτελεσματική λύση
για την δημιουργία προφίλ χρηστών αποτελείται από δύο βασικά μέρη:

- Αναπτύξαμε μια πλατφόρμα, με το όνομα Mediator, η οποία είναι διαδικτυακή και ανοιχτού κώδικα δημιουργώντας ένα πλήρες προφίλ χρήστη που περιλαμβάνει ένα ευρύ φάσμα πολυτιμών ψυχο-δημογραφικών πτυχών που προσδιορίζονται στη βιβλιογραφία. Ειδικότερα, οι ψυχο-δημογραφικές πτυχές που υποστηρίζονται από την πλατφόρμα μας κατηγοροποιούνται ως δημογραφικές και ψυχογραφικές. Πιο συγκεκριμένα, περιλαμβάνει πληροφορίες σχετικά με την ηλικία των χρηστών των κοινωνικών μέσων, το φύλο, το συναίσθημα, τα χαρακτηριστικά της προσωπικότητας που υπάρχουν στο μοντέλο των Big Five, την τοποθεσία, την εθνικότητα και τα ενδιαφέροντα. Τα δεδομένα που χρησιμοποιούνται για την εξαγωγή των πληροφοριών του ατόμου προέρχονται από τον λογαριασμό τους στο Twitter. Από όσο γνωρίζουμε, είναι το πρώτο έργο που εντοπίστηκε στη βιβλιογραφία και επιτρέπει τη δημιουργία προφίλ χρήστη και, συνεπώς, ένα αποτελεσματικό User Profiling.

- Αναπτύξαμε μια επεξεργασία προσέγγιση για το προφίλ χρηστών χρησιμοποιώντας αλγόριθμους μηχανής μάθησης. Καθώς σε αυτήν την εργασία δημιουργούμε το προφίλ χρηστών, ψάχνουμε να αναπτύξουμε τεχνητή νοημοσύνη και μηχανική μάθηση χρησιμοποιώντας προσωπικά δεδομένα, οπότε προκύπτουν τα ζητήματα λογοδοσίας και διαφάνειας τα οποία είναι ιδιαίτερα σημαντικά, και αν δεν είμαστε σε θέση να παρέχουμε αυστηρή επεξήγηση, στους αλγόριθμους μας, τότε περιορίζει τις πιθανές επιπτώσεις της τεχνητής νοημοσύνης για τέτοιες λύσεις. Στην τεχνητή νοημοσύνη και το υποπεδίο της μηχανικής μάθησης, υπάρχει το δικαίωμα για επεξήγηση που σημαίνει το δικαίωμα να δοθεί εξήγηση για ένα αποτέλεσμα του αλγορίθμου και αποτελεί θεμελιώδες ατομικό δικαίωμα που αντανακλά τα δικαίωμα για αποφάσεις που επηρεάζουν σημαντικά ένα άτομο, ιδίως νομικά ή οικονομικά).

Αυτή η μεταπτυχιακή διπλωματική εργασία ακολουθεί την ακόλουθη δομή: Το Κεφάλαιο 1 εισάγει το πρόβλημα, δίνει εισαγωγικές πληροφορίες σχετικά με το προφίλ χρηστών των κοινωνικών δικτύων, συζητά τις προκλήσεις και τα κίνητρα στο προφίλ των χρηστών των κοινωνικών δικτύων και, τέλος, τονίζει τη συμβολή της προτεινόμενης εργασίας. Το Κεφάλαιο 2 παρέχει μια συστηματική βιβλιογραφική ανασκόπηση σχετικά με τις μεθόδους και τις τεχνικές δημιουργίας προφίλ χρηστών στα κοινωνικά δίκτυα, περιγράφει τα προκαταρκτικά της υπάρχουσας έρευνας προκειμένου να καταστήσουμε την εργασία μας πιο κατανοητική για τον αναγνώστη, συζητά ζητήματα δεοντολογίας δεδομένων στο προφίλ χρηστών και καταλήγει στην επισήμανση του ερευνητικού κενού σχετικά με τις τρέχουσες μεθόδους δημιουργίας προφίλ χρηστών στα κοινωνικά δίκτυα. Το Κεφάλαιο 3 παρουσιάζει λεπτομερώς τη μεθοδολογία και τις φάσεις της προτεινόμενης προσέγγισης, καθώς και τη ροή εργασιών και την αναπτυγμένη αρχιτεκτονική του συστήματος. Επιπλέον, περιγράφει λεπτομερώς όλες τις φάσεις που ακολούθησαν τα δεδομένα μέχρι τα τελικά αποτελέσματα (π.χ. προεπεξεργασία δεδομένων, εξαγωγή χαρακτηριστικών), καθώς και χρησιμοποιεί αρχιτεκτονικές μηχανικής μάθησης (για κάθε πτυχή) που σχεδιάσαμε και αναπτύξαμε χρησιμοποιώντας τις ψυχο-δημογραφικές πτυχές για έναν χρήστη Twitter. Στο Κεφάλαιο 4, παρουσιάζουμε την πειραματική διαδικασία που ακολούθησαν προκειμένου να αξιολογηθεί η απόδοση της πλατφόρμας Mediator που αναπτύξαμε στα πλαίσια της
ις της εργασίας αυτής. Επιπλέον, παρέχει μια ολοκληρωμένη παρουσίαση των εξαγόμενων αποτελεσμάτων για κάθε ψυχοδημογραφική πτυχή, και τέλος παρέχει τα αποτελέσματα επεξηγήσεων (χρησιμοποιώντας την τεχνική LIME) για τις πρότυπες προβλέψεις που εξήχθησαν από έναν λογαριασμό Twitter μέσω της πλατφόρμας μας. Το Κεφάλαιο 5 συζητά τις τεχνολογίες που χρησιμοποιήσαμε για την ανάπτυξη της πλατφόρμας μας (δηλαδή, της πλατφόρμας Mediator). Συγκεκριμένα, συζητάμε το σενάριο περίπτωσης χρήσης, τις απαιτήσεις συστήματος και την αρχιτεκτονική της αναπτυγμένης πλατφόρμας και την υλοποίηση της πλατφόρμας λεπτομερώς, δηλαδή, αρχιτεκτονική backend, την τεχνική για την εγαγωγή των επεξηγήσεων και διεπαφή χρήστη - frontend. Τέλος, στο Κεφάλαιο 6, παρέχουμε τα κύρια συμπεράσματά μας, συνοδευόμενα από μερικές μελλοντικές προτάσεις για την περαιτέρω βελτίωση και επέκταση της πλατφόρμας.
Acknowledgements

This chapter in my life would not have been possible without all the emotional and scientific support of many people who have accompanied my route through these last years.

First of all, I would like to demonstrate my sincere gratitude to my advisor, Professor Athena Vakali for giving me the opportunity to pursue my research work and work on my thesis under her supervision. I am very grateful for her encouragement, guidance and for the opportunity of working with this research team. I am also very thankful to my co-adviser, Dimitra Karanatsiou, for all the knowledge and enthusiasm which allowed me to achieve my goals and set milestones throughout my project research and development.

Special thanks to my friends Paschalis, Kostas and Tekos (Dimitris) for the great teamwork that we had two years now and all these memories we created during all these nights of endless study and work, for all the work and projects that we successfully achieved.

At last, I would like to thank all my family as you were the people who played the most important role all these years, during both undergraduate and postgraduate studies. To my mum and dad who continuously supported, believed in me and pulled me up when I needed most, I will always be grateful. Without your psychological support, I would not be able to follow my dreams and expectations.
Contents

Declaration of Authorship ......................................................................................... i
Abstract .................................................................................................................. ii
Εκτεταμένη Περίληψη ............................................................................................... iv
Acknowledgements ................................................................................................... viii

List of figures ........................................................................................................... xi
List of Tables ............................................................................................................ xii

1. Introduction ........................................................................................................... 1
  1.1 Social Media User Profiling: background and fundamentals ....................... 2
  1.2 Challenges and Motivation in Social Media User Profiling ......................... 4
  1.3 Contribution of the thesis ................................................................................ 7
  1.4 Structure of the Thesis ..................................................................................... 9

2. User Profiling Methods in Social Networks: A Systematic Literature Review ........................................................................ 10
  2.1 Reviewing methodology of User Profiling Methods in Social Networks .......... 10
  2.2 Overview of User Profiling Techniques in Social Networks ......................... 14
    2.2.1 Data gathering - Data Sources ................................................................. 15
    2.2.2 Aspects of user profiling in social networks ........................................... 19
    2.2.3 User profiling proposed algorithms ......................................................... 23
  2.3 User profiling in social networks: an overview of the most popular and open available tools .................................................................................. 26
    2.3.2 A comparison among user profiling platforms and tools ......................... 32
  2.4 Data Ethics Issues in User Profiling ............................................................... 34
  2.5 Research Gap on Current User Profiling Methods in Social Networks .......... 37

3. Proposed Methodology & Approach for User Profiling in Social Networks ........................................................................... 38
  3.1 Workflow – System Architecture .................................................................... 38
  3.2 Pre-model Phases .............................................................................................. 40
    3.2.1 Data preprocessing ..................................................................................... 40
    3.2.2 Feature Extraction ..................................................................................... 41
  3.3 Building Classification/Regression models ....................................................... 43
    3.3.1 Gender Classifier ...................................................................................... 44
    3.3.2 Age Classifier ............................................................................................ 45
List of Figures

Figure 1.1: Social media types (“Social Media Marketing”, 2021) ................................. 3
Figure 1.2: Psycho-demographic aspects (“CB Insights”, 2021) ...................................... 5
Figure 2.1: Systematic Literature Review Stages .................................................................. 11
Figure 2.3: A New Classification Schema of user profiling proposed by (Farid et. al., 2018). .......................................................... 18
Figure 2.4: Instance from language analysis in personality space ...................................... 27
Figure 2.5: Features of URL Profiler. .................................................................................. 31
Figure 3.1: User interaction with the platform – Basic overview ........................................ 39
Figure 3.2: Machine learning process for User Profile building. ....................................... 39
Figure 3.3: Gender model instance prediction ..................................................................... 45
Figure 3.4: An example of Voting Classifiers ..................................................................... 46
Figure 3.5: Sentiment classifier training steps ..................................................................... 47
Figure 3.6: Emotion prediction of an account ..................................................................... 48
Figure 3.7: Sentiment VS Emotion User Profiling Analysis. .............................................. 49
Figure 3.8: The Big Five personality traits, acronymed Big Five. ....................................... 50
Figure 3.9: Lime prediction example (Bars show the importance of the features in each class. Red predict class 0 and green class 1) .............................................................. 53
Figure 4.1: Gender Classifiers results ............................................................................... 59
Figure 4.2: Age predictor results ....................................................................................... 61
Figure 4.3: Sentiment analysis results ................................................................................ 63
Figure 4.4: Emotion’s predictors results for RF classifier .................................................. 65
Figure 4.5: Relationship between Explainability and Accuracy .......................................... 68
Figure 4.6: White box vs black box performance demographics ........................................ 72
Figure 4.7: White box vs black box performance - emotion .............................................. 72
Figure 4.7: Shap explainability results for Openness model (top-10 features) ...................... 76
Figure 5.1: System Architecture ....................................................................................... 80
Figure 5.2: Technologies of the platform .......................................................................... 83
Figure 5.3: Mediator - Home screen .................................................................................. 85
Figure 5.4: Mediator - Age classification .......................................................................... 86
Figure 5.5: Mediator - Gender classification ...................................................................... 86
Figure 5.6: Mediator - Sentiment prediction ...................................................................... 87
Figure 5.7: Mediator - Emotion prediction ......................................................................... 88
Figure 5.8: Mediator - Emotion prediction word-clouds .................................................... 88
Figure 5.9: Mediator - Personality prediction ..................................................................... 89
Figure 5.10: Mediator - Topics 1 prediction ....................................................................... 90
Figure 5.11: Mediator - Topics 2 LDA .............................................................................. 90
Figure 5.12: Mediator - Ethnicity/Location prediction ....................................................... 91
List of Tables

Table 1-1: Psycho-demographics aspects included in Mediator ................................................................. 7
Table 2-1: Search terms and examples given in DBs ................................................................................. 13
Table 2-2: Example of paper details screened in literature review ......................................................... 14
Table 2-3: Aspect categorization of literature. ......................................................................................... 22
Table 2-4: Capabilities of the existing platforms ..................................................................................... 33
Table 3-1: Features extracted by the tweet content .................................................................................. 43
Table 4-1: Dataset’s overview .................................................................................................................. 55
Table 4-2: Gender models hyper-parameterization ................................................................................ 58
Table 4-3: Performance results for different classification algorithms. ................................................... 59
Table 4-4: Performance results for text model and context model ......................................................... 60
Table 4-5: Class imbalance techniques results ........................................................................................ 61
Table 4-6: Random Forest in preprocessed and not preprocessed data .................................................. 62
Table 4-7: Hyperparameters of sentiment models .................................................................................... 62
Table 4-8: Performance metrics with data preprocessing, with data preprocessing and using pretrained models ................................................................................................................................. 63
Table 4-9: Performance metrics for SVM, Naive Bayes, Logistic Regression, Random Forest and Voting Classifier ML algorithms. ................................................................................. 64
Table 4-10: Performance results for Random Forest classifier ............................................................... 65
Table 4-11: Performance results for SVM classifier. ............................................................................... 65
Table 4-12: Performance results for Naive Bayes classifier. ................................................................. 66
Table 4-13: Performance results for Logistic Regression classifier. ....................................................... 66
Table 4-14 Personality (OCEAN) models hyper-parameterization ......................................................... 66
Table 4-15: Decision Tree and Random Forest for age prediction .......................................................... 69
Table 4-16: Decision Tree, Random Forest and VC in black box .............................................................. 69
Table 4-17: Random Forest performance metrics for emotion .............................................................. 70
Table 4-18: Logistic regression performance metrics for emotion .......................................................... 70
Table 4-19: Decision tree performance metrics for emotion ................................................................. 70
Table 4-20: Logistic regression, Random Forest and a voting schema for sentiment prediction ............ 71
Table 4-21: OCEAN model comparison (RMSE) of explainable and non-explainable models ............... 71
Table 4-22: Emotion – Sadness Classifier results ..................................................................................... 74
Table 4-23: Emotion – Top-5 words per prediction ................................................................................... 75
Table 4-24: Gender Explainability ~ Male prediction ............................................................................. 75
Chapter 1

1. Introduction

These days, as the evolution of social media has reached unprecedented levels, new opportunities for extraction and collection of essential knowledge about certain social media users have been raised. There is no doubt that this new era of social media has a high daily impact in the lives of modern individuals providing a wealth of data that can reveal interesting information about persons, aiding the extraction of personality aspects and preferences after proper analysis. For example, the number of the followers of a social media user could possibly indicate valuable information about their personality, such as whether a person is extroverted or introverted, while their reactions on certain publications can reveal whether the user is conservative or open to new ideas. Thus, social media users are ripe sources of information regarding personalization.

In fact, the recent growth of the available information on the internet, as well as the huge diversity of social media users have led to a high priority toward personalization which can facilitate several purposes such as social media advertising\(^1\) and targeted marketing\(^2\). These practices can direct companies to identify their target group(s), and focus their advertising campaigns exclusively at a niche audience for the promotion of their products (Celli et. al., 2017). On top of that, these data can be leveraged by companies to understand whether the product or service they have developed and promoted satisfies users or not, by analyzing their user profiles. In this context, user profiling remains a crucial issue for both information and service personalization. A key research venture in these areas would be the automatic extraction of personality information and their visualization in effective ways in order to better understand their needs and provide personalized services, for example better serve marketers (FlexMR, 2021). Hence, building an automated user profile based on accumulated information is the main challenge in developing an adaptive personalized application for a wide spectrum of purposes. Research in personalization is used in several applications such as recommendation, e-learning, and personalized information retrieval, etc.

Based on the above statements, this work is motivated by the necessity to analyze a bigger image of social media users including their psycho-demographics information along with their preferences in order to provide accurate analysis of users profiling without requiring direct input of data. This process is called \textit{user profiling} and from the provided results multiple stakeholders on different domains can be benefit. For instance, some indicative use

\(^1\)https://www.bigcommerce.com/blog/social-media-advertising/#what-are-the-benefits-of-advertising-on-social-media-channels

\(^2\)https://www.thebalancesmb.com/target-marketing-2948355
cases are: (i) advertising (Celli et al., 2017), since users’ demographics affect their purchasing choices, (ii) employee hiring with the purpose of matching specific candidates’ traits with position’s requirements (Eke et al., 2019), (iii) marketing, as consumer/customer profiling is the only way to gather the insights needed to identify, segmentate and define the target audience for a specific product etc (Choi & Lim, 2020). In this chapter we present fundamental information on the topic of User Profiling, as well as the research gap in recent literature that this master thesis is aiming to address.

1.1 Social Media User Profiling: background and fundamentals

In recent years, social media is used by an ever-growing part of the population\(^3\) in a daily basis. Their capabilities are constantly expanded, allowing them to gain very significant assets such as internet connectivity, education, sharing knowledge and advertising, according to Akram & Kumar (2017). Specifically, these assets affect businesses, education and society as social media are proven to wield the power to influence all human activities. More specifically, in the area of culture and education, they can offer alternative ways of groups’ constitutions and interactions, improving the quality of learning and conversation. Furthermore, in a sociological spectrum, Social Networks (SNs) represent a “virtual society” since users connect with others in order to communicate, exchange opinions and form groups or communities. On top of that, a key characteristic of social media is that information and news can be disseminated much faster than traditional ways such as newspapers or websites. Thus, the amount of data being exchanged daily in social media is considerably vast. In fact, as most users usually have accounts on more than one social networks (Kaushal et. al., 2019) they generate data via their active participation (User-Generated Content). As a result, social media have undoubtedly become an integral part of our daily lives. Of course, as users present different personality traits, opinions and interests, social media have to adapt and cater to every need, resulting in different types of media being developed.

There are many types of social media and each one is used for different purposes and contains different content. as shown in Figure 1.1. The most popular types are blogs, microblogs, social news, media sharing, sites that have to do with opinion sharing, reviews and ratings as well as online social networking. Their key features are Participation, Openness, Conversation, Community and Connectedness\(^4\). Currently, the most popular social media platforms are Facebook, YouTube, LinkedIn and Twitter, among others\(^5\). Twitter and Pinterest are social media sites and belong to the category of microblogs. YouTube is a web application where users can upload videos they have created and watch other’s content. It focuses on videos and each person can follow channels that are created by the users (“YouTubers”). Users can rate and comment the content of the platform, creating a direct feedback model and exchanging knowledge. On the other hand, LinkedIn focuses on individuals’ professional careers and contains more formal and professional content.

\(^{4}\) 7 Features Of A Successful Social Media App - Young Upstarts
\(^{5}\) https://www.adobe.com/express/learn/blog/top-social-media-sites
Specialized in scientific topics, job opportunities and academic education. All of them can be used in advertising and communication as well. In addition, there are many more such as Instagram, Flickr, Quora etc. At this point we should mention that all these types of social media are primarily used to fill or create gaps and needs which the modern, demanding and rapidly changing societies dictate, while all together have formed a new type of "digital society".

In this digital society, users are interested in different types of content and topics while also wanting to belong in groups with likeminded individuals. This way, they feel part of a bigger group that accomplishes common goals and shares common beliefs, ideals and personality traits. Platforms, like Facebook and Twitter provide some techniques to select the teams or groups that are more relevant to an individual’s interests. This facilitates the analysis of individuals and allows them to be grouped according to some characteristics like age, gender, buying behavior etc.

Moreover, the information found on a Social Media profile can be used to extract insights about the psycho-demographic traits of a user. This information consists of input explicit data provided by the user (e.g., account data such as name, surname, date of birth, religion etc.) or implicit data that the system can extract from a user’s activity (e.g., likes, posts or comments to posts). All these data can be gathered and analyzed creating a user profile that contains the primary digital characteristics of an individual. User profiling has attracted wide research attention as it can yield valuable information about a social media user which can be utilized for several purposes such as social media advertising\(^6\) and targeted marketing\(^7\), as discussed previously. On top of that, these data can be used by companies to understand whether the product or service they have developed and promoted satisfies users or not by analyzing their comments and reactions.

\[\text{Figure 1.1: Social media types (“Social Media Marketing”, 2021)}\]

\(^6\)https://www.bigcommerce.com/blog/social-media-advertising/#what-are-the-benefits-of-advertising-on-social-media-channels
\(^7\)https://www.thebalancesmb.com/target-marketing-2948355
In the following sections, this chapter gives a brief overview of the basics of this work so that readers can be familiarized with User Profiling in Social Media. Basic definitions about Social Media and their applications are given descriptively. In the next section the problems being addressed are widely described and afterwards thorough discussion about our contribution and novelty in the existing research and scientific field is provided in detail. Finally, the structure of the remaining chapters of this master thesis is presented.

1.2 Challenges and Motivation in Social Media User Profiling

Nowadays many people participate in social media, providing a lot of data which can be processed to extract valuable information. As mentioned before, users can explicitly include information about themselves such as age and gender. This information can be analyzed and used by advertising campaigns to identify their target group, and afterwards concentrate only at this specific audience. For instance, a teen-friendly product should not be recommended to elder customers. However, sometimes people do not include publicly available information about their personality and psycho-demographic aspects in their posts or in their profiles since they may want to preserve their privacy. On the other hand, in some cases, the explicit information provided by the individuals may be fake or incomplete. Furthermore, some Social Networks (SNs) do not directly support explicit personal data. For example, Twitter accounts do not contain information about the user’s gender. This causes many challenges in user profiling analysis that may lead to inaccurate results. Hence, there is a need to extract implicit information about the users based on their activity in social media networks which is influenced by information not provided directly by them.

In particular, these implicit data can be extracted using the available information in the user’s account, also referred to as User Profiling. User Profiling is the scientific field which extracts useful information explicit, implicit or a combination of both (hybrid) (Kanoje et al, 2014). User’s information, specifically personality and psycho-demographics aspects, such as preferences, behavior, opinion and personality, are possible to be extracted from their social media (van Engelen & Hoos, 2019) and online activity (Belini T., 2021). In particular, these features are part of two main categories, demographics and psychographics. The first category includes aspects such as age, gender, location, ethnicity, religion etc. Psychographics are the emotion, sentiment, personality traits, interest etc. The psycho-demographic aspects can be used by companies, organizations and individuals in their advertising efforts. More specifically, User Profiling can potentially be used in scenarios such as targeted advertising. For example, a service provided by a company which targets people who are over 50 years old or additionally, the males in this age group, will certainly require some User Profiling analysis to pinpoint the characteristics of this target group and devise a well-rounded advertising strategy. Sentiment and emotion analysis is also important for a company, to recognize the reactions of the customers in changes and plan ahead for future improvements.

Regarding the development of such an analysis, user profiling can be achieved using ontology-based approaches, machine learning approaches and statistical-based approaches (Sugiyama et. al., 2004; Mendis, 2021). Machine Learning (ML) is the scientific field which automatically acquires new knowledge and is able to make decision or predict a value. With
these techniques it is possible to train machine learning models and extract/predict results for uncertain values which are not explicitly available. There are three main machine learning approaches, namely supervised, semi-supervised and unsupervised (Kanoje et al., 2014; van Engelen & Hoos, 2019). In general, quality data are required to train the ML algorithms in order to create efficient and effective models for accurate predictions. These data might arise from a variety of sources and tools. Specifically, the insights for the user can be exported through social media (Kazai, Yusof & Clarke, 2016; Volkova et al., 2016; Vasanthakumar et al., 2019), games (van Lankveld et al., 2011), website surf behavior (Catalin et al., 2019; Celli et al., 2017), eye movements (Ma et al., 2017) etc. Social media content consists of text data, derived from user’s tweets and profile info, images, video and audio that the user may have posted in the social media platform. In this work, we deal with the information that can be extracted from the text: direct information from a user’s profile and posts (i.e., name, short description), as well as indirect information that is a product of the analysis of a user’s profile and posts (i.e., sentiment and emotion analysis). From all these data, features can be extracted and passed through to the models for valuable results and insights about any social media user.

![Image](image.png)

*Figure 1.2: Psycho-demographic aspects (“CB Insights”, 2021)*

However, the data may include or be prone to bias and therefore the predictions made by machine learning models may not be based on appropriate information (Gilpin et al., 2018). For example, gender extraction from a user’s account, could rely on existing information about their education which may not be the appropriate criteria for the decision. To deal with this concern, and to alleviate any confusion caused to the user regarding the predictions of a model, decisions made by the ML models must be explainable with proper and reasonable justifications so as to be understandable by the end user and thus gain their trust (Ribeiro et al., 2016; Choi & Lim, 2020). A new concept introduced and widely used in Artificial Intelligence to deal with such problem is the Interpretability / Explainability. These are techniques that can be used to explain the predictions made by agents. Ribeiro et al. (2016) refers to explainability as “presenting textual or visual artifacts that provide
qualitative understanding of the relationship between the instance’s components (e.g., words in text, patches in an image) and the model’s prediction”. Specifically, a model’s result/decision, is influenced by certain features that may or may not be more important than others. These features, however, do not affect every result with the same degree of importance. Feature importance varies and may influence a decision at high level or low level. Using Explainability techniques, we can provide the reasons that a decision was taken by a predictive model. These techniques could prevent erroneous or discriminatory predictions. Furthermore, by applying these techniques and providing the reasons for a model’s predictions, they would be more trustworthy and acceptable by the public.

Last but not least, it is worth to mention that a large spectrum of analysis tools has been developed for social media user profiling over the last years, exporting psychodemographics from explicit or implicit data. In particular, existing tools are mainly categorized into two broad groups, which are social listeners and social profiling tools. Social Listeners are platforms which analyze explicit information and provide statistics as results for psycho-demographic aspects. These tools can monitor conversations of the users about products or events. In addition, these tools can be used to track products and events and are limited to analyzing implicit information. Social Profiling tools collect information and create an overall user profile. Specifically, they focus on each individual separately and extract implicit data. They also require consent from the user because they collect sensitive and personal data in contrast to social listeners. On top of that, each organization or service or even individual wants to learn different aspects of an account, so there are different needs for analyses for each of them that requires the corresponding tools.

The motivation of this work is a result of the aforementioned gaps and challenges towards an efficient and effective user profiling. More specifically, the main challenges identified are as follows:

- Valuable information about individuals can be extracted through social media analysis, serving several purposes and applications such as advertising and marketing. However, research must be carried out about which data we should take into account, as well as build accurate predictive models using data of any form.
- Information provided by the users may not be sufficient or it might be incomplete or false, leading to inaccurate user profiling. So, advanced techniques that could automatically extract information needed, utilizing implicit information.
- User profiling is a commonly used method for efficiently creating an explicit digital representation of a person’s identity using machine learning. However, research must be carried out about proper tools, adjustments and optimizations that must be done to be efficient and effective.
- The existing tools focus on limited aspects of user profiles. For example, most common are the basic demographics characteristics such as gender, age and ethnicity. However, to provide a complete and rich image of users more psycho-demographics should be included serving all kind of user’s needs.
- Psycho-demographics aspects should be presented not as simple categories, numbers or information, but the end user need to know why a specific label is assigned to them, in order to account for both transparency of results and robustness. For this reason, we need our results to be explainable.

The challenges mentioned above, are the basic motivation of this study and consequently the development of the platform. The effort of this work focuses to handle these existing gaps,
covering a wide range of psycho-demographic aspects and providing explanations on them using accurate machine learning models. In the following section, we discuss the contributions of the current study in detail.

1.3 Contribution of the thesis

As we discussed previously, the initial phase of user profiling requires sufficient information gathering about the user. However, when the information is not sufficient (e.g., it is false or incomplete), it will lead to a cold-start problem. The cold start problem is common in learning and adapting dynamic user profiles for personalization, where the system is not capable of providing an effective personalization service in order to learn the user profile. This is very common in content-based personalization systems. Novel and more sophisticated user profiling approaches need to be developed to address such a problem.

To combat this problem, as well as in order to develop a robust and efficient user profiling solution addressing to the best the aforementioned challenges identified in the literature, presented in the previous subsection, we used machine learning techniques to extract psycho-demographics traits. On top of that, we used explainable machine learning techniques to extract psycho-demographics traits of Online Social Network (OSN) users by developing an online, open-source, explainable, user profiling platform. The primary aim of this work is to cover all of the psycho-demographic aspects found in the literature. Specifically, we focus on the features presented in Table 1-1:

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Psychographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Sentiment</td>
</tr>
<tr>
<td>Gender</td>
<td>Emotion</td>
</tr>
<tr>
<td>location</td>
<td>Personality</td>
</tr>
<tr>
<td>Nationality/Ethnicity</td>
<td>Interests</td>
</tr>
</tbody>
</table>

Our developed platform, Mediator, is based on data provided by Twitter. Twitter is one of the most popular social media platforms worldwide with the number of daily active users exceeding 350 million. A lot of the existing research in the literature focuses on Twitter. Especially, they concentrate on extracting implicit data and creating datasets to be used in new research ventures. Those datasets can be used for supervised techniques. In addition, it provides an open API (Application Programming Interface). This way, there is an interface for communication between the same application (Twitter) and external parties.

The purpose of this work is to develop an online, open-source, explainable, user profiling platform. The literature review revealed a gap in the existing User Profiling approaches. In particular, the existing user profiling methodologies focus only on a few psycho-demographic aspects, and thus, they do not export a complete and accurate user profile creation. On top of that, they do not provide explanations for their models, rather than their ML models are often thought of as black boxes that are impossible to interpret. It is crucial to understand that in the end, these models will be used by humans who need to trust them, understand the errors they make, and the reasoning behind their predictions.

---

8 https://en.wikipedia.org/wiki/API
Towards this direction, the contribution of this work in the direction of an efficient and robust user profiling solution is twofold:

- We developed a platform, named Mediator, which is online and open source creating a complete user profile including a wide spectrum of valuable psycho-demographics aspects identified in the literature. In particular, the psycho-demographics aspects which are supported by our developed platform are categorized as demographics and psychographics. More specifically, it includes insights about social media user’s age, gender, sentiment, emotion, personality traits that exist in Big Five model, location, ethnicity, and interests. The data used to export the individual’s insights is obtained from their Twitter account. To the best of our knowledge, it is the first work identified in the literature that enables user profiling based on numerous valuable insights about the user. Previous works used only a limited number of insights about the social media user. Therefore, this work contributes towards a novel solution for a more complete user profile creation, and thus an efficient and effective user profiling.

- We developed an explainable and interpretable user profiling approach using machine learning algorithms. As in this work and user profiling works generally, we are looking to deploy artificial intelligence and machine learning using personal data, questions of accountability and transparency are particularly important, and if we are unable to properly deliver explainability, in our algorithms, we will seriously be limiting the potential impact of artificial intelligence for such solutions. According to the European Commission “Explainable AI: the basics” report⁹, there are a few reasons why some form of interpretability explainability in user profiling solutions might be desirable. These include: (a) giving users confidence and trust in the system, (b) prevent bias, (c) meeting regulatory and policy requirements, (d) improving system design, (e) assessing risk, robustness, and vulnerability, and (f) understanding and verifying the outputs from a system. On top of that, the European Union General Data Protection Regulation¹⁰, known as GDPR, extends the automated decision-making rights in the 1995 Data Protection Directive to provide a legally disputed form of a right to an explanation, stated as such in Recital 71¹¹: "The data subject should have the right ... to obtain an explanation of the decision reached". In particular, in artificial intelligence and its subfield of machine learning, a right to explanation¹² (or right to an explanation) which means a right to be given an explanation for an output of the algorithm constitutes a fundamental individual right (i.e., right to be given an explanation for decisions that significantly affect an individual, particularly legally or financially) and should be seriously taken into consideration.

For the above reasons, we used machine learning methodologies to extract the insights. On top of that, along with the user’s psycho-demographics aspects, we have included proper and reasonable explanations for each model’s decision to provide transparency and clarity, as well as avoid any kind of bias or discrimination. It is worthwhile to highlight that clarity ensures transparency, proper functioning, and non-discrimination.

¹⁰https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
¹²https://en.wikipedia.org/wiki/Right_to_explanation
Finally, we deal with the problem that arises when researchers are trying to implement efficient machine learning algorithms ensuring the property of interpretability/explainability in their developed solutions. In some cases, it is not possible to use those techniques properly with advanced model structures. Thus, a trade-off between explainability and the accuracy of the models is presented.

1.4 Structure of the Thesis

The current study consists of six chapters, each of them organized in sub-sections. The remaining chapters are structured as follows. Chapter 2 provides a literature preview and preliminaries of the existing research in order to make our work more comprehensible to the reader, while Chapter 3 presents, in detail, the methodology and the phases of the proposed approach and the developed architectures of the models. In chapter 4, we discuss the technologies that we used to develop our Mediator platform and a short presentation of the results is given. In Chapter 5, we present the experimental procedure followed in order to evaluate the performance of the Mediator platform and, finally, in chapter 6, we provide our main conclusions, accompanied by some future work suggestions.
Chapter 2

2. User Profiling Methods in Social Networks: A Systematic Literature Review

The recent growth of the available information on the internet, as well as the huge diversity of social media users led to a high priority towards personalization. In this context, user profiling is the crucial issue for both information and service personalization, such as personalized advertisement (Farid et. al., 2018; Eke et. al., 2019). Building an automated user profile is the main challenge in developing an adaptive personalized application for a wide spectrum of purposes (Kazai et. al., 2016; Celli et. al., 2017). Towards this direction user profiling has gained the attention of researchers in this area. It is worthwhile to highlight that effort has been also put towards the effective, efficient and accurate user profiling.

In this chapter we present our study on the available relative bibliography on user profiling following a systematic literature review approach, and afterwards we provide a discussion on the results. In addition, psycho-demographic aspects that already have been researched are described. In particular, Section 2.1 presents the research methodology that was used in order to include and study the background literature which is based on the Kitchenham process (Kitchenham, 2004). The next section 2.2 provides a literature review of user profiling and describes the methodologies that have been developed in recent years. Section 2.3 describes a variety of tools on the market for performing user profiling. An extensive description of these tools and the possibilities they provide is given. Section 2.4 provides a description of data ethics issues in user profiling. Finally, the results from the systematic literature review are summarized to provide a foundation for organizing research efforts towards the design and development of proper user profiling solutions for a large number of applications.

2.1 Reviewing methodology of User Profiling Methods in Social Networks

In this section we describe the literature review process on user profiling, as well as give some basic definitions for user profiling. Specifically, the first subsection describes the
Kitchenham methodology on which we based to perform the literature review on user profiling for the development of our platform. The second subsection analyzes the methodologies proposed by researchers in the literature for efficient user profiling. Then, the existing gap of the current user profiling approaches and the necessity for novel and more sophisticated solutions are discussed.

2.1.1 Search methodology

In this section we describe the research methodology which applied in the current work. The different stages of the process described in the following subsections. The methodology which was selected is Kitchenham method (Kitchenham, 2004), a popular approach that organizes the research providing more accurate research directions avoiding bias and misleading. Kitchenham consists of three main phases: a) Plan Review, b) Conduct Review and c) Document Review, which are described below. In Figure 2.1, the workflow of the whole process is represented.

**Plan Review**
The goal of a systematic literature review is to provide a qualitative research work. Describing each of the steps, followed by the researchers in their work, this method ensures transparency in terms of the output results. The organization which is offered by these systematic methodologies serves to avoid mistakes that can potentially happen, as well as misunderstandings. In addition, it provides an overview of the research that has already been done by the researchers.

![Systematic Literature Review Stages](image)

Our research concentrates on multi-aspect user profiling from social media with emphasis in the psycho-demographics of each social media user. For this reason, we have set four research questions to address. These questions constitute the basis for this work and the current research on user profiling.
Q1: What are the research topics and tools/platforms in user profiling which have already been addressed?

The first question we came up with is about the already developed tools that exist. Also, what capabilities those tools provide as well as how easily a user could handle them. The research of papers is critical to understand the existing methodologies, detect existing literature gaps and future directions to improve the current approaches. In addition, a better understanding would lead to improvements on current tools/platforms giving better results and further research in the scientific field.

Q2: What tools have been developed for user profiling (for Social Media and not)?

Predicting user aspects in social media is a handy way to get insights, providing valuable information about future clients and partners, and many other applications. On top of that, those tools could lead to better communication between people whose characteristics differ. For this, it is important for a tool to provide a useful way of interaction with the end user.

Q3: What are the research gaps in user profiling?

The systematic literature review was used to identify gaps of the existing proposed methodologies of the scientific field. The organization provided by the methodology helped us to extract results objectively. After identifying the gaps, we could improve/expand current tools, create novel approaches, or lead the scientific field to a new era.

Q4: What are the future directions of the user profiling?

Previous questions reveal future directions and gaps of the current state. This question is an important factor about the proposed approach of this thesis as well as future improvements.

At the second step of plan review, we developed a protocol, to exclude papers which did not imply our requirements for this research. This can be helpful in finding relevant and proper literature review, as well as making the research more objective and avoid any kind of bias. So, in our work we excluded papers which didn’t meet the following requirements:

a. papers with less than four pages  
b. papers which were not in the English language  
c. papers which were about other scientific field/topic  
d. papers that weren’t in full text and  
e. papers similar to others that we included and had similar topics.

For example, regarding papers which were about the age profiling aspect, we screened and studied only the ones that have been published recently. To screen a paper, it has to pass the three phases described in the next section. In our protocol, we didn’t set rules about most cited papers, authors or universities because we considered this as a bias. As referred by Bray in (2020), gender bias could affect the bibliometrics of an individual. Also, Reingewertz & Lutmar (2018) approve that some journals publish papers that contain their affiliation to the paper. For example, someone educated in a well-known university is more probable to be accepted by the journal. Also, we passed from phases one and two with further attention, papers which were published before 2013 but we didn’t set a strict rule about the publication year, as we also considered papers published earlier.

Conduct Review - Paper collection

As mentioned previously the second phase of the process is to collect relevant papers and extract the required data on the specified topic. We performed an extensive search for
relevant research papers in many bibliographic databases and used the protocol, which we
developed, to screen or reject the papers that presuppose or not our requirements
respectively. The research protocol that we created, used to collect the most relevant
research papers for our research. In order to screen the relevant papers, we defined some
“strong” keywords about the topic we are interested in. We combine the keywords identified
the most suitable for our research in a boolean logic to get the most qualitative results. The
main keywords which we used are shown in Table 2.1 alongside with some search
examples. We have to note that the search was not limited to these keywords, but they were
the most important and returned the most relevant results. Other secondary keywords
applied, after the results from the main ones were collected. In addition, operators provided
by search engines such as “AND” were used to locate more than one word. Also, phrase
search was achieved using the quotation marks which searches for expression in the text, or
else the search return results that contain most relevant to the query.

<table>
<thead>
<tr>
<th>MAIN KEYWORDS</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER PROFILE</td>
<td>“user profiling”</td>
</tr>
<tr>
<td>PERSONALITY</td>
<td>“user profiling” AND “personality”</td>
</tr>
<tr>
<td>PSYCHOMETRICS</td>
<td>user profiling psychometrics</td>
</tr>
<tr>
<td>DEMOGRAPHICS</td>
<td>“age classification”</td>
</tr>
</tbody>
</table>

Table 2-1: Search terms and examples given in DBs.

We observed that many of the results were about social sciences and specifically when we
used “personality” as the search keyword or within a search phrase. Nevertheless, our
research is about user profiling in social media using Machine Learning techniques, so
our searches focused on the combination of these terms.

After we specified the search keywords, we had to choose the databases which we will
conduct our searches. We chose some of the most important publishers since they have
important research articles and all of them goes through a peer-review which ensures the
quality of published articles. Five databases have been used for our main searches: Science
Direct (powered by Elsevier), Scopus, IEEE Xplore, ACM Digital Library and Springer. We
focused on these tools and we avoided web searches, like google search, which could lead to
articles of unknown quality, although in some cases that was not avoidable.

The results given by the databases included many non-relevant papers or papers that
emerged from other scientific fields. First of all, we screen the titles of the papers in the
results. Titles which were not relevant were rejected. Again, the papers which had relevant
titles and passed the previous step are checked again in their abstract. We read the abstracts
so that we can get a deeper understanding about the topic and contribution of each one.
Also, we check for keywords (not only the search keywords) that could highlight other
aspects of the topic. Finally, after we collected all the papers that were relevant, we read
again the abstracts, conclusions and/or discussions in order to include them in our research.

Finally, posterior to the paper screening, we extracted required data from the papers for our
work. These data collected in an excel table to keep track of the literature that we screened
previously. The table was used as a form to organize the information gathered in order to
detect gaps in research and future directions. The table filled with the data in each column.
The data which was collected attempt to answer the question which we defined at the first
steps of this study. An example of a paper collected is shown in Table 2.2 including the most important information that we kept track.

Table 2-2: Example of paper details screened in literature review.

<table>
<thead>
<tr>
<th>#</th>
<th>INFORMATION</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Title</td>
<td>Birds of a feather locate together? Foursquare checkins and personality homophily</td>
</tr>
<tr>
<td>2</td>
<td>Year</td>
<td>2016</td>
</tr>
<tr>
<td>3</td>
<td>Author</td>
<td>Nyala Noe, Roger M. Whitaker, Martin J. Chorley, Thomas V. Pollet</td>
</tr>
<tr>
<td>4</td>
<td>Venue</td>
<td>Computers in Human Behavior</td>
</tr>
<tr>
<td>5</td>
<td>Keywords</td>
<td>Spatial homophily, Personality, Location-based social network Foursquare</td>
</tr>
<tr>
<td>6</td>
<td>Focus</td>
<td>Personality homophily and preferences (location)</td>
</tr>
<tr>
<td>7</td>
<td>Objective</td>
<td>whether people with similar personality traits have a preference for common locations</td>
</tr>
<tr>
<td>8</td>
<td>Method</td>
<td>network and content analyses (clustering method and emotion valence estimation)</td>
</tr>
<tr>
<td>9</td>
<td>Features</td>
<td>#friends, #followers</td>
</tr>
<tr>
<td>10</td>
<td>Sample</td>
<td>174 anonymous Foursquare users</td>
</tr>
<tr>
<td>11</td>
<td>Dataset</td>
<td><a href="http://www.libs.uga.edu/multisearch.html">http://www.libs.uga.edu/multisearch.html</a></td>
</tr>
<tr>
<td>12</td>
<td>Performance</td>
<td>p-value=.0001, p-value=.016, p-value=.71</td>
</tr>
</tbody>
</table>

Document Review
The last step of Kitchenahm methodology is Document Review. This consists of the report creation and the validation. In the Section 2.2, Section 2.3 and Section 2.4, we present user profiling methodologies, technology review and data ethics along with the existing techniques respectively. These sections provide useful information about the existing User Profiling approaches.

2.2 Overview of User Profiling Techniques in Social Networks

Nowadays, the rapid development of information caused an overload on the web, which increases the demand for personalized approaches for finding the required information effectively and efficiently minimizing the time needed for the end users. Research in personalization is used in several applications/use cases such as recommendation, e-learning, and personalized information retrieval, etc.

Profiling of a social media user is the process of identifying the data about a user’s interest domain that constitute the user model (Farid et.al., 2018). A user profile is generally used for personalization and user modeling. It reflects personal information or interesting facts about an individual user, demographic information, e.g., name, age, education level, country, etc. It should capture the user’s behavior (interesting topics, ratings, patterns, goals, etc.) when interacting with the various social media platforms. User modeling is the process of gathering information about a user’s interests, constructing, maintaining and using user profiles. For example, in e-business systems, it captures online users’ characteristics, finds similar online users, provides customized products and services, and therefore improves the user satisfaction.
There is no doubt that in recent years, creating the profile of individuals has become increasingly important. Companies of various areas, regardless of their size, try to target a specific audience based on their products or services\(^\text{13}\). For this, they need to know more information about the profile of an individual to adjust their speech to him/her. For example, interviews could be adapted in an appropriate format based on the candidate’s profile in order to be a more friendly process. In recommendations systems, such techniques are used combining individuals’ purchasing choices and their demographics (Kanoje et al., 2014). User profiling has also been used in domains such as healthcare (Bhardwaj et al., 2017), smart home (Cesco netto et al., 2020), buying behavior (Celli et al., 2017) and social networks. User Profiling is widely used to achieve those goals, and thus several researchers have been consistently proposing approaches for an effective and efficient User Profiling in their scientific articles. User profiling techniques have widely applied in the various web search, user-adaptive software systems, web user identification, personalization, recommendation, e-market analysis, intelligent tutoring systems, intelligent agents, as well as personalized information retrieval and filtering. The content and amount of the information within a user profile can differ depending on the application area. The accuracy of the user profile based on how the user information is gathered and organized. Moreover, how accurate this information reflects the user while building the profile. Farid et. al. in (2018) refers that there are two approaches in user modeling process: The generation of an initial user profile for a new user and the continuous update of the profile information to adapt to the continuous changing of the user’s interests, preferences, and needs (Farid et al., 2018). A major challenge within user profiling is how the user profile can be built to accurately reflects users’ preferences. In the following, we categorized these articles which we have studied in the context of this work through the process described in section 2.1. On top of that, Kitchenham methodology was used to reveal new research directions and identify the gaps that may exist in this scientific field. Our main findings are: a) about data sources and b) profiling aspects. The first concerns the collection of data from various sources like Twitter (Volkova et. al., 2016; Greco & Polli, 2020), Facebook (Kanoje et al., 2014), gameplay (van Lankveld et. al., 2011; Ma et al., 2017) etc. as well as a combination of these data. The information collection is important because it determines how the user profile will be created using the available features provided by the platforms (van Lankveld et. al., 2011; Volkova et. al., 2016; Ma et al., 2017). The second, relates to the basic idea of this work's psycho-demographic aspects. We categorize articles according to the different aspects of the person they are focusing on. In section 2.2.1 we present the data sources and data gathering methodologies which have been proposed. In section 2.2.2 we describe psycho-demographic aspects and the prediction labels that have been used in the existing papers and in 2.2.3 we analyze the innovative methodologies for each of the human aspect.

### 2.2.1 Data gathering - Data Sources

The principal phase of user profiling is to gather data about the user. Obtaining and extracting user data and knowing his or her interests is a major task for making personalized systems. There are several types of data that can exist in a user profile: user personal data, user background knowledge and skills; user needs and goals, which are static data types (Farid, et. al., 2018). Alternatively, user’s behavior, user’s interests and preferences, and context are dynamic. Data gathering, initially, includes the gathering or relevant and suitable

\(^{13}\) [https://en.wikipedia.org/wiki/Market_segmentation](https://en.wikipedia.org/wiki/Market_segmentation)
data sources which are used to extract data. This data is different across different platforms, i.e., different data is provided by Twitter compared to the data available in Instagram. Two types of data are used in the literature in order to extract results from the data sources. The first one is to use the data from a platform individually to get insights (Ikeda et al., 2013). For example, Twitter data can be used to build a user’s personality. The second one relies on data fusion (Kanoje et al., 2014). In this case, a combination of different data sources is used, which could lead to better and more accurate results. Specifically, a combination of Twitter and Foursquare accounts of an individual can be analyzed together to get better insights about an individual. For instance, in the combination of these two platforms, text content from Twitter along with locations from Foursquare can be used to provide a more complete user profile about user’s preference and interests (Farseev et al., 2015). We have to mention that social media platforms may not be able to provide personal data due to restrictions which have been applied from the Cambridge Analytica scandal.

There is no doubt that the data collection is an important task of any machine learning supervised technique, basically, because data needed in all three algorithm processes. The first one, which has to be completed before the predictions, is the dataset creation to train machine learning models in case that supervised technique is used. The second includes the statistical analysis that need to be done. In this case, data can be collected directly from users using questionnaires for example. However, this is not preferred as users most of the time do not want to disclose their private data and thus, they may provide false, inaccurate or incomplete data. The third technique is performed by collecting data that are not obvious using advanced techniques for analyzing user’s activity and drawing conclusions (Eke et al., 2019).

The results extracted from ML techniques may come from a variety of data sources, such as social media, blogs, exhibitions, browsing applications, etc. It is fundamental that the data used to train these models are similar to the data that will be used to make the predictions. For example, to predict a tweet, the features given to the model must have been extracted from the same data. This is because, for example, in an academic article the way of writing is more formal than in a tweet. Twitter is a widely used data source of the researchers (Ikeda et al., 2013; Ahmad & Siddique, 2017; Vasanthakumar et al., 2019) because: a) provides an easy-to-use free API, b) is widely used (Statista, 2021), and c) is able to provide accurate results. Nevertheless, the main limit of using Twitter data is that there is only public data which adds a lot of bias. That occurs because there are many people who upload content only for their friends, while the public accounts are more possible to be organizations or public people (Kaushal et al., 2019).

Data from OSN (Twitter, Foursquare, Instagram and Facebook):

Verhoeven et al. (2016), created a corpus for MBTI personality models using 6 languages, namely TWISTY. This dataset contains 18,168 individuals on twitter. They searched for users that have one of the MBTI in their profile and manually add labels for gender. Afterwards, they used Twitter API to collect tweets from the users that were found in the previous step. The corpus contains gender labels as well as indicators for Myers-Briggs personality type. In another study, tweets are also collected, and the labels are applied manually by the authors, but they kept only the Japanese text (Ikeda et al., 2013). Farseev et al. (2015) created a dataset from Twitter, Foursquare, Instagram and Facebook. The integration of the data sources is widely known as data fusion. Those platforms used to get
an extensive overview of the profiles. The account mapping is achieved using the Foursquare API.

Data from different OSNs:

Kaushal et. al. (2019) focused on collecting data from many social networks. To make a complete profile could combine all the available data for a user. To achieve a more accurate result integration of different OSNs are very important. In this paper the author suggests 5 methods to integrate the data from different sources. These methods are Social Aggregator, Self-disclosure, Cross platform sharing, Advanced Search Operator and Friend Finding Feature.

Dataset from questionnaires:

The collection of the data can occur by questionnaires as well (Catalin et. al., 2019; van Lankveld et. al., 2011). Catalin et. al. (2019) used a questionnaire filled in by the individuals who participated in the survey. The questionnaire contained psycho-demographics aspects such as age, sex, location, the university where they are studied, the year of study, questions about the time spend on Facebook and the knowledge about the Facebook interactions of each user. Also, they tracked the behavior of users. In detail, scrolling, mistakes written in text, way of correcting text, deletion of text, mode which the text read, how they write text and how they scroll. Questionnaires may not be so accurate sometimes because each person can fill in some information that he believes for himself, but it might be inaccurate.

Finally, Farid et. al. (2018), introduce explicit, implicit, and hybrid data collection methods, while they considered that data collection can be manual, semi-automatic, and automatic profiles respectively. The accuracy of the user profile depends on the amount of generating data through user-system interaction. Moreover, where the gathered or processed information is maintained; the personalization process in terms of when and where the information is available to be exploited for personalization. Figure 2.3 presents their proposed classification schema of research studies on user profiling. The modeling process studies first the information constituting the user profile and the gathering of raw data about the user and how it can be extracted. Second, the building and maintaining process are investigated, which way for representing a user model and the techniques of its construction. Third, how to identify the user and how to explore the applications in the context of user profiling in order to provide personalized services.
Explicit Data
The explicit user information collection approaches require manual techniques through user intervention and interaction including registration forms, questionnaires or in cases of recommendation systems, by asking users to rate items, or by tracking users’ query words. For instance, eBay asks users to provide their opinions and to give ratings for the services and products. Afterwards, the company makes uses of these data to improve the personalized recommendations that it gives to future customers. Generally, in recommender systems, explicit rating data provided by the services users are widely used to profile users’ preferences. Tags are also commonly used as direct interest keywords when they are attached by the user to web content or using social websites (Farid et. al., 2018). In some research works, researchers utilize users’ posts and comments in social media platforms directly to extract keywords to represent users’ interests, while others directly use the rated items as an indication of users’ interest (Farid et. al., 2018). The explicit profiles have also a static nature and are valid only until the user changes their interest and preference parameters explicitly. To overcome this problem, extraction of the information implicitly by tracking the users’ activities and/or behavior is required (Farid et.al., 2018).

Implicit Data
The implicit method attempts to extract the user’s data from the information collected automatically without any effort or intervention from the user. Unlike static profiling, automatic profiling uses the implicit method and analyzes the user’s behavior pattern. The implicit user profiling can be achieved in client-side or server-side. Farid et.al. (2018) refer that browser agents may gather user data interactively, while the user browses. For this, an application or a browser extension must be installed on the user’s desktop computer to capture user activity and actions performed.

Hybrid Data
Hybrid user profiles are gained by semi-automated techniques with low user involvement. The system collects data combining implicit and explicit methods. This approach helps to profile more efficient and maintains the accuracy of temporal information as information gets updated temporally (Farid et. al., 2018). For example, both explicit and implicit profiles can be generated independently and then combine into a single profile. In which the explicit
user profile would be limited to asking the users to specify their interests in terms of keywords. The implicit user profile could be, for example, created by constantly monitoring user’s activities by storing the browsing data, the time spend on each page, and additional actions like printing or saving. Such approach, it combines explicitly state interests with the observation of user behavior.

2.2.2 Aspects of user profiling in social networks

Extracting a person’s profiling aspects is important for both companies and individuals, as they can compose a complete user’s profile that might be used for several purposes and applications. Some of these reasons are targeted ads, age limit required by certain professions, way of communicating with people through knowledge of his personality, etc. Sentiment and emotion used widely to get insights after launch of products, services or events from organizations. A user’s psycho-demographics aspects can be categorized as demographic and psychographic. The first category includes age, gender, ethnicity and location, while psychographic category includes buying behavior, interests, sentiment, emotion and personality traits of individuals such as the characteristics of included in the Big Five model. This subsection describes the features that have been proposed in the literature as well as the predictions that can be made based on them. For example, which labels will be used to predict age, for example “25 y.o.” or “33 y.o.” or its preferable to divide in age groups such as “25-33” and “34-55”.

There is no doubt that many different profiling aspects of each person are used in the research of user profiling. Most of them deal with a specific characteristic while others deal with a characteristic category (e.g., demographics). The most common of these are age and gender. In addition, common tasks are the personality predictions based on the Big Five model (Bayot et. al., 2015; Celli et. al., 2017; van Lankveld et. al., 2011), as well as the purchasing behavior of the users. Additional characteristics are emotion/sentiment (Chatzakou et. al.2017; Greco & Polli, 2020; Shamantha et. al., 2019), needs/values (Ma et al., 2017), gender (Verhoeven et. al., 2016; Ikeda et.al., 2013; Bayot et. al., 2015; Volkova et. al., 2016; Ma et al., 2017; Qiu et. al., 2019; Noecker et. al., 2013; Jurgens et. al., 2017; Ma et al., 2017), age (Ikeda et. al., 2013; Bayot et. al., 2015; Volkova et. al., 2016; Ma et al., 2017; Qiu, Li & Zheng, 2019; Jurgens et. al., 2017; Ma et al., 2017; Kanoje et al, 2014), ethnicity (Ma et al., 2017; Noecker et. al., 2013), interests (Kazai et. al., 2016; Cui et. al., 2019; Mantilla et. al., 2013), topics (Vasanthakumar et. al., 2019; Ma et al., 2017) and places of interests (Ikeda et. al., 2013). The datasets that are used, differentiate in the focused aspects and the data sources. In the following, the two big categories of psycho-demographics aspects are analyzed (i.e., demographic aspects and psychological aspects), while other aspects of profiling are also discussed.

**Demographic Aspects**

Demographic analysis is the study of a population based on factors such as age, race, and gender etc. Demographic data refers to socioeconomic information expressed statistically including employment, education, income, marriage rates, birth and more. However, in the context of this work and based of the most common demographic aspects that are used in the literature for user profiling in social media, age and gender demographics have been considered and they are analyzed further in the following.
Age
Predicting the age of individuals is very important\(^{14}\) in applications such as recommendations systems, advertising, etc. (Jiang et. al., 2015). It can also be used in Human Computer Interaction (HCI) and sensitive content restrictions. In particular, in an advertising campaign, it would be preferable to display items such as specific fashion clothing brands for customers who are younger. Furthermore, age prediction can be applied in cases such as employment to police and military (Trivedi & N., 2020). Usually, predicting age can be achieved through age segmentation which means that the predictions are about groups of age. Specifically, by age group we mean that the models predict age approximately, for example “Customer X is in age group 25-34”. To predict the age of individuals, Ikeda et al. (2013) used the age groups of 10s, 20s, 30s and over 40. Bayot et. al. (2015) train ML models to predict age groups of the age groups 18-24, 25-34, 35-49 and 50-XX while in Volkova, et. al (2016) and Ma et al. (2017) the groups consist of the binary classification “Below 25” and “Above 25”. In another approach, categorization is based on year groups in which someone was born (Qiu et. al., 2019). The groups consist of a) -1979, b) 1980-1989, and c) 1990+. Usually, the distributions are not equal in size and especially in the cases of younger ages there is a disadvantage as the samples from the datasets are not representative (Jurgens et. al., 2017). We can observe that there are many ways for age grouping, and to find the best for our applications, we have to consider the specific goals and purposes of each scenario.

Gender
Extracting a person's gender is a very common practice in advertising in order to target specific audiences. Also, it can be used in order to prevent discrimination\(^{15}\). Gender aspect prediction is about predicting “male” or “female” class (Verhoeven et. al., 2016; Ikeda et. al., 2013). Therefore, it is a binary problem and there are several approaches to predict it. Ikeda et. al. (2013), used about 100,000 Twitter users for predictions. The accuracy in the binary predictions, like gender, was very high (84.5%) but in the aspects which had many labels especially in hobby the results were very poor (37.7%). Intuitively, a gender prediction is possible to occur from text, comparing the writing style, topics and interests.

Psychological Aspects
In the articles that we have screened, the most commonly used model for personality is the Big Five model (Bayot et. al., 2015; Celli et. al., 2017; van Lankveld et. al., 2011). The others which we found in the papers are the DISC (Ahmad & Siddique, 2017) and MBTI (Verhoeven et. al., 2016; Noecker et. al., 2013) models. Jurgens et al. (2017), focus only on Extroversion/Introversion aspects, while, Ma et al. (2017), focus on extrovert, intuition and feeling personality psychological aspects. Also, many other personality models proposed in the bibliography which we will not discuss at this moment as they are out of our scope. There is no doubt that tasks such that to get the users personality, is a challenging task and many approaches have been proposed.

The **OCEAN** model, also known as Big Five, is derived from the initials of the following individual characteristics:

---

\(^{14}\)https://www.uspsoig.gov/blog/age-matters-advertisers

• **Openness**: This term concerns how receptive a person is to new things / activities and how easily he / she can adapt and understand them.

• **Conscientiousness**: This term is about how efficient one can be and how consistent one is.

• **Extraversion**: The third characteristic, Extraversion, concerns sociability with other people and activities, such as having a large or small participation in discussions and express their opinion.

• **Agreeableness**: Agreeableness shows how friendly one is and how likable can become to others.

• **Neuroticism**: This term shows how sensitive and nervous one is. It is the only negative feature contained in the model.

Another model used for user profiling identified in the literature is the DISC (Dominance, Influence, Stability and Compatibility). This model takes into account that the individual's behavior is dynamic and can change depending on the situations he/she experiences. MBTI, also known as ISTJ or ENFP, apart from the following features: a) Intuition / Sensing, b) Introversion / Extraversion, c) Feeling / Thinking and d) Perception / Judging. In some cases, it has not been considered effective for user profiling (Bailey et al., 2018).

**Emotion - Sentiment**

Awareness and knowledge of people's emotion and sentiment is fundamental in order to understand how certain individuals feel about an activity, event, service or product, how they like or dislike. In this way the improvement of certain activities can be analyzed, or, on the other hand, this kind of situations can be avoided in order to ensure users’ comfort keeping them away from uncomfortable and inconvenient situations. On top of that, an individual’s emotion about a certain situation or product may reveal valuable information regarding other demographic or psychological personal aspects. According to Sailunaz et al. (2018) “human emotions can be related to their age, gender, time, location, ethnicity, political views, educational qualifications, in addition to some other individual, social, temporal and spatial parameters”. Furthermore, Chatzakou, et al. (2017) predicts to which category the text belongs, specifically which emotion characterizes the sentence / text. It is desirable to approach emotion-sentiment aspect as a multi-class problem because a single sentence can contain more than one emotion-sentiment state. For example, "It's a nice product but too expensive", this sentence has two emotions. We have to mention that, sentiment can be classified as positive, neutral and negative (Shamantha et. al., 2019) while for emotion, most commonly used classification, comes from Ekman’s theory, which classifies in fear, anger, joy, sadness, disgust, and surprise (Ekman, 1992).

In order to predict certain emotion-sentiment in a sentence, the use of good and accurate dictionaries is essential. In particular, the dictionaries that have been suggested and used to predict sentiment give a percentage to each sentence, and, in most cases, classify as neutral, negative or positive. Vader gives the result for the neutral, positive and negative along with compound. Compound score based on the official site\(^\text{16}\) is “computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 and +1”.

---
\(^{16}\)https://github.com/cjhutto/vaderSentiment
Table 2-3: Aspect categorization of literature.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Psychological aspects</th>
<th>Demographics</th>
<th>Social aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Emotion/Sentiment</td>
<td>Personality</td>
<td>Gender</td>
</tr>
<tr>
<td>(VERHOEVEN ET. AL., 2016)</td>
<td>MBTI ✓</td>
<td>✓ ✓ ✓ ✓</td>
<td></td>
</tr>
<tr>
<td>(IKEDA ET. AL., 2013)</td>
<td>✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(AHMAD &amp; SIDDIQUE, 2017)</td>
<td>DISC ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CHATZAKOU ET. AL., 2017)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CATALIN, ALIN, MADALINA &amp; BOGDAN, 2019)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(BAYOT ET. AL., 2015)</td>
<td>Big Five ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(KAZAI ET. AL., 2016)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CELLI ET. AL., 2017)</td>
<td>Big Five ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(VAN LANKVELD ET. AL., 2011)</td>
<td>Big Five</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(VOLKOVA ET. AL., 2016)</td>
<td>Life Satisfaction, Narcissism ✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(QIU, LI &amp; ZHENG, 2019)</td>
<td>✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(VASANTHAKUMAR ET. AL., 2019)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CUI ET. AL., 2019)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NOECKER ET. AL., 2013)</td>
<td>MBTI ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(NO ET. AL., 2016)</td>
<td>Big Five ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(JURGENS ET. AL., 2017)</td>
<td>Extroversion-Introversion ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(MA ET AL., 2017)</td>
<td>✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(MANTILLA ET. AL., 2013)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(GRECO &amp; POLLI, 2020)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(KARANATSIOU ET AL., 2020)</td>
<td>Big Five ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(SHAMANTHA ET. AL., 2019)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(KANOJE ET AL, 2014)</td>
<td>✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROPOSED WORK</td>
<td>✓ Big Five ✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

USER PROFILING ASPECTS
Other aspects of profiling
Furthermore, we provide some information for aspect which their labels consist of text which is not standard. Specifically, someone’s interest can be sports but specifically football which is also acceptable. The following aspects are part of the previous categories, psychographic or demographic. Ma et al. (2017) and Noecker et. al. (2013) predict nationality/ethnicity. Labels can be based on continental prediction or even countries of origin (Ma et al., 2017). Interests are predicted from Kazai, et. al. (2016) and they rely on the location and feed of individuals. In another approach location of the tweets extracted to get the interests of the individuals (Cui et. al., 2019). The predicted values are keywords like “Traveling” and “dining”. The predictions of topics also contain appropriate keywords of discussion (Vasanthakumar et. al., 2019). It is obvious that the predictions of these aspects are more flexible and have no standard labels.

Table 2-3 presents a summary of the research work that we have collected and studied regarding what user profiling aspects have been used by researchers. It is worthwhile to highlight that all scientific papers studied have used machine learning techniques to predict values and extract results. However, the method of collecting personal information and the data used for training the ML algorithms and predicting values that authors in the literature have used are quite different depending on the application and the purpose of user profiling. In our work we collected users’ data from social media, and specifically Twitter accounts.

The columns of the Table 2-3 indicate the user profiling aspects of the individuals used according to the literature which we screened. In particular, the categories of user profiling may fall into the following categories:

- Psychological aspects, including emotion/sentiment and personality aspects,
- Demographics, including individual’s gender, age and ethnicity aspects,
- Social aspects, including interests, topics, places of interest, buying behavior etc.

Most approaches focus on the user profiling aspects of personality, age and gender. It is obvious that any of the existing works deals with all the available features. A subset of the total aspects is used in each study for User Profiling. In our work we have added all the available user profiling features in order the develop platform: a) to provide a complete view of each user's profile, and b) to be easily adapted to any application and use case purpose.

2.2.3 User profiling proposed algorithms

The user profiling process is particularly difficult and requires complex algorithms to extract an efficient result. Several methodologies have been proposed for the aspects mentioned in the previous section. Some of them can become quite complex. We can categorize them into two super-categories based on the data that needed for the analysis. The first one is knowledge extraction from the content (multimodal) and second is the knowledge extraction from the social network. In some cases, both of them are combined to achieve a better and more accurate result (Ikeda et. al., 2013). In the first category, the data that are used to get the insights can be of any type, text, images, video or audio. For the social network data, the connections and interactions of each node (person) are used. In both categories, machine learning techniques are mainly used. These are supervised learning (Catalin et. al., 2019; Qiu, Li & Zheng, 2019), unsupervised learning (Catalin et. al., 2019; Greco & Polli, 2020)
and semi-supervised learning (Celli et al., 2017; Ma et al., 2017). Below we present novel approaches for the user profiling aspects. In supervised and semi-supervised approaches, a machine learning model has to be trained from existing data. These data, consist of the features and actual label/s. Features are extracted from the content (text, image, and video). The label is an annotation which included either manually or from automated process (van Engelen & Hoos, 2019). With that process the models are trained to make predictions for new instances. Unsupervised approaches are self-trained and the results deriving from clusters created by the models (Kanoje et al., 2014).

Verhoeven et al., (2016), the corpus creation outlined, and some baseline predictions of each language presented. Their datasets are about the following six languages “Dutch, German, French, Italian, Portuguese and Spanish”, while (Bayot et al., 2015) predict for Dutch, English, Italian and Spanish. Text features extracted after preprocessing following a typical ML training process in both works. Gender prediction is promising while MBTI personality is a demanding task. Specifically, two of the characteristics of MBTI, I-E and T-F, can be extracted from text. The efficiency in the other two is very low (Verhoeven et al., 2016). Bayot et al., (2015), predict age, gender and personality traits (Big Five model) from models which have been trained from text features.

Furthermore, Ikeda et. al in (2013) focused on the predictions of demographics such as Age, Gender, Area, Occupation, Hobby, Marital status combining tweets and the community that users belong to. They use a hybrid approach which is based on both context and social media community for collecting their data. In total, authors collected 244,345 twitter user's profiles. From the profiles which were gathered only about 3.5% had some of the demographic’s aspects in their twitter profile. Some terms were extracted as predictors such as school or university which can match to a teenager or a 20s person, respectively. This kind of topics can provide insights for further grouping the users based on their interests. Each of the aspects had proof that can be grouped/predicted with some keywords.

Catalin et. al (2019) focused on user’s behavior and activity (i.e., scrolling, typing and reading) in social media and grouped them into clusters. A total number of 359 participants took part for the research purposes in the context of this work. Specifically, user behavior tracked in order to collect data which will be used in order to get user profiling insights. The behavior was examined as the way that a participant uses the social media application. For this reason, two apps were developed and embedded in Facebook web app or mobile app to track user’s behavior. Furthermore, Ma et al. (2017) used an application that records the user’s eye reactions. The person's eyes are recorded while watching videos to analyze their personality and demographic characteristics analyzing data features which came out from the eyes of the individuals. In detail, demographics and personal traits (e.g., extrovert, intuition, and feeling) were able to be predicted. This is an instant method to get the user profile without a data retrieval from different data sources, or the manual data input of the social media users.

A novel and interesting approach is given by (van Lankveld et. al., 2011) which is about personality user profiling using games. In particular, a game was used to extract features which would be used to analyze an individual's profile through their behavior in the particular game. Players were invited to answer the "NEO-PI-R" questionnaire before playing the game. Afterwards, a total number of 44 participants were asked to play the game in order to collect traits and analyze their behavior in it. From this analysis, a total of 275 features were collected. The researchers categorized the collected features into the following
two categories: pooled and unpooled.

Unpooled variables were 226 in number and were those which were behavioral based, such as movement, conversation and miscellaneous. On the other hand, pooled variables are those which have an effect across areas of the game. All characteristics were related to any of the five characteristics of Big Five and therefore they concluded that the personality of each person can be distinguished through the way a game play. Celli et. al., in (2017), developed a platform which is able to extract Big Five using demographic info and multi-modal context to make the predictions. Semi-supervised, deep learning and SVMs classification algorithms were used to train models. Ahmad & Siddique in (2017), used DISC model in order to create the profile of a user's personality. Each user’s account was retrieved to group them in the aspects based on some keywords. The personality of each user included characteristics, emotions, behavior as well as the thought of each individual that differs him from the others.

Volkova et. al, (2016) analyzed whether the user's interests are related to psychometric and demographic characteristics. The data that they used collected from Twitter. Except the context data from tweets the analysis also included the users that follows. The aspects that were predicted in this work were for gender, age, educational background, political and personal aspects. They created the dataset and the labels added by Mechanical Turk. Many accounts had no content and, in some cases, researchers used features from their “neighbors” to be able to use their profile for their application. This article focused on user interest in demographic predictions (gender, age, educational background, political stand and personality). Predictions for personality aspects were narcissism and satisfaction. Kazai et. al, (2016), proposed a mobile app for news recommendations. In this app, news feeds were recommended to users based on their interests which implicitly were exported by their “location, social media (Twitter, Facebook) feed and their interactions with the app”. Finally, the interactions of the users were used in order to correct and improve their model.

Another proposed approach identifies the activities that a user does during the posting of a tweet (Cui et. al., 2019), while Noecker et. al, (2013) considered that the way each person writes can include some elements that can be attributed to his personality. However, the content of the tweet may differ from his activity. Specifically, semantics can be exported in order to connect content of text to a user activity when he is posting a status in social media platforms. For example, an individual can exercise or take his breakfast while writing about a recipe. This can be useful because it is directly related to the interests of individuals. As a result of building the user profile on the aspect of “interests”, a company could preview related ads to certain individuals. In most research papers the analysis of a person's characteristics is done through the personal data in his profile and especially the text in comments or posts. Jurgens et. al, (2017), however, take into account what person's connections say about themselves. A challenging task is to finally combine all features that are introduced by the user content and the network structure. Ma et al. (2017), combine data from user's connections as well as from user context. Furthermore, Vasanthakumar, et. al (2019), predict topics/interests of discussion based on certain clusters. In particular, tweets are clustered to the interpretative group. The result is based on the total count of the discussion of the individual. The higher related keywords with the higher count number classify the individual’s topic and label in the appropriate cluster.

Nevertheless, only a small number of users post their location in online platforms. Gated Recurrent Neural Networks are used for such analysis. The assignment of the tags to the
tweets is done automatically and is based on the location. The authors use Hybrid-LSTM and analyze some more architectures. In another study, data ingestion model analyses the content of incoming streams (Mantilla et. al, 2013). The content is passed through quality filters and media and other features are extracted. Duplicates or similar contents are removed after filtering. Logistic Regression classifiers are used to assign topic labels. Content is then indexed and stored in Elasticsearch.

Chatzakou et al. (2017) proposed a hybrid approach to detect emotion/sentiment of an individual. Considering that both lexicon and ML techniques present various disadvantages, they used a hybrid approach. Both techniques are combined in order to overcome deficiencies that arise when using these techniques separately. Their process was based on text from different social media platforms. In particular, they collected results from lexicons and they passed through these results to machine learning models as features. Additionally, they extracted sentimental features along with text. The predictions were the emotion which exists in each sentence as a class, for example “This is a lovely place” could be predicted with class joy. Furthermore, they predicted the primary emotions based on Ekman's theory and the social ones (Ekman, 1992). Ekman's theory contains disgust, fear, joy, anger, surprise and sadness while the social ones are enthusiasm, anxiety, calm, interest, rejection and shame. Shamantha et. al., in (2019) trained a Naive Bayes Classifier, Random Forest and SVM machine learning classification algorithms to apply sentiment analysis in Twitter stream. A basic task of the sentiment analysis is the recognition of emotional states.

Studying the literature review we found that there are a variety of methodologies for user profiling. In addition, the way that the data are collected and the data sources used differ for each of these works. Nevertheless, our conducted Systematic Literature Review revealed existing gaps in the scientific field of user profiling. Initially, none of the existing tools create a complete user profile. On the contrary, most of them only focus on a small number of profiling aspects. In the following, the most popular available tools that have been implemented for social media user profiling are discussed in detail.

2.3 User profiling in social networks: an overview of the most popular and open available tools

This chapter reviews some of the most popular tools that have been implemented for social media user profiling. These tools can be useful in many different use cases, such as in recruiters who need to analyze the personality of their candidates or marketers (FlexMR, 2021) who need to understand the personality and demographics of their targeted audience. In general, a deep understanding over the users to facilitate the process needed for these applications, these platforms offer rich capabilities and features in understanding human profiles and behavior by categorizing individuals after collecting or predicting their data and building a profile with the desired profiling aspects. In the sections below, we describe social media profiling tools and some social media listeners as well as some modeling (mockup) tools with emphasis on creating not only a user-behavior-on-platform profile but also socio-psychological image of individuals. We focus on social media profiling tools which provide more details and implicit intellective information about social media users. These tools are widely used and could impact large audience.

There are many tools available online. However, most of them require a subscription to use
them. In our work we include a number of them which are freely available or with a free trial. Many tools were rejected from the provided analysis because they do not focus on socio-psychological features or they have similar functionality to those included in our work.

**World Well-Being Project (WWBP)**

This project was developed at the University of Pennsylvania for the analysis of demographic and psychometric characteristics of individuals. It is the result of a collaboration between scientists of several scientific fields. Specifically, computer scientists, psychologists and statisticians collaborated to achieve the completion of this project. The OCEAN personality traits (Openness, Conscientiousness, Extraversion, Agreeableness and Emotional Stability) as well as demographics, age and gender, are used for the analysis. The dataset consists of 75000 Facebook users who participate in the process and have their status analyzed. The frequency of each occurrence is shown at different ages. It comprises a research tool and is free for use by anyone without requiring a subscription. The features of this tool are divided into 5 apps. The first concerns predictions of demographic characteristics, age and gender. The text in the status of each Facebook profile is used and, based on the words it contains, an analysis is feasible. The user has to connect to his account to give access to this app and permit it to conduct the analysis. This feature is no longer supported after Facebook forbade the access to the API due to the Cambridge Analytica data scandal.

![Instance from language analysis in personality space](https://wwbp.org/langTopics.html)

The word clouds app displays the most frequent words used categorized by age, gender or personality. The most common uses per category are provided for brief analysis. There is the possibility for the user to select the age range he wants and to display the corresponding

17 https://wwbp.org/langTopics.html
results. Similarly, he can choose a specific gender or personality and, in all cases, retrieve relevant Word clouds. The interaction map shows measurements about the personal characteristics of each place in the United States analyzing the most recent public tweets. In the last application each user can, by entering some words, retrieve a result about a possible age that could contain these words in his text. This is based on the word frequency that appears in the dataset.

**Crystal**

Implemented since 2015, its main purpose is to reveal if people could work efficiently together based on their personality. Crystal is a tool that allows users to understand their personality or the personalities of other people and compare each other. The tool provides the results about the personality with an analysis of the content of each individual. A user that wishes to assess his personality or see someone else’s profile can manually do the self-assessments that are available in the platform, completing a plethora of questions. There are two options to fill the personality information in the profile, those that are verified (having completed the questionnaire) and those who are predicted using a resume or their LinkedIn profile. There is also the option to get the data from LinkedIn using the site or via an extension for the Google Chrome browser and make the prediction automatically without having to fill in all the answers. The results, of the extension, are shown in a browser window. Also, it is used for communication between individuals or customers using the appropriate writing for different personalities. For example, if an email adapts to a person's personality, they will be more likely to accept the purpose of the message. This is possible due to the plethora of results and analytics offered through the platform.

The predictions produced are made for several personality model, namely:

- DISC,
- 16-Personality,
- Enneagram and
- Big Five (OCEAN).

These complement each other by covering each other's shortcomings. There is a limit of ten searches for the trial version. There is a subscription for unlimited predictions and some other privileges like personality comparison and the prediction of another individual’s personality.

**SocialBakers**

This is a project that provides a thorough analysis over the data of many social media platforms such as Facebook, Twitter, Instagram, YouTube etc. The analysis is based on data provided by each platform which means that this tool uses the available explicit data in the user’s social media account to provide valuable insights. This tool can be described as a social listener which provides insights such as location, age, gender, interests etc. about the individuals. The results can be used by targets companies which are willing to increase the popularity and influence over a specific audience of users. On top of that, tweets, likes, reposts etc. also facilitate the analysis in this platform. The analysis contains the interests of each individual and demographics (age, gender). Also, sentiment analysis provided. A free
trial offered for 14 days. After the expiration of the trial version, a subscription is required that one may use the full capabilities of the tool.

**Xeerpa**

Xeerpa is a Spanish company based in Madrid. This is a complete tool that offers many options for analyzing people's insights on their social network. Provides capabilities for companies to control their entire clientele. A profile is created for each user which includes a lot of information. More accurately, it offers an analysis of demographic characteristics, psychometrics, people's activities, behavior and the influence of each individual. It can be connected with social login in many social networking platforms such as Facebook, Twitter, LinkedIn, Instagram etc. Xeerpa is then executed to perform all the actions desired by the user. In addition, data is possible to be received from the Wifi hotspot. Anyone who is trying to connect to the internet provided by a business is asked to connect to the social login to gain access and receive some user information. When a user logs in to a social network, he/she is asked to provide some information from his/her profile.

For the analysis of individuals' personality, Xeerpa uses the "Watson Personality and Customer Behavior Insights" from IBM. This makes an extensive analysis of personality which is analyzed below. Furthermore, an analysis can be conducted based on the point in time that someone started writing for a company, i.e., it offers a history analysis. Through its user-friendly interface, many options are offered to the user for better data analysis and based on what interests him. This is achieved by watching people who are interested in two categories and analyzing only those categories. It is a commercial product and requires a subscription to render its features accessible.

**IBM Watson Personality Insight**

IBM Watson Personality Insights is a product of IBM, one of the most well-known companies in the field of computer science. It is one of the many solutions offered by the company through its cloud. It is not limited to connecting to any social media, like the other tools, but can analyze through the text that will be given by the users. The features which are offered by the service are Personality Characteristics, Consumer preferences and needs. There are up to fifty-two psychographics which provide a deeper understanding of the people involved in any case. The consumer preferences consist of 8 categories and more than 40 preferences. Creates a reasonably good profile of a person just passing some text in the API.

A free tool for some controls provided by bluemix which serves as a tutorial. Also, a twitter account can be parsed. This is done by analyzing the tweets and comments of each person. Permission may also be granted to analyze the user's own account. After the analysis, the characteristics of the individual's personality, consumer needs and values are displayed. A small description about the individual is also provided as a result along with some habits that someone has. It is also possible to provide an analysis via a text. This text should be of a reasonable size so that the results are valid and accurate. This product is available through an API and an account login is required for results. The number of calls is limited depending on the subscription that users will pay. There is a free trial for 30 days so that one can see the capabilities of the tool.
TalkWalker

TalkWalker is a CRM platform which is used in business or organization’s analytics and can integrate their social media into the platform. This tool provides some capabilities such as social listening, social media analytics, reporting and data intelligence. Social media analytics are more demanding and provide some implicit information of the customers to the company. Provides insights concerning the customers among other analytics about the company. These insights include behavior analysis, age, gender, sentiment and interests. These customer aspects are provided through social media as well as the influence of each brand. A brand can divide customers into meaningful groups with demographics data like interests and occupations which leads to better handling of the audience. The data can be collected from many social media types. Social networks include Twitter and Facebook, social media sharing like Instagram, Flickr and Youtube, discussion forums and others.

One important feature of the tool is image recognition that can be provided. For example, a company’s logo can be detected by analyzing users' images which provides important information for campaigns or the preferences of users. Provides an API to retrieve results in a timeframe of 30 days. It is an industrial product which is provided with a subscription. Some free tools are provided like alerts from social media, hashtag trends and some more which complement the main platform.

Roialty

Roialty is a customer profiling tool. It targets companies which are using the available information and can create a user profile to make it easier to analyze the interest and buying behavior of their customers. There are more than three thousand topics which can be assigned in a user’s profile. It can track profiles to analyse the behavior of each customer and segment them to clusters of relevant interest and behaviors. Additionally, it analyses demographics, interactions, locations and influence of individuals. The data are gathered from different social media. There are many options for the company for a marketing campaign. It can pinpoint the relevant audience and the right influencers, optimize and personalize the content in each customer and spot new marketing trends. The core of the platform is a REST-like interface which can be used. Through an API all the functionality is provided. This tool is commercial and needs subscription for an individual or a company to be able to use it.

Sonar Platform

It’s a platform, based in Indonesia, which offers analytics in companies to improve their customer suggestions. It offers market intelligence, influencer analysis and user profiling among others. It includes several capabilities which are a complete solution for a business. It uses text analysis and machine learning methods for the predictions of insights. Demographics (age, gender), interests and user places of interest. It consists of two main tools: a) Sonar Analytics and b) Sonar Influence. Sonar Analytics provides automated analysis of conversations about a company. Offers sentiment analysis over the conversations about the products’ events. It can also discover trends and find out which people are talking about the trend. Sonar Influence finds people who have a great influence on Instagram. The detection can happen based on the relevant topics of the company. Daily reports provided to have a complete view of the campaign. The platform can integrate information from many
resources. Supports more than 670 Indonesian online media portals, Facebook, Twitter, Youtube and Instagram profiles. It's an industrial product and in order to use it a subscription is required. There are two options for subscription: the white label and on-premise.

**URL Profiler**

URL Profiler tool is a desktop application providing many capabilities in data collection and analysis. As referred in the website, this tool is useful in situations like content audits, backlink analysis, competitive research and penalty audits, as depicted in Figure 2.5. The collection of the data can be completed from different resources and many tools can be combined for a more fleshed out result. It supports many social media platforms, like Twitter, Instagram etc. but including all of them could lead to delayed results. Also, it can connect to many other data sources (uClassify, Google analytics etc.) which need special permissions and access.

![URL Profiler Features](image_url)

**Figure 2.4: Features of URL Profiler.**

In order to specify the desired data for user profiling there is a list to import files in a proper format or add URLs. The output of the program is in an xlsx format. This file contains all the previous added options (data sources) and provides many insights through the analysis from the resources. The insights include sentiment analysis, MBTI personality model, topics, preferences of the user and language detection. URL Profiler is available under a subscription. There are two plans, monthly and annual payment schedules. Also, the price of the product rises as the more users are added to the plan.

**Other tools**

There are many tools available similar to those described above which are about user profiling. Most do not offer implicit resolutions but are limited to the data provided by social networks through the API or the user explicitly states them. Other useful tools based on the personas like the xtensio tool, which only offers a design of a user’s profile. This tool provides templates that include a lot of information for individuals but these must be entered manually and doesn’t provide any insight of a user or a customer. Hootsuite is a platform to
2.3.2 A comparison among user profiling platforms and tools

We have analyzed more than 10 tools/platforms in our research study. These tools offer personality and insight analysis to companies which are interested in their audience. Some of them provide analysis in characteristics like demographics, interests and no-textual data of the people while some others are more complex and provide a more in-depth personality analysis. At this point, it should be noted that the information has been obtained by a demo or a description of the website, in the case of tools that could not be tested in action.

As can be seen in Table 2.3, there is a gap in the already implemented tools with emphasis on the personality of individuals. In particular, only virtually half of them provide insights about personality, while most of the ones that they do, they use the Big Five model. The most common insights are about demographics (age, gender) because it’s easier to predict through the context provided by the platforms like Twitter. Furthermore, interests and topics about the individuals are a common feature of the tools. An interesting observation is about the data sources. Social login is required by some tools to add many of the platforms and some other uses the open API of Twitter and other blogs, news sites etc. Except WWBP project, the other are industrial products which are available under a subscription. The most complete tool which we have found is Xeerpa. However, this cannot be used by the general public as it needs subscription to be able to use its capabilities.

Due to the fact that the tools are industrial, no explanations about the functionality of each tool are given. Therefore, the efficiency of each tool is not known. The information/insights are affected by the data source. For example, in cases in which the text is formal probably the predictions could be misleading. Also, we observe that any of the available tools don’t provide explanation/interpretability about the insights that each tool provides. Explainability is an important task in cases someone wants to understand the reason why a decision was made. Explainability can prevent human bias and in such a case understand the people’s personality and how affected by several factors.

We propose a novel tool which overcomes the previous described shortcomings. In detail, we develop a tool with improved models for the socio-psychological aspects. In addition, provides interpretability/explainability about the aspects. Also, it is a free to use tool, without subscription requirement, and open source.

Most of the social media platforms now restrict the access to the data provided. Some tools are suspended due to their involvement in Cambridge Analytica scandal\(^\text{18}\). Due to this, restrictions and laws applied to protect privacy. This can lead to a lack of knowledge because many of the information that is provided by social media, now are hidden.

Table 2-4: Capabilities of the existing platforms

<table>
<thead>
<tr>
<th>Tool</th>
<th>User Profiling Aspects</th>
<th>Data info</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Psychological aspects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social aspects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source</td>
<td>Availability (Open API)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explainability</td>
<td>Functionality</td>
<td></td>
</tr>
<tr>
<td>Emotion / Sentiment</td>
<td>Personality</td>
<td>Other (needs, values etc.)</td>
<td>Gender</td>
</tr>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mediator</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2-4 summarizes the capabilities of existing User Profiling tools. Each column represents psycho-demographic aspects which the platform predicts. Apparently, the age, gender and emotion / sentiment characteristics are predicted by most platforms. Also, we have to remind the reader that the age and gender are the most common according to the literature review of the previous subsection. Personality predictions are made on four platforms and are about Big Five model personality traits. The rest of the features are user preferences and behavior. In addition, the data used by the platforms which we set as “SNs” are obtained from the account of individuals in the social network. So, an account is needed to provide insights from the SNs with the user’s consent.

Our tool, Mediator, aims to fill the gap that exists in the literature by covering almost all aspects that: (a) we have identified and analyzed in the literature review, and (b) the existing
tools which we presented in the section are using for user profile analysis and modeling. In addition to the plethora of insights about an individual’s profile, another important feature of our developed platform is the ability to provide explanations for every prediction made by the AI models. These explanations provide information so that the user is aware of the reason of why every decision was made by the machine learning algorithms. Also, as we highlighted previously, except WWBP, all existing user profiling tools are industrial and thus subscription is needed to be able to be used for user profiling analysis and modeling; none of them is open source. On top of that, for the development of our platform, we are using Twitter to obtain user’s data and the insights extracted from user’s posts (tweets) and their profile data.

2.4 Data Ethics Issues in User Profiling

In recent years, artificial intelligence (AI) has been increasingly used to support many applications including recruiting and human resources (HR) departments. For example, in the past, the usual way to complete a process, such as taking out a loan or applying for a job, was to be checked by someone and be accepted or rejected it regarding your personal information or by peer in reviews. In recent years, however, to speed up these processes machine learning techniques are used for immediate results. These techniques/models, also known as agents, can deliver thousands of results in a matter of seconds, facilitating the workload of employees who would otherwise have to do the work themselves.

Nowadays, the idea that AI will replace humans in the workplace is widespread. These automated results are the predictions of the pre-trained machine learning models used for each use case scenario. Nevertheless, predictions made by these models often lead to uncertainty. On the other hand, these predictions are not always correct because there is a possibility of software failure or lack and/or inaccurate data. Machine Learning algorithms mainly rely on accurate and trustworthy data for, firstly, developing efficient models to, afterwards, predict values. A failure can be considered a False Positive, e.g., accept a loan which the recipient will never be able to repay, and False Negative, for example reject a candidate who was fit exactly to a company’s open position. Even the most efficient models have a chance of giving the wrong result, and there is no 100% accurate machine learning model.

In addition, supervised learning, as mentioned in the Chapter 2, is based on datasets which may contain problems or misinterpretations. These faults can emerge by the number of instances that datasets contain, possible biases and misinterpretation of the models. In the case of unsupervised models, the accuracy is lower and is usually categorized based on similarity of the provided data (“Supervised vs. Unsupervised Learning | IBM”, 2021). Furthermore, class imbalance, missing data, and distribution inconsistency could lead to non-appropriate results due to misleading training of the models (Haibo He & Garcia, 2009). These problems can arise from physical or human factors and special handling required. These techniques cannot detect errors and misleadings, only in some cases where there is a small number of “outliers” they can be able to overrun and improve the machine learning model without being affected.

Even if a prediction is correct, for example a person correctly predicted as female, there is no guarantee that this decision was taken based on appropriate criteria. In other words, the
prediction may be correct but the characteristics that have led to this decision may be the wrong ones. In such cases it is desirable to apply explainability techniques to expose irrelevant behavior. There are some popular explainability/interpretability techniques such as LIME (Ribeiro et. al, 2016), SHAP (Lundberg et al., 2020) and surrogates. LIME perturbates the input features to detect the most efficient features. SHAP\textsuperscript{19} "it connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions". Surrogate models are a combination of a white box and a black box. Specifically, a white box model is trained to approximately provide explanations from the predictions of a black box model. The LIME technique is used in our work which we describe below.

Critical applications necessarily have to interpret the models to avoid any forms of prejudice and discrimination. Another characteristic advantage of using explanatory is the avoidance of prejudice. Fairness can be affected by several reasons such as:

1. Features that may include bias.
2. Dataset creation - training of the model.
3. Parameters of the models.
4. Human bias.

There is no doubt that by choosing some features that are considered important factors for the predicted results, prejudices or discrimination may be added in the model. For example, if gender is included in the feature set which will train the machine learning model, bias also included because that feature is not a factor to any decision. Furthermore, prediction of the incomes of an individual should not contain as parameter the origin, because this introduces race bias and contributes to prejudice and thus it is desirable to avoid such cases. Nevertheless, there are specific cases where such feature selection cannot be avoided, such as whether there is a pregnancy or not, which gender is required as a feature.

In addition, the creation of datasets can be also affected by factors such as whether the collected sample is representative. Specifically, the distribution must be equal and contain each case. It is quite a challenging process and especially when it involves people coming from different geolocations (i.e., from all over the world). In this case, the predicted results may be misleading, so this requires further caution.

Furthermore, during the training of the machine learning models, the parameters and features that have been added should be taken into account. In the first case, a better result can be achieved by slightly modifying the parameters of the model to the appropriate ones. For the second case, values that are very different from each other must be escalated in the same value range, also refer to as data normalization. If a feature such as the income - working hours, has a large difference in scale the first feature may affect a model less than the others. This feature produces results that are not correct. So, all values should be normalized and the value range must be the same for all features. Finally, outliers or missing values should be removed from our datasets as either they do not contribute at all in adding accuracy to our model, or, in the worst case, they may lead to inaccurate results.

\textsuperscript{19}https://github.com/slundberg/shap
Last but not least, special attention must be given to human bias. Based on Wikipedia, there are many kinds of human biases which while the creation of a dataset could be included, on purpose or without intention. In particular, humans generally tend to not have a neutral viewpoint, rather than a more subjective position. This can be caused by differences in their culture, personal experience or social discrimination. In some places, laws are existing in order to prevent human bias and discrimination. An advantage of using ML for predicting values is that computers do not add bias in their prediction, but they objectively predict values according to certain features and parameters. On top of that, when they are able to present explainability descriptions along with their results, they can ensure transparency of their predictions.

An example of bias could be prejudice against women if for a company’s open position selection if mostly the candidates got hired are males. In this case the accuracy percentages of the ML models may be lower for women than men. By applying explainability techniques, problems can be identified and resolved so that they do not lead to wrong decisions. Furthermore, these techniques can be used to debug and find the right features for model training in code level. In addition, we can see that the accuracy percentages are better for lighter individuals than darker. Nevertheless, this is likely to be product of less or inaccurate datasets and data sources rather than software bias.

Some machine learning models by their structure can be easily interpretable with no need for further explainability. For example, linear models and decision trees can be represented in a human-readable format. Linear model’s predictions can be a simple illustration of the features which contribute to the result, while decision trees can be a representation of the feature’s importance in a tree schema. These models are called White Box models. Usually, they contain simple calculations, and they are not very complicated in terms of their structure which can lead to loss of accuracy. In contrast, black box models are more complicated models, perform complex calculations and require an adequate amount of data to be effective and efficient. Therefore, their interpretation might be difficult requiring time and special techniques to be achieved. Examples of these machine learning models are Random Forests, Neural Networks, and SVM. On top of that, these models are not feasible to represent in a simple schema based on their structure such as white box model.

In a recent study, the LIME technique has been proposed which is widely used (Ribeiro et. al, 2016). LIME is a model-agnostic technique which means that every ML model can be used explained despite their complex implementation and structure. LIME perturbates the input features to detect the most efficient features. This way, it is able to provide which of the features are more valuable. The most valuable feature, probably leads to the predicted results. However, it concerns predictions of an instance and does not interpret the whole model. These techniques are called Local interpretable models.

As discussed previously, the explainability is one of the most important parts when developing ML models. In the regulation of algorithms, particularly artificial intelligence and its subfield of machine learning, a right to explanation (or right to an explanation) is a right to be given an explanation for an output (prediction) of the algorithm. Such rights primarily refer to individual rights to be given an explanation for decisions that significantly

---

20 https://en.wikipedia.org/wiki/Cognitive_bias
affect an individual, particularly legally or financially. For example, a person who applies for a job and is rejected based on user profiling information may ask for an explanation, which could be "Your activity in social media platforms shows that you are a dogmatic person, while the specific job position requires a more open-minded individual. Thus, you are not suitable for this job, but at the same time, it seems that you wouldn’t like this position as well; this is the main factor in considering you too likely to quit the job in a short time, and thus we will not hire you for the job position you applied for." Many libraries have been widely used for this purpose. In our work, we will make use of the LIME library (Ribeiro et al., 2016), which is available online, as well as it is implemented in python as the rest of the developed components of our platform. LIME library implements techniques for enabling explainability over AI results.

2.5 Research Gap on Current User Profiling Methods in Social Networks

Studying the literature review we found that there are a variety of methodologies for user profiling. In addition, the way that the data are collected and the data sources used differ for each of these works. In particular, data sources are social media, or various platforms such as videogames and webpages that track user’s behavior, as presented in detail in section 2.2.

Nevertheless, the Systematic literature review and Technology review revealed existing gaps in the scientific field of user profiling. Initially, none of the existing tools create a complete user profile. On the contrary, most of them only focus on a small number of profiling aspects. In particular, regarding the tools that are freely available, none of them contain as many aspects for an individual. Finally, we noticed that in the existing works or platforms, there is no the capability for explainability of the predicted values.
Chapter 3

3. Proposed Methodology & Approach for User Profiling in Social Networks

This chapter describes the methodology and approach that we have followed to develop our User Profiling Platform Mediator. In the following sections, an overview of the system, the architecture of the models, as well as the process flow which is required until the final results are provided. As already described in the previous chapter, although there are several tools and platforms that make predictions of psycho-demographic aspects, none of the existing approaches enables explainability / interpretability over the predicting values. However, in our case, the models follow specific architectures in order to be able to use techniques to enable explainability/interpretability while achieving comparable prediction accuracy with the state-of-the-art models. At first, we present the system architecture and the components that we developed in the context of this work, as well as how the ML models are "loaded". Afterwards, we describe how the data were collected, the features were selected and how each model was developed separately, for each psycho-demographic aspect that we studied. In particular, the processes of data preprocessing and feature extraction that we used to develop our machine learning model are described in detail. On top of that, we present the features that are used for each model prediction. Furthermore, we describe the model’s approaches that were used and the labels that were used for the prediction. Finally, we refer to the explainability/interpretability function and how it works, as well as the LIME technique, as a common technique used for local explainability.

3.1 Workflow – System Architecture

The purpose of this research is to provide predictions about a person's psycho-demographic aspects along with their explanations on how this classification occurred. These predictions require clarity, i.e., the reason why the models gave these results, as well as which features are more important and contribute more to the final results. Those results have to be in a format that is easily readable and perceivable by the users, so the results should be given in a user-friendly interface. Through our developed app, the user can enter the Twitter profile (@username) of the person who wishes to analyze his profiling aspects from their Twitter account. Our application works only for public accounts and not for the cases of restrictions and private accounts. The platform gets the required data that are needed for this analysis. A visual representation of the user interaction with the tool is provided in Figure 3.1.
As depicted in Figure 3-1, the developed platform, after entering the username, communicates via the Twitter API and receive relevant information about the specific user for further analysis of their profile (see steps 1.1 and 2.1 respectively). This information includes the recent tweets, specifically the last twenty, which have been published by the user and his account information such as the number of people following or followed by the account. With the appropriate pre-processing, it will be possible to provide the results to the end user of the application. This could last a few seconds until the final results delivered to the user. Machine learning process is described in the following while the system architecture presented in Chapter 5.

![Figure 3.1: User interaction with the platform – Basic overview.](image)

![Figure 3.2: Machine learning process for User Profile building.](image)
The complete machine learning process/flow, that is followed by our developed platform is represented in Figure 3.2. Initially, the user inserts the username of the user that he/her want to analyze to the platform. After, at step 1, the data preprocessing takes place. In particular, this process converts the gathered data from the user’s Twitter profile to a specific desirable format and removes the redundant words and symbols. In case that the data is tabular, which means is in a format of an array, step 1 is skipped. Furthermore, the feature engineering referred to the process which is used to extract features. Also, features have to be transformed and normalized in a common format and value range. These features will be used in model training at step 3. Each model corresponds to a specific prediction, for example there is one model to predict age and another one to predict sentiment. Finally, the predictions of the models combined to build the user’s profile. User Profile is sent to the user’s User Interface (UI) and is, then, represented in an appropriate and understandable format.

All the steps followed (i.e., pre-model phases, building classification/regression model, and explainability/interpretability), as shown in Figure 3-2, are described in more detail in the following sections of this chapter.

### 3.2 Pre-model Phases

This section describes in detail all the phases that the data followed until the final results. In particular, these phases consist of: 1) dataset preprocessing, 2) feature extraction, 3) model training and 4) user profile build. In our case, the process of ML predictions requires the data collection from the Twitter social media platform. The user’s input should be a username which corresponds to a Twitter account. This account’s data will be used to make several different predictions, such as user’s age, user’s emotion state in certain tweets etc. After this process, our platform merges the results of each distinct model in order to build the user’s complete profile and redirect the output to the UI or API.

#### 3.2.1 Data preprocessing

As described in the previous section, we gathered publicly available datasets to train each ML model (i.e., for every different psycho-demographic aspect) separately using certain features based on the literature, and then we collect tweets and data available in user’s profile to export these features and apply our developed ML models in order to make some predictions about the user’s psycho-demographic aspects. However, tweets contain many punctuation marks, unfamiliar words, acronyms, numbers, links, etc, that might add difficulty to the understanding and analysis. To address this issue, we follow a procedure for cleaning the data. Specifically, this procedure is performed using the well-known Bag of Words (BoW) techniques as features in order to be used by the ML model. As we will also discuss in the following, in certain cases, we have extracted features based on the user profile or text which is in a tabular format. Such features might be the number of mentions, numbers of hashtags, character flooding, etc., and they do not require to be “cleaned” because in that case, we would have a significant loss of data.
In particular, cleaning data consists of the following steps (Vijayarani et al, 2015):

1. **Text tokenization**: Tokenization is one of the most common tasks when it comes to working with raw text data. In particular, in order to get the data into a “more” structured form, a tokenization process must take place. According to tokenization, every sentence is represented as an array of tokens. Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of the smaller units are called tokens. This process helps to extract textual content in the form of bag-of-words, a form necessary in most cases for further data preprocessing facilitating the understanding of raw data by a software. After tokenization, the document is presented as an array of strings separated by comma.

2. **Remove stopwords**: Stopwords are common words that are non-valuable for interpreting the meaning of a phrase. In most cases, they do not provide any semantic value to the text, on the contrary, they are just fillers for making the text more aesthetic for human readers. For tasks like text classification, where raw text should be analyzed in order to be classified into different categories, stopwords are removed and/or excluded from the given text so that more focus can be given to those words which are more valuable towards defining the meaning of the text. Stopwords include, for example, words like the, he, have etc.

3. **Remove punctuation and technical marks**: This step concerns the deletion of punctuation marks that may be contained in a tweet as well as urls, mentions, numbers etc. These elements do not add any information when it is to be used with BoW. Instead, they can reduce the performance of the model, and thus they are removed from the text analysis.

4. **Lemmatization of words**: It is the process that turns words into lemma form, because, in social media platforms content, usually, text in tweets, posts or comments is not in a formal format. In this way, it is possible for machine learning models to recognize the word and give it the same meaning effectively. Otherwise, the words would be completely different because machine learning models cannot understand and interpret semantics.

5. **Remove empty strings**: After removing stopwords, characters and marks in the previous phases, an empty string might be created. This string, then, should be removed from our dataset because it only adds extra time and effort for the machine learning models to analyze it, while it adds nothing to the accuracy of the model.

6. **Remove words with less than 3 characters**: we noticed that words that have less than three characters either do not offer any information or are not a dictionary word. So, in some cases, we removed all these words in order to improve the performance of our machine learning models.

### 3.2.2 Feature Extraction

A vital aspect in every prediction task that utilizes machine learning algorithms is choosing meaningful features that distinguish predictions effectively and efficiently\(^{22}\). There will be

---

\(^{22}\)https://cogitotech.medium.com/what-are-features-in-machine-learning-and-why-it-is-important- e7f9905b54d
features that improve the performance in terms of accuracy of the machine learning algorithms, while others might worsen the results and thus, they should be excluded from the training of the ML model. In the context of this work, we extracted two types of features: a) **text features**, which are represented by words detected in text (i.e., tweets), and b) **content features** which are data extracted directly from the Twitter profile of the user and refer to their behavior on the platform. Text features can be extracted with Bag of Words (BoW) techniques. The two most common techniques for this type of feature are TF-IDF and Count Vectorizer, which are covered in the following. Furthermore, content features include features that can be extracted from data, user profile and text, such as the total number of hashtags/mentions, number of capitals letters or number of emojis that detected to an individual’s profile. Our feature selection is based on the most common used in literature as well as in our conducted experiments. More details are provided below.

**TF-IDF**
In data science, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that intends to reflect how important a word is to a document in a collection or corpus (“tf–idf - Wikipedia”, 2021). In this project, we are using the TF-IDF representation of the textual content of the dataset in order to be able to identify the words in our datasets with significant value to the output and exclude others that they do not offer anything to the accuracy of our model. In this way, we can apply further data preprocessing, if needed, and efficiently train our machine learning models. More precisely, TF-IDF is a Transformer which takes sets of terms and converts those sets into fixed-length feature vectors. In text processing, a “set of terms” might be a bag of words.

**Count Vectorizer**
CountVectorizer (CV) transforms text documents collection to a matrix of token counts. Those token counts are the total number of occurrences that exist in a document. This technique creates a sparse table to store the attributes.

After testing separately for every prediction model, the most efficient feature extraction techniques (i.e., either CV or TF-IDF) have been selected respectively, that are the ones that gave the highest accuracy. In each case, in addition, parameterization has been done, for example, words that appear less than 5 times, as well as those that exist in a percentage greater than 85% are not taken into account. All these techniques have been applied at the word level. These are automated processes and do not require further processing of the text. Also, we have created words from unigrams and bigrams in all cases.

Apart from BoW, we have also extracted features from the user profile. In order to extract as many features as possible for better accuracy, we used both profiles and tweet’s data. In particular, we extracted features which are based on the content of the profiles and the context of tweets. For example, a feature of the first category constitutes the number of social media user’s followers provided directly by Twitter using its API. The second category is based on the data that are retrieved from user’s tweets, such as the number of mentions in the text (tweets) or hashtags. The combination of these two data categories can give very useful results and, in this work, we will refer to them as behavioral features (BF). Context features used are presented in Table 3-2, while in Table 3-3 the content features which extracted by user’s profile are shown. It is worthwhile to highlight that each prediction model performs better with a specific set of features. So, using all the available features on all prediction models does not necessarily mean that they will perform better.
Table 3-1: Features extracted by the tweet content.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Gender</th>
<th>Emotion</th>
<th>Sentiment</th>
<th>Personality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of mentions</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>number of hashtags</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of URLs</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>character Flooding</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>number of punctuations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>number of capitals letters</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Words avg length</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Sentiment pos/neg score</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>number of emojis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>emojis emotion</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>avg words in tweets</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Size of tweet</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of stopwords</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negation in tweet</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years Active</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Favorites Count</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Followers Count</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friends Count</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statuses Count</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Verified</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Likes</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description Length</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Username length</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

The Table 3-1 presents all the features (see Column 2) found in the literature that can be retrieved from the tweets and metadata of a Twitter social media user in order to predict one or some of the following ``labels´´ about the social media user: (i) age, (ii) gender, (iii) emotion, (iv) sentiment, and/or (v) personality. The type of the features, content or context as mentioned previously, are shown in Column 1.

3.3 Building Classification/Regression models

In this section, we are presenting, in more detail, the machine learning models and architectures that we designed and developed to predict the psycho-demographic aspects about a Twitter user. Our primary goal was to create an application with models that can be interpreted by the end user. Therefore, in order the developed models to be interpretable, a range of more sophisticated solutions such as the Artificial Neural Network which have been shown to be highly accurate, should be excluded as they function as a ``black box´´,
which is not easy to be interpreted by the end user. Towards this direction, we have strategically chosen to use simpler ML models that achieve acceptable accuracy while ensuring the significant advantage of interpretability balancing the two desirable but incompatible features of accuracy and interpretability in our solution. In the following, we also present details about the tradeoff of the two approaches integrating LIME technique and providing explainability/interpretability in our results.

In order to achieve the best efficiency, all model’s hyper-parameter optimization has been performed using the technique RandomSearchCV. This is a commonly used technique which tests different parameters for each model in order to identify the one(s) with the best result.

3.3.1 Gender Classifier

As mentioned in Chapter 2, gender prediction is a binary classification problem, which means that the prediction values, also refer to as labels, consist of two discrete values: male or female. Binary classification models, are usually, more accurate because there are only two classes in the prediction and thus, a model can distinguish those labels more efficiently. In particular, they have better results in terms of accuracy on their predictions because only one of the two labels is about to be predicted.

It is worthwhile to mention that it is too complex to combine behavioral features and BoW in one model and, for certain model, to integrate and use the explainability technique. Nevertheless, the explainability technique which we have used to explain the predictions is able to support several types of input values. In our case, we used only two of them, one concerns content data and the other Text data. It is obvious that the text data is referred to the BoWs, while the tabular are about the set of behavioral features.

In order to predict the gender aspect, we combined available data of both types, namely behavioral and text data. The extracted results prove that the text model itself could perform much better and thus we selected it as our main classifier. Gender model was trained by using the features extracted from the Count Vectorizer feature extraction technique with data contained from the PAN-18 dataset (Daneshvar & Inkpen, 2018). We concatenate the tweets of each Twitter account (Bayot et. al., 2015). After that, we performed preprocessing steps to clean text. More specifically, we perform a) tokenization, b) stopwords removal, c) reformat urls, mentions, and hashtags to tokens URL, MENTION and HASHTAGS respectively, d) we applied lemmatization, e) removed words with size less than 3 characters, and finally removed empty strings (Daneshvar & Inkpen, 2018). After that, features extraction is done using the CountVectorizer technique. As reported by Alowibdi et. al., (2013) and Abdallah et. al. (2020), POS tags could improve the results but due to time delay upon the results and not significant performance improvement we didn’t use in our model, although the results are given in Section 4. The BoWs that were created by the Count Vectorizer, then, passes to the classifier along with the labels. Our selected classifier is the well-known and commonly used Random Forest classifier which we selected throughout a performance process comparing with other classifiers.
Figure 3.3: Gender model instance prediction.

To provide explanations of the predictions, we used LIME in the model and the results are provided to the user in a user-friendly interface. So, the user should be able to understand the decisions made by the model and recognize whether a model predicted correct or not, given appropriate reasons that led to the particular decisions.

In Figure 3.3, a new instance’s prediction depicted. Data are preprocessed and the relevant features are extracted and then the model is applied to make the final prediction alongside with the explanations provided. For the training of the model, the process is similar except the last step of the explanations that provided by LIME.

3.3.2 Age Classifier

The second aspect which we added to our developed platform was the age prediction. This challenge can be a regression one by the exact age or classification one, using age groups as classes (Bayot et. al., 2015). For our application, we address the classification problem using datasets with age groups data. At first, we trained a classification model which predicts age groups. There are several approaches to divide age into age groups as mentioned in Section 2.2. The groups that were used for our results are:

- 18-24
- 25-34
- 35-49 and
- 50-99

The dataset used as ground truth, namely PAN-15, contains a very small volume of data. In this case, the results would not be very accurate. Firstly, we concatenate the most recent tweets for the user that we desire to analyze (Bayot et. al., 2015; Daneshvar & Inkpen, 2018). After that, we preprocessed the tweets following the process described in Section 3.1. By properly tuning a random forest classifier, we achieved about 77% accuracy in the test set.

Nevertheless, to make sure the results are not affected by small volume of training data itself, we used a semi-supervised technique to label some more instances (Bilgin and Senturk, 2017). Specifically, increasing the size of the dataset makes the model more flexible in new instances that previously were unknown. Based on the first model which was trained from the initial dataset, we made predictions using instances from a second dataset which we created with random users collected though Twitter API. Regarding the new instances’ labels, we set an acceptance limit which refers to the probability that the prediction from the ML model is accurate, for instance, the higher the value is, the more certain we are that the model has correctly predicted the snapshot. Afterwards, the number
of instances that exceeded the limit was added to the original dataset. The model is, then, trained again with all the new labeled data. This process occurred a great number of iterations. After that, and when the percentage of predictions is below the limit for all new instances, the process ends. The final data set includes the original data, as well as those ones that were labeled during the semi-supervised process, as described previously.

Figure 3.4: An example of Voting Classifiers\(^2\). At first, for the semi-supervised process, we used a random forest (RF) classifier and we set the probability of 75% to accept a new instance in the new dataset. We observed that the RF did not resulted probabilities with more than 70%. So, finally we used a Voting Schema (Abdallah et. al., 2020) to train the model for age prediction, and expand our dataset, because it gives results with a higher degree of accuracy than an individual model. Voting classifiers combine the predictions of many classifiers to predict an instance. There are two options: a) ‘soft vote’ which means that the probabilities of the models are combined for the final result and b) ‘hard vote’ which predicts based on the most predicted label. Four tuned models are used to perform the best possible and after our experiments we choose soft voting performed better. These are: Random Forest, SVM, Naïve Bayes and Logistic Regression.

3.3.3 Sentiment Classifier

As described previously, sentiment classification is about predicting if a tweet contains positive, negative or neutral opinions. To achieve sentiment analysis, we are using exclusively text data. The process is similar to the one which described in Section 3.1. In particular, the text retrieved from user’s tweets are transformed and cleaned up from redundant data. After that, a Count Vectorizer is used to extract features while discards words that appear with a percentage of more than 85%. Furthermore, we discarded tokens that do not exist more than 3 times in the dataset. After applying this, we discard uncommon words and reduce the size of the dictionary which is used by our BoW technique. A web application requires efficiency and fast results so we had to consider the time response along

with the results. Along with the features extracted by the BoW we extracted features from the context of tweets. To train our model we include the features that presented in Table 3.1 which are those that are used the most common in literature. More specifically, we extract number of emojis, negation in text, number of punctuations and emojis emotion scores, number of positive/neutral/negative words and negations (Bahrainian & Dengel, 2013).

![Diagram](image)

*Figure 3.5: Sentiment classifier training steps.*

We used these features to train the sentiment model. The dataset which we used as ground truth to train our model is the Sem-Val (SemEval, 2016) which we briefly describe this in Chapter 4. Specifically, we trained four classifiers with the appropriate tuning in order to achieve the best results. We combine their results with a voting schema as it is shown in Figure 3.5. More specifically, the class with the higher percentage is the result. The predictions are made for each available tweet. The final results consist of a percentage of each category and the most important tweets that extracted by the model. We have to mention at this point, that with this approach to provide explanations we are able to use Vader, as well as any other model interchangeably along with our model without much effort, attaching them with a few changes to our platform.

### 3.3.4 Emotion Classifier

The emotions prediction from the profile data may also be predicted as classification or regression problem. In classification approaches, the existence or not of a particular emotion in a certain tweet might be labeled with 0 or 1. On the other hand, in regression approaches, the percentage of each emotion could be calculated and provided. In our work, we have used the classification approach, so that we can extract the percentage of each emotion for the user profile accounts that we desire to analyze. Specifically, our goal is to provide the percentage of each emotion category that is identified in the Tweets that social media users have posted in their Twitter accounts.
Using the dataset from Sem-Val, we selected only the basic emotions in order to create a simple, but more understandable and logical model. In particular, the basic emotions, as found in the literature, are anger, disgust, fear, joy, love and sadness (Chatzakou et. al., 2017; Sailunaz & Alhajj, 2019). In this case, we could predict the categories as a multi-class problem. That means a post can belong to more than one emotion, with a certain percentage. On top of that, each post may belong or not to one of the categories.

In our model, the data collected by Twitter are used to predict the six emotions. For each emotion, a model has been trained to make predictions only for the specific value. The used features were extracted from text data (i.e., the tweets) (Chatzakou et. al., 2017). On top of that, behavioral features were excluded, because they did not perform efficiently for the particular classification problem. Although, we used emoticons, punctuation marks, URLs, mention and hashtags (Chatzakou et. al., 2017). Also, we replace negation words with token SEMANTIC_NEGATION as it is obvious that such words could be important for this task. Furthermore, TF-IDF has been used to create the features for the predictions (Bostan, 2018). For each tweet a prediction is made, and all the output results were added together to extract the percentage of the user profile. There are many predictions for each Twitter account. For these predictions, in order to provide explanations, there is also the need to summarize all predicted values. Specifically, for each decision, the feature importance returned by LIME is added, and finally the most important and valuable features are returned to the user. In Figure 3.6, the architecture of the model for emotion prediction is depicted, while the difference between sentiment and emotion user profiling analysis is presented in Figure 3.7.
In particular, sentiment can be defined as “the underlying positive or negative” feeling, attitude, evaluation, or emotion associated with an opinion” (Liu, 2015). Based upon the definition of sentiment, the terms “emotion” and “sentiment” do not significantly differ from each other as a sentiment seems to be an underlying emotion. However, the difference is, emotions are states of an individual, whereas sentiments are assigned properties of an object, e.g., an opinion. For example, a person might say about a user interface that it would be “frustrating”, which means she associates the interface with an emotional state of frustration. Thus, the property assigned to the user interface is “negative.” The person expects that the interaction with this UI leads to frustrating emotions (Frijda, 1994; Brave & Nass, 2009). Figure 3.7 visualizes this difference.

### 3.3.5 Personality Regressor

For the personality traits of the individual, the characteristics of the Big Five model have been selected and create a separate regression model for each of the personality traits. In this case, the labels have a continuous value between 0 and 5. The higher the value is, the more the person is characterized by this aspect. The prediction of one trait does not affect the others, that is, one can have all 5 values high or all five low or a mix of them. In Figure 3.8, each of the Big Five traits given with examples for better understanding of what those continuous values mean for a person. Low and high score contain examples of each trait about an individual while the trait (center column) describes the domain of each of them.
For the purposes of this work, only behavioral features have been used for these features. This is because models that output continuous value predictions are not compatible with the LIME library for ensuring clarity and transparency over the predictions. As our primary goal was to develop an interpretable model and due to this compatibility problem, the final set included only the behavioral features. More specifically, we extracted behavioral features (see Table 3-1) from the user’s profile using the dataset that are included from Karanatsiou et al, (2020) which have 243 user ids along with their personality traits score. After that, we used a robust scaler to normalize the features values to contribute to each prediction equally and avoid possible loss (“Minimax - Wikipedia”, 2021). In our models we trained five Random Forests for each of the personality trait. We used random searchCV on each of them to optimize their tuning. Finally, the results are pass through the LIME library to provide explanations to the users.

3.3.6 Location, Topics & Ethnicity

Due to limited available datasets, it was impossible to develop and train accurate ML model, and thus for the predictions of these location, topics and ethnicity aspects, we strategically chose other techniques to identify them. In addition, extracting the location by LDA technique, as preference would be highly questionable as it would provide additional information that was not directly linked with location. Due to user privacy concerns, we chose methodologies that are based on statistics using the text that are posted by each user.

---

Figure 3.8: The Big Five personality traits, acronyed Big Five.24

24 https://media.cheggcdn.com/study/e11/e115c2a2-ac03-41c5-8602-ff5ca900c50a/715614-11-47PAQEII.png

for topics and ethnicity, and user’s geolocations coordinates for location extraction.

First of all, a person’s location may be extracted from their profile, if exists, as most users do not fill this field explicitly. In addition, using Geocode, coordinates are returned automatically for the given location through tweets. On the other hand, ethnicity can be derived from the language which is used in user’s profile. So, the tweets received are processed to identify the languages an individual uses to their posts. The language usage percentage is displayed to the user. However, in this case, we could not claim that the language with the highest percentage would indicate a person's ethnicity with accuracy 100%, because the common use of English, as an international language worldwide, would negatively affect the output results.

3.4 Explainability / Interpretability

These two concepts, namely Explainability and Interpretability, are often used interchangeably. But there is a significant difference. According to KDnuggets (2021), “interpretability is about being able to discern the mechanics without necessarily knowing why. However, explainability is being able to quite literally explain what is happening”. Explainability has been touted as a way to enhance transparency of ML systems, particularly for end users. Often releasing (or forcing organizations to release) the data that models were trained on or the accompanying code is challenging due to user privacy issues. Moreover, end users are generally not equipped to be able to understand how raw data and code translate into benefits or harms that might affect them individually. By providing an explanation for how the model made a decision, explainability techniques seek to provide transparency directly targeted to human users, often with the goal of improving user trust. The importance of explainability as a concept has been reflected in legal and ethical guidelines for data and ML. In cases of automated decision-making, Articles 13-15 of the European General Data Protection Regulation (GDPR) 26 require that data subjects have access to “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject” (GDPR). In addition, technology companies have released artificial intelligence (AI) principles that include transparency as a core value, including notions of explainability, interpretability, or intelligibility. 27

Towards this direction, there is no doubt that explainability in machine learning (ML) and artificial intelligence (AI) is becoming increasingly important composing a fundamental requirement so that the emerging ML and AI applications to gain the trust of all involved stakeholders and reach their full potential in the market. More precisely, explainable machine learning seeks to provide various stakeholders with insights into model behavior via feature importance scores, counterfactual explanations, and influential samples, among other techniques. In the recent years, with the increased demand for explanations and the number of new approaches out there, it could be difficult to know where to start. In particular, several techniques have been developed to explain the performance of machine learning model predictions. This can be explained for a variety of reasons as was discussed

26 https://gdpr.eu/
27 https://deepai.org/publication/explainable-machine-learning-in-deployment
When it comes to explaining the models and/or their decisions, multiple approaches exist. Generally, there are two types of explanations for Explainability: a) global (overall) model behavior and b) local explanation (i.e., explain the decision of the model about each instance in the data). In particular, the first concerns the explanatory nature of the features in the model, namely how important and how much they contribute to the predictions. In the second case, the explanation is given for a specific snapshot, i.e., the characteristics that contributed more to this particular snapshot may be different for other predictions etc. Apart from that, some approaches are applied before building the ML model, while others after the training of the model (post-hoc). In addition, some approaches explain the data, some others the model, while some might be purely visual and others not. It all depends on the final application that we are aiming to develop and what it better fits to it.

For the purposes of this work and taking into consideration all the requirements about explainability in our ML models set previously, the output of the ML models, before being provided to the end users, passed through the LIME explainability technique. LIME is short for Local Interpretable Model-Agnostic Explanations. Each part of the name reflects something that we desire in explanations of our ML models. Local refers to local fidelity - i.e., we want the explanation to really reflect the behaviour of the classifier "around" the instance being predicted. This explanation is useless unless it is interpretable - that is, unless a human can make sense of it. LIME is able to explain any model without needing to 'peak' into it, so it is model-agnostic technique. Consider a loan approval model, what if a user request is declined, then the user has the right to question WHY? and authorities should know why the model has declined the user request and communicate the same to the user as they just can’t say that their system has rejected it instead, they need to explain on what factors(features) the request is rejected. Furthermore, LIME can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model (e.g., linear regression, decision tree, etc.). It tests what happens to the predictions when we feed variations of the data into the machine learning model, and it can be used on tabular, text, and image data.

Regarding the algorithm’s implementation, there are two types of ML models: a) white box and b) black box models. White box are models that are easy to be explained and their structure facilitates the representation of decisions, such as decision tree or linear classification algorithms. While black box models are difficult to be explained as they perform complex predictions. In our work, we mainly used the second case because they seem to have better accuracy. In addition, they can usually be used with a large number of features effectively and efficiently.

On our platform, we have used the LIME tool which gives explanations for a snapshot. As we said previously, LIME is model-agnostic, meaning that it can be applied to any machine

---

28 https://medium.com/@ODSC/how-to-explain-your-ml-models-e7d7f6a65d96
29 https://www.analyticsvidhya.com/blog/2021/06/beginners-guide-to-machine-learning-explainability/
learning model. The technique attempts to understand the model by perturbing the input of data samples and understanding how the predictions change. LIME assumes a black box machine learning model and investigates the relationship between input and output, in order to give explanations over the results. It can be used with any machine learning model, as well as for different types of data input. Specifically, text, images and tabular data can be efficiently imported.

Figure 3.9: LIME prediction example (Bars show the importance of the features in each class. Red predict class 0 and green class 1)

In conclusion, in Chapter 3, we discussed the proposed methodology & approach for user profiling in social networks. First of all, we provided the proposed workflow and the system architecture of the models which this dissertation proposes for multi-aspect user profiling in Twitter social media platform. The purpose of this research is to provide predictions about a person’s psycho-demographic aspects along with their explanations on how this classification occurred. These predictions require clarity, i.e., the reason why the models gave these results, as well as which features are more important and contribute more to the final results. Those results have to be in a format that is easily readable and perceivable by the users, so the results should be given in a user-friendly interface. The overall architecture consists of five machine learning models, namely a) age, b) gender, c) sentiment, d) emotion and e) personality along with three statistical models for topics, ethnicity and location. This multi-aspect approach combines the results of the models and present them to the user across with the most important features that affect the results. In addition, we provided the pre-model phases (i.e., data preprocessing and feature extraction), as well as the analysis of the building of the classification/regression models. Finally, we provided a comprehensive analysis on the explainability/interpretability definitions, and we discussed LIME explainability technique as the most suitable technique for explaining our ML models. In the next Chapter 4, the experiments that conducted for each model are reported along with the results and experiments on user profiling in social networks implementations.
Chapter 4

4. Results and Experiments on User Profiling in Social Networks Implementations

In this chapter, we present the stages of experimentation for each aspect and the models that were developed from the phase of the data collection of ground truth datasets to the evaluation of performance, describing also in detail the performance metrics that have been used to evaluate the process of the proposed methodologies and model architectures. This process is separated in the following steps:

1. The datasets that are used for training and evaluation of the performance and the modifications that applied to fit the problems.
2. The hyperparameter optimization for the model of each aspect.
3. The results of conducted experiments and the metrics that were used to highlight performance and the effectiveness of the models.
4. Discussion on the explainability findings.

As described in the previous chapter, the implementation of the platform, machine learning models, data manipulation, and the performance evaluation are performed using Python programming language. Moreover, the Microsoft Excel digital tool was also used in this chapter for the visualization results and the plots as Excel is a powerful, flexible tool for analytics and visualization activities, providing quality results in minimum time. Moving forward, the datasets and the changes that applied on each of them, to improve performance, are described. Toward building an effective and efficient web app, we also had to consider the performance of the platform. So, there is a significant tradeoff between the performance of the machine learning models (running on the background) and the platform itself (visible to the end user).

4.1 Datasets

The datasets contain instances along with the label, i.e., the aspect of profiling that we want the model to predict. They are used in order to develop and train our machine learning algorithms. In our work, we have used different datasets for the different profiling aspects that we want to predict. We developed and trained distinct ML models for predicting every profiling aspect and in the end, we gather all the results in order to provide the final outcome to the end user. We did this in order to provide predictions and explainability over the results for every different aspect, as well as because for every aspect, we needed a specific dataset to predict the particular aspect. Furthermore, all of the datasets consist of Twitter data because our predictions are about an individual’s profile based on data that retrieved from the Twitter. Using data that retrieved from a different source, e.g. headlines, could lead to inaccurate results.

In particular, each dataset contains prediction for only a single aspect, so a different dataset has been used for each of different personal aspects that we want to predict for a certain user. It is worthwhile to mention that the dataset’s size varies, as well as the content and the features that were available. Table 4-1, presents the datasets which were used for predicting every different psycho-demographic trait of a particular user, as well as some additional information about them.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Dataset/s</th>
<th>Size</th>
<th>Format</th>
<th>Label/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>(Pardo et. al., 2015)</td>
<td>294</td>
<td>Text</td>
<td>Age groups</td>
</tr>
<tr>
<td>Gender</td>
<td>(Kestemont et. al., 2018)</td>
<td>4500</td>
<td>Text</td>
<td>Male/Female</td>
</tr>
<tr>
<td>Emotion</td>
<td>(Mohammad et. al., 2018)</td>
<td>6838</td>
<td>Text</td>
<td>Anger, Disgust, Fear, Joy, Sadness, Surprise</td>
</tr>
<tr>
<td>Sentiment</td>
<td>(Nakov et. al., 2016)</td>
<td>30632</td>
<td>Text</td>
<td>Positive/Negative/Neutral</td>
</tr>
<tr>
<td>Personality</td>
<td>(Karanatsiou et. al., 2020)</td>
<td>242</td>
<td>Tabular/Text</td>
<td>Continuous values [0-5] for Big Five model</td>
</tr>
</tbody>
</table>

For age prediction was used one available dataset which was labeled with ``age groups``. In this case, model training for individuals' age group predictions was difficult to achieve a good performance and accuracy because the original data set was too small. The prediction of the age of the individuals is done by defining four age group categories. This dataset consists of age groups, as follows:

- 18-24 years old,
- 25-34 years old,
- 35-49 years old, and
- >=50 years old.

Nevertheless, the dataset that was available included a very small volume of data. In such case, the results would not be very accurate. By properly tuning a random forest classifier, we achieved about 80% accuracy on the test set. However, in order to ensure that this result was not affected by the data size (probable adding bias) itself, we used a semi-supervised
technique to label some more instances and further train our model. In particular, we used a public dataset which was used for another purpose to create a new labeled dataset about age for our developed ML model. We tested again our model using the same test set as previously in order to compare the model’s accuracy and identify whether has improved or not by training the model with more data. We used this from the initial training of the model with the original data and the same in each addition of new snapshots. The accuracy was the same in both cases. The process followed is described in the next section in more detail.

Furthermore, two different datasets have been used to predict a person's gender aspect. The first one contained only text (Tweets) and the label, while the second included text and behavioral features which comes from the twitter profile data. In both cases, the label is male or female. These two are used to train two different models for gender prediction. The first one’s predictions used to train the second models alongside with other features extracted from Twitter profile.

Datasets that were used for sentiment and emotion are from Sem-Val (SemEval, 2016) and consist of tweets directly collected from Twitter in the format of text. For sentiment analysis, each tweet is marked with a label. Sentiment consists of three labels negative, positive and neutral which one value negates the other. On the other hand, emotion analysis has several possible labels, but we used a subset of the total labels. These are anger, disgust, fear, joy, sadness and surprise. In the used dataset, there can be more than one label creating a multi-label problem and confusion, e.g., a tweet can be marked as anger and sad. So, in our developed model, for each aspect the tweet can either correspond to a class or not.

On top of that, a set of data containing the characteristics of the Big Five psychology model was used to predict the individual's personality. As described in Chapter 2, Big Five model consists of tweets and labels regarding Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism of a certain Twitter user. These labels include continuous values with an interval from 0 to 5. We have also modified a dataset to include tabular data due to incompatibility of the text with the LIME tool for explainability which we were aiming to use. So, the final data set includes data from the user profile, as well as some that have been extracted from the raw text of their profile.

4.2 Experimentation and Discussion on psychodemographic models

Now, after the detailed description of the datasets which are crucial to the training and performance of ML models, the experiments that conducted for each aspect are presented.

4.2.1 Experimental Environment

The computer system on which our developed platform is running and used has a major role in its performance evaluation due to the long times that needed for: a) the training of the models and b) the building/debugging of the platform as well. Due to these concerns, an 8-core and 16-threaded CPU at 3.8GHz clocked and 16 GBs of RAM are utilized. An M2 hard disk is used to load the already trained ML models for the building of the platform’s UI. In
comparison with an old 4-core CPU the training times reduced from ~6 hours to ~3 hours in some cases (hyper-parameterization) which is really crucial considering the amount of machine learning models that we train and use in our application.

4.2.2 Hyperparameter optimization

The hyperparameter optimization of the models is complex and time-consuming procedure as there are multiple parameters that need to be configured and optimized. We experiment with two techniques to automate this process, namely a) Grid Search and b) Random Search (Brownlee, 2021). The first one is an automated process for testing all the combinations of the model’s parameters. However, this was not effective because every run lasts a few days for each model. For this reason, we used the second one which makes a random selection of the parameter values. It’s a heuristic approach which provide a very good result close to optimal.

The hyperparameters of each model are presented in the following section. Different values applied for each value according the RandomSearchCV best results. Age, gender, emotion, sentiment and personality contain a different set of hyper-parameters due to the difference of their models. Location, Ethnicity and Interests traits are extracted by statistical models and thus no hyper-parameterization was needed. In the following section we present the results given by each developed model.

4.2.3 Textual and contextual models performance

As already mentioned, some models are based on text data for their predictions while others are based on context data (gender, OCEAN). Text data are extracted from user’s tweets that are collected from the Twitter account that we desire to analyze. After that, preprocessing and word tokenization are applied to these data to transform them in a suitable form for our models, as described in the previous chapter. Context models take as input data that can be extracted by a Twitter account in order to make predictions and are not based on text. It is worthwhile to mention that Ethnicity and Topics models are statistical models while the remaining are ML models which we briefly discuss this in the following.

In statistical models, the results are product of the percentage that is exported by the statistics. Specifically, for the ethnicity of the individuals, we process the text that has been collected from their Tweets and, based on the percentage of each identified language used by the user, we make an assumption about their nationality. As result, the percentage of every identified language is presented to the end user. Furthermore, to extract topics, we used user’s hashtags, most used common words, n-grams and LDA. These data provide a clear overview of the topics and interests of an individual and, on top of that, are easy to visualize.

On the other hand, the datasets that were described in the subsection 4.1 are used to train the machine learning models to efficiently and effectively predict user’s age, gender, sentiment, emotion and personality. The text, in this case, pass through the phases of preprocessing and

31https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925
feature extraction that accompany every model separately. Unigrams and bigrams words are the input for the models. In each case, the datasets split in two subsets: a) train and b) test with 70% and 30% ratios respectively, using a specific method which is provided by SK-learn library.

In order to calculate the performance of our models, we carefully selected a set of metrics based on the literature and the related work. Specifically, accuracy, recall, precision and F1 score are selected for model’s performance evaluation purposes. Relied on the best scores, the appropriate predictors were selected, but mainly we concentrated on accuracy and f1-score evaluation metrics. Also, these metrics are the User Profiling researcher’s most common choice. Performance metrics for these aspects are presented in the following figures. Furthermore, we have to mention that the tweets are fed individually to the model and an average of their results is then presented to the end user. In the following, we describe the results of the predictors for every model.

4.2.3.1 Gender
First, we have to mention again that for gender prediction two classifiers were trained with different types of data. Therefore, it is a hybrid approach, where the results from the text models are given as input to the context models to make the prediction. The first one - text classifier - was trained with text only, while the second one - context model - was trained with context data. In particular, the results for the text model were obtained as follows: for each tweet, a distinct result whether the user is man or woman was predicted and, in the end, all these results were added up to get the final predictions for a certain profile. In addition, the following results show the case where we have added up all the separate results for the tweets of each user and produce a result, the final forecast.

Each classifier was tested individually. Text model was trained from tweets contained in Pan-15 dataset. We applied preprocessing and cleanup to the data before the model training. In addition, the hyper-parameters of Random Forest Classifier for Gender Text prediction and Gender Context, as identified from the hyperparameter optimization, are presented in the following table:

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Gender Text Prediction</th>
<th>Gender Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_ESTIMATORS</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>MIN_SAMPLES_SPLIT</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>MIN_SAMPLES_LEAF</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MAX_FEATURES</td>
<td>&quot;auto&quot;</td>
<td>&quot;auto&quot;</td>
</tr>
<tr>
<td>MAX_DEPTH</td>
<td>80</td>
<td>50</td>
</tr>
<tr>
<td>BOOTSTRAP</td>
<td>False</td>
<td>True</td>
</tr>
</tbody>
</table>

On top of that, we trained five ML models, namely SVM, Naïve Bayes, Logistic Regression, Random Forest, and Voting Classifier with the four previously mentioned algorithms to select the one with the best performance for gender classification. Comparing the performance metrics (see Table 4-3), RF achieved slightly better results than logistic regression, while Naïve Bayes performed the worst. As it is shown in Table 4-3, Voting
Classifier using the four other algorithms, namely SVM, Naïve Bayes, Logistic Regression, Random Forest shown the best performance achieved – as expected – due to the fact that it trains on the ensemble of the four models and predicts an output (class) based on their highest probability of chosen class as the output.

Table 4-3: Performance results for different classification algorithms.

<table>
<thead>
<tr>
<th></th>
<th>RANDOM FOREST</th>
<th>LOGISTIC REGRESSION</th>
<th>SVM</th>
<th>NAÏVE BAYES</th>
<th>VOTING CLASSIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.784737</td>
<td>0.784734</td>
<td>0.779474</td>
<td>0.754737</td>
<td>0.797895</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.784876</td>
<td>0.784737</td>
<td>0.779474</td>
<td>0.754741</td>
<td>0.797928</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.784737</td>
<td>0.784733</td>
<td>0.779474</td>
<td>0.754737</td>
<td>0.797895</td>
</tr>
<tr>
<td>F1</td>
<td>0.784711</td>
<td>0.784737</td>
<td>0.779474</td>
<td>0.754736</td>
<td>0.797889</td>
</tr>
</tbody>
</table>

In Figure 4-1, we can observe that all of the performance metrics are close to 64%. On the other hand, we experimented also with datasets with context data. With the appropriate tuning, a better performance can be observed through in the Figure 4.1. We observe that the context model shows better performance. More specific, accuracy was about 72.43% while F1-score is 71.97%, which those metrics are our main criteria for model selection. Furthermore, recall and precision metrics are also close with 72.27% and 71.86%, respectively. Using both classifiers, we utilize all of the available data providing from Twitter. Table 4-4 presents the performance results for text model and context model separately.

![Gender Classifiers](image-url)
This approach significantly increases the execution time to our platform. Text models running slower than context, however utilizing all the available data, performance can be maximized. In production environment, data can be collected through user’s feedback improving both of our models’ performance evaluation.

### 4.2.3.2 Age

As referred on the first section of this chapter, the age dataset contains tweets of 294 twitter accounts. However, only the amount of 154 accounts was used for training our model, while the remaining was used as test set, following the competition rules and in order to be able to compare our performance. Tweets of each account concatenate to a single instance (Bayot et al., 2015). In our model training, we performed preprocessing, as described in Chapter 3, which improve the initial results. Briefly, we removed stopwords, punctuation marks and words that were less than three characters considering that they do not add any accuracy to our model, while, in the opposite, they decrease its performance. After that, we applied lemmatization in order to get the formal words.

On top of that, we trained four ML models, namely SVM, Naïve Bayes, Logistic Regression and Random Forest, to select the one with the best performance for age prediction. Comparing the performance metrics, RF achieved the highest results. We have to mention that each model passed through a hyper-parameterization phase to get the maximum performance. Having in mind that models were trained with a small dataset and thus, the performance in production environment could be very different, we considered to combine the results using a Voting Classifier with soft vote. The hyper-parameters of models used in Voting Classifier were the following:

- Random Forest => #estimators=200, min samples split=5, min samples leaf=2, max features="auto", max depth=90, bootstrap=False
- svm => kernel='poly', gamma=1, C=1, probability=True
- Multinomial NB => alpha=1e-05
- Logistic Regression => solver="newton-cg", penalty='none', C=100

The results are presented in Figure 4-2. We can observe that our model achieves accuracy 73.93% and f1 score 47.32%. The performance metrics varies to each run of the model as it was expected due to very few instances. Besides that, we can observe that using Voting Classifier a higher performance was achieved – as expected - except in recall metrics which

<table>
<thead>
<tr>
<th>TEXT MODEL</th>
<th>CONTEXT MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.6400</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.6476</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.6442</td>
</tr>
<tr>
<td>F1</td>
<td>0.6390</td>
</tr>
</tbody>
</table>
is approximately 3% lower than the RF classifier alone.

![Age Results]

**Figure 4.2: Age predictor results**

Having so little data to train our model, we experimented semi-supervised techniques as it was described in Section 3. To test the new instances labeled by the model, we used the same test set to compare the results. So, we used the new instances along with the already contained in the PAN-15 dataset and trained our model again. F1-score was slightly decreased about 4% and accuracy 6%. Also, we observed that setting the probability for a new instance to 75%, most of the new instances added to the dataset came only from 2 two classes (i.e., age groups), while the rest of the classes had very few instances. For instance, most of the predictions for class '18-24' and ‘50-XX’ were rejected because they had a probability lower than the set threshold.

**Table 4-5: Class imbalance techniques results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOTE</td>
<td>0.753521</td>
<td>0.625992</td>
<td>0.629249</td>
<td>0.623447</td>
</tr>
<tr>
<td>RANDOM OVERSAMPLE</td>
<td>0.802817</td>
<td>0.730303</td>
<td>0.694858</td>
<td>0.705304</td>
</tr>
<tr>
<td>BORDERLINE SMOTE</td>
<td>0.774648</td>
<td>0.716065</td>
<td>0.650677</td>
<td>0.65667</td>
</tr>
<tr>
<td>NEARMISS 1</td>
<td>0.5</td>
<td>0.493128</td>
<td>0.557482</td>
<td>0.482</td>
</tr>
<tr>
<td>NEARMISS 2</td>
<td>0.556338</td>
<td>0.514048</td>
<td>0.593504</td>
<td>0.508973</td>
</tr>
<tr>
<td>NEARMISS 3</td>
<td>0.443662</td>
<td>0.408572</td>
<td>0.447999</td>
<td>0.350132</td>
</tr>
<tr>
<td>SMOTE + NEARMISS 2</td>
<td>0.683099</td>
<td>0.592664</td>
<td>0.585683</td>
<td>0.573483</td>
</tr>
<tr>
<td>SMOTE + NEARMISS 3</td>
<td>0.612676</td>
<td>0.588286</td>
<td>0.564778</td>
<td>0.495354</td>
</tr>
</tbody>
</table>

This resulted from the fact that there is a class imbalance in the data. In these two minor classes we used imbalance techniques, as can be shown in Table 4-5, but the probabilities
given by the model were low to get into the new dataset. We observed that performance metrics slightly improved, while the best metrics were output of the Random Oversampler class imbalance technique. However, the probabilities extracted by the Voting Classifier were lower and thus, the instances were rejected.

In the following Table 4-6 we can observe how the data preprocessing, and especially the data cleaning, improves the accuracy of machine learning models. However, it can be seen that not all performance metrics were improved, on the contrary recall was significantly decreased. This is a proof that although data preprocessing is a crucial stage in machine learning, we should experiment with data pre-processing steps that are appropriate for our specific data to see if and when we can get that desirable boost in model accuracy, as well as in the rest of the performance metrics.

<table>
<thead>
<tr>
<th>WITH DATA CLEANING/PREPROCESSING</th>
<th>AGE WITHOUT CLEANING</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACCURACY</strong></td>
<td>0.7254</td>
</tr>
<tr>
<td><strong>RECALL</strong></td>
<td>0.6166</td>
</tr>
<tr>
<td><strong>PRECISION</strong></td>
<td><strong>0.4762</strong></td>
</tr>
<tr>
<td><strong>F1</strong></td>
<td>0.4632</td>
</tr>
</tbody>
</table>

### 4.2.3.3 Sentiment

As mentioned before, in order to predict sentiment, we used a text model which predicts user’s sentiment for each tweet of their account. For the sentiment prediction, we used Voting Classifier with the Random Forest, SVM, Naïve Bayes and Logistic Regression classifiers, for which the hyper-parameters, identified from the hyperparameter optimization, are presented in Table 4-7:

<table>
<thead>
<tr>
<th>Random Forest</th>
<th>SVM</th>
<th>Naïve Bayes</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_estimators=1600</td>
<td>kernel='linear'</td>
<td>fit_prior=True</td>
<td>solver='newton-cg'</td>
</tr>
<tr>
<td>min_samples_split=10</td>
<td>gamma=1</td>
<td>alpha=1</td>
<td>penalty='l2'</td>
</tr>
<tr>
<td>min_samples_leaf=2</td>
<td>C=1</td>
<td></td>
<td>C=10</td>
</tr>
<tr>
<td>max_features=&quot;auto&quot;</td>
<td>probability=True</td>
<td></td>
<td></td>
</tr>
<tr>
<td>max_depth=None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bootstrap=True</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the Random Forest classifier are presented in Figure 4.3 that follows. We can
observe that the accuracy and recall are 62% and 63%, respectively. At this point, we have to mention that these results were extracted from the text without preprocessing. Experiments conducted with preprocessing were performed directly in text. Specifically, we removed words that are not presented in a corpus, hashtags, URLs and any word with less than 3 characters. The performance of the predictors was diminished as it is shown in the Figure 4.3. It can be seen that the recall was slightly improved, but all the other performance evaluation metrics were decreased up to 13%.

Figure 4.3: Sentiment analysis results

Considering that further improving can be achieved using pre-trained models, we experimented also with Vader Sentiment (Hutto & Gilbert, 2014), as it can be seen in Table 4-8, where the performance metrics are presented for sentiment data without preprocessing, with preprocessing and using pretrained models. In particular, we experimented with polarity score values to get these results. Especially, we set negative as < -0.45, positive > 0.45 and neutral between the -0.45 and 0.45. We observed that, in most metrics, the results presented lower performance. Especially, recall was 51.29% which is almost 14% lower than our text model and accuracy about 5% lower. Precision was slightly better in comparison to our model, about 4% improvement. Finally, we selected our text model with text cleaning based mainly on accuracy and F1 metric.

Table 4-8: Performance metrics with data preprocessing, with data preprocessing and using pretrained models.

<table>
<thead>
<tr>
<th></th>
<th>SENTIMENT WITHOUT CLEANUP</th>
<th>SENTIMENT WITH CLEAN UP</th>
<th>VADER SENTIMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.6066</td>
<td>0.6258</td>
<td>0.5699</td>
</tr>
<tr>
<td>RECALL</td>
<td><strong>0.6500</strong></td>
<td>0.6243</td>
<td>0.5129</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.4878</td>
<td>0.5299</td>
<td><strong>0.5783</strong></td>
</tr>
<tr>
<td>F1</td>
<td>0.4712</td>
<td><strong>0.5369</strong></td>
<td>0.5237</td>
</tr>
</tbody>
</table>
On top of that, similarly to the age prediction, again for sentiment prediction, we trained four ML models, namely SVM, Naïve Bayes, Logistic Regression and Random Forest, to select the one with the best performance for sentiment prediction. Comparing the performance metrics (see Table 4-9), RF achieved the highest results. Having in mind that models were trained with a small dataset and thus, the performance in production environment could be very different, we considered to combine the results using a voting schema with soft vote. As it is shown in Table 4-9, a higher performance achieved – as expected - except in recall metrics which is approximately 3% lower than the RF classifier alone.

Table 4-9: Performance metrics for SVM, Naive Bayes, Logistic Regression, Random Forest and Voting Classifier ML algorithms.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>NAÏVE BAYES</th>
<th>LOGISTIC REGRESSION</th>
<th>RANDOM FOREST</th>
<th>VOTING CLASSIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.6040</td>
<td>0.6043</td>
<td>0.5972</td>
<td>0.6295</td>
<td>0.6324</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.5732</td>
<td>0.5798</td>
<td>0.5697</td>
<td>0.6491</td>
<td>0.6185</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.5710</td>
<td><strong>0.5965</strong></td>
<td>0.5662</td>
<td>0.5432</td>
<td>0.5890</td>
</tr>
<tr>
<td>F1</td>
<td>0.5718</td>
<td>0.5861</td>
<td>0.5677</td>
<td>0.5566</td>
<td><strong>0.5973</strong></td>
</tr>
</tbody>
</table>

4.2.3.4 Emotion

In Figure 4.3, the results of the emotion predictors are presented, including the six basic emotions that we aim to analyze: anger, disgust, fear, joy, sadness and surprise. In our predictors, we used Random Forest with the appropriate hyper-parametrization. In particular, the hyper-parameters of Random Forest Classifier, identified from the hyperparameter optimization, were the following:

✓ n_estimators=400,
✓ min_samples_split=2,
✓ min_samples_leaf=1,
✓ max_features="sqrt",
✓ max_depth=20,
✓ bootstrap=False.

In particular, Figure 4.4 shows that all the models perform with a high percentage of the selected metrics. Predictors for fear and surprise return the highest accuracy of 89.52% and 92.25% respectively. The lowest value observed in disgust, which is 73.05%. The remaining varies from 77% to 80%, which is quite high considering that our predictors are simplified to be the most explainable and efficient in time manner. Despite that the highest accuracy provided by Surprise predictors, we can observe that it has the lowest F1-score of 58.41%. F1-score for the other models is in range of 66.42% of sadness and 80.12% that model of fear achieves.

Furthermore, experimentations with Classifier Chain (Read et. al., 2009) and One VS Rest took place. In both cases, predictions which made for a class affected the predictions for the other classes. Especially, Classifier Chain predicting a class value, uses this result as input to the next predictor. On the other hand, in One VS Rest ‘the class is fitted against all the other classes’ which is similar to our approach, but it extracts all the classes of an instance. Considering that it will be really complex to provide explainability using these techniques,
we finally use our initial approach.

![Figure 4.4: Emotion's predictors results for RF classifier.](image)

In the following performance metrics are presented for Random Forest, SVM, Naïve Bayes and Logistic Regression classifiers. We can observe that LR classifier performs slightly better in “anger”, “disgust”, “joy” and “surprise” than the RF classifier, while the RF classifier performs better in “fear” and “sadness” than the LR classifier. SVM and Naïve Bayes perform the worst accuracy. According to Google, “Logistic regression performs better when the number of noise variables is less than or equal to the number of explanatory variables and the random forest has a higher true and false positive rate as the number of explanatory variables increases in a dataset.”

In general, we chose the RF classifier for emotion prediction in our model, as it is simpler, fast and it presents better accuracy with more categorical data than numeric, while logistic regression is a little confusing when comes to categorical data.

<table>
<thead>
<tr>
<th>ANGER</th>
<th>DISGUST</th>
<th>FEAR</th>
<th>JOY</th>
<th>SADNESS</th>
<th>SURPRISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.781189</td>
<td>0.730507</td>
<td><strong>0.895224</strong></td>
<td>0.800195</td>
<td><strong>0.771930</strong></td>
</tr>
<tr>
<td>RECALL</td>
<td>0.720070</td>
<td>0.680473</td>
<td><strong>0.769624</strong></td>
<td>0.753323</td>
<td>0.648892</td>
</tr>
<tr>
<td>PRECISION</td>
<td><strong>0.789564</strong></td>
<td>0.727987</td>
<td>0.851323</td>
<td>0.811111</td>
<td>0.760864</td>
</tr>
<tr>
<td>F1</td>
<td>0.734435</td>
<td>0.687269</td>
<td><strong>0.801251</strong></td>
<td>0.767252</td>
<td>0.664272</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANGER</th>
<th>DISGUST</th>
<th>FEAR</th>
<th>JOY</th>
<th>SADNESS</th>
<th>SURPRISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.778265</td>
<td>0.729532</td>
<td>0.878168</td>
<td>0.783626</td>
<td>0.761209</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.715654</td>
<td>0.664066</td>
<td>0.688047</td>
<td>0.720239</td>
<td>0.614051</td>
</tr>
<tr>
<td>PRECISION</td>
<td><strong>0.787319</strong></td>
<td><strong>0.753632</strong></td>
<td><strong>0.871303</strong></td>
<td><strong>0.820879</strong></td>
<td><strong>0.781433</strong></td>
</tr>
<tr>
<td>F1</td>
<td>0.729808</td>
<td>0.667383</td>
<td>0.733446</td>
<td>0.734318</td>
<td>0.618875</td>
</tr>
</tbody>
</table>
Table 4.2: Performance results for Naive Bayes classifier.

<table>
<thead>
<tr>
<th></th>
<th>ANGER</th>
<th>DISGUST</th>
<th>FEAR</th>
<th>JOY</th>
<th>SADNESS</th>
<th>SURPRISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.787524</td>
<td>0.748538</td>
<td>0.848441</td>
<td>0.787037</td>
<td>0.760721</td>
<td>0.944932</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.748604</td>
<td>0.708654</td>
<td>0.634313</td>
<td>0.735513</td>
<td>0.633768</td>
<td>0.511116</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.774852</td>
<td>0.741973</td>
<td>0.775709</td>
<td>0.799726</td>
<td>0.739363</td>
<td>0.563911</td>
</tr>
<tr>
<td>F1</td>
<td>0.757693</td>
<td>0.716481</td>
<td>0.664710</td>
<td>0.748979</td>
<td>0.645752</td>
<td>0.511032</td>
</tr>
</tbody>
</table>

Table 4.3: Performance results for Logistic Regression classifier.

<table>
<thead>
<tr>
<th></th>
<th>ANGER</th>
<th>DISGUST</th>
<th>FEAR</th>
<th>JOY</th>
<th>SADNESS</th>
<th>SURPRISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>0.799220</td>
<td>0.753411</td>
<td>0.884990</td>
<td>0.814327</td>
<td>0.767057</td>
<td>0.953704</td>
</tr>
<tr>
<td>RECALL</td>
<td>0.760749</td>
<td>0.716056</td>
<td>0.720469</td>
<td>0.775731</td>
<td>0.654508</td>
<td>0.576124</td>
</tr>
<tr>
<td>PRECISION</td>
<td>0.789408</td>
<td>0.746054</td>
<td>0.859232</td>
<td>0.818449</td>
<td>0.737903</td>
<td>0.798786</td>
</tr>
<tr>
<td>F1</td>
<td>0.770672</td>
<td>0.723789</td>
<td>0.763067</td>
<td>0.788342</td>
<td>0.669713</td>
<td>0.614041</td>
</tr>
</tbody>
</table>

Having described the text models for age, emotion and sentiment, in the next section context models performance will be presented, which include gender and personality traits. These models are based in features that are being extracted basically from the Twitter account metadata. The datasets which were used for these models contain such information along with their text (tweets).

4.2.3.5 Personality

Last but not least, personality model predictor was included in our platform to perform analysis over Big Five personality traits. This model is based on context data which means that takes as input data that can be extracted by a Twitter account (metadata). The hyperparameters for the models (i.e., Agreeableness, Conscientiousness, Extraversion, Neuroticism, Openness), identified from the hyperparameter optimization, are presented in the following:

Table 4.4 Personality (OCEAN) models hyper-parameterization.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Big Five</th>
</tr>
</thead>
<tbody>
<tr>
<td># estimators</td>
<td>400</td>
</tr>
<tr>
<td>Samples split</td>
<td>3</td>
</tr>
<tr>
<td>Min samples leaf</td>
<td>2</td>
</tr>
<tr>
<td>Max features</td>
<td>sqrt</td>
</tr>
<tr>
<td>Max depth</td>
<td>None</td>
</tr>
<tr>
<td>bootstrap</td>
<td>false</td>
</tr>
</tbody>
</table>

For the individual's personality prediction, a set of data containing the characteristics of the Big Five psychology model was used. As described in Chapter 3, Big Five model consists of tweets and features regarding Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism of a certain Twitter user. The features that were used to train the model are
presented in Chapter 3. In the following, the Mean Absolute Error (MAE) is given for every model separately. The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. In simple words, the MAE is used to measure the closeness of the prediction to the eventual outcomes, and thus model’s accuracy. We can observe the MAE is quite high and we should optimize our models to get lower values.

<table>
<thead>
<tr>
<th></th>
<th>Openness</th>
<th>Conscientious</th>
<th>Extraversion</th>
<th>Agreeableness</th>
<th>Neuroticism</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAE</strong></td>
<td>0.585375</td>
<td>0.701573</td>
<td>0.920786</td>
<td>0.628115</td>
<td>0.800690</td>
</tr>
</tbody>
</table>

Comparing the results with those presented by Karanatsiou et al. (2020) we considered to include their text model in our platform in order to get better results. Specifically, their results are 0.2 RMSE for the independent model and 0.19 for the holistic model (in label scale 0-1) which are quite better. Furthermore, a comparison of the results is given in the following section (see Table 4-20). We considered necessary to include their model because the results are higher than our simple model and text utilization is our main focus for this platform. However, to keep our platform consistent and provide explanations about the aspects we kept our own model complementary but the final results are given by the text model and as explained in our platform the explanations consist only possible reasons that could reveal the personality aspect.

To conclude, in this section we described some of the experiments that conducted to enhance the best suitable model to our platform. Also, we would like to mention that our experiments were focused on dataset modifications such as predicting the age of specific tweets and merging the results of a single account in contradiction with the majority of recent works that try to predict the age of the account at once. Furthermore, we conducted and presented several experiments with sets of features in order to select the most efficient model. We also trained several models, like neural networks. But we included only a part of them due to space limitations. Finally, even a promising model with high percentage scores may not work efficiently when the new data for prediction are unknown. So, our platform is designed to be easily adjustable to incorporate modifications over the models in the future. In the next section 4.2.4, we present the trade-off that we made for our models.

### 4.2.4 Performance and Explainability Tradeoff and SOTA Comparison

Given the extensiveness with which AI systems are being developed and deployed to upend from personalizing the user experience, to recommending products and content for customers, to automating two-way conversations with customers to nurture relationships, it becomes critical for researchers to understand how to achieve explainability for different types of AI models, assess the tradeoff between prediction accuracy and explanation associated with different choices, and develop and deploy trustworthy AI systems that meet business and fairness objectives.

Tradeoff aims to achieve a balance between two desirable but incompatible features, as depicted in Figure 4-5; in our case ML model’s accuracy and its results explainability. Many
practical applications of machine learning systems call for the ability to explain why certain predictions are made. But such explanations are not always compatible with our machine learning model.

When choosing a machine learning model, we usually think in terms of two choices (Loyola-Gonzalez, 2019):

- **accurate but black box**: The best classification accuracy is typically achieved by black box models such as Gaussian processes, neural networks or random forests, or complicated ensembles of all of these. Black boxes are often criticized because they are very complex and thus, their inner workings are really hard to understand. They do not, in general, provide a clear explanation of the reasons they made a certain prediction, they just spit out a probability.

- **white box but weak**: On the other end of the spectrum, models whose predictions are easy to understand and communicate are usually very impoverished in their predictive capacity (linear regression, a single decision tree) or are inflexible and computationally cumbersome (explicit graphical models).

So which ones should we use: accurate black-box models, or less accurate but easy-to-explain white-box models? Explainability is not a property of the model, but based on our application, we should consider which explainability models are suitable and applicable.

![Figure 4.5: Relationship between Explainability and Accuracy](image)

At this point, that the results for every single aspect have been presented and analyzed, we will discuss the tradeoff between explainability and performance for our developed platform. Although many architectures have been proposed for predicting personal traits effectively, as shown in the literature review, there is no doubt that complex approaches are difficult to be explainable to the end user. More specifically, using complex architectures, models may be explainable, but they might not be easily understandable by users. Also, although they might improve model’s accuracy, they may be difficult to be managed by our platform. For example, including POS tags in our text models could further improve the performance but the explainability to the end user is not obvious. Furthermore, adding different types of data and integrating different personal traits in one single model is a strong point of this work. However, this can potentially add complexity in our system in terms of explainability. For this reason, we developed different models for predicting each personal trait using different datasets and types of data, and then we integrated all results.

On top of that, a fundamental difference between this work and prior research is that we
took into account explainability as an initial requirement towards the development of our platform. That is, along with the user profiling outcome, we present an approach to provide a textual description that summarizes the system’s understanding of the model’s results about user’s personal traits. In this way, it makes it easier for the model to be validated and adjusted, building user trust and understanding.

In the following, we can see tables for each aspect presenting the accuracy that we achieved with the black box which has been presented in Section 4.2.3, and those that we achieved with the white box models providing the result explanations in detail.

For the Age prediction, we can observe that the results are differ in a high percentage. Specifically, accuracy is 16% lower and f1-score about 5% which are our main selection criteria. Furthermore, we provide results of a more complex model (CM1) that we experiment but it was not able to use LIME. Specifically, for CM1 we trained four classifiers (RF, SVM, Naïve Bayes, and LR) with proper tuning and we used that in a voting schema. The training of these models occurred combining text and context features. We observe that the results are quite higher than our RF approach but, in this way, it is impossible to provide explanations for each prediction. Specifically, accuracy achieved is about 77.46% and f1 score is 65.66%. Furthermore, precision and recall are also higher, 71.6% and 65.06 respectively. Despite the fact that better results are achieved with CM1, LIME utilization was not able because no explanations could be given.

<table>
<thead>
<tr>
<th>Age</th>
<th>Decision Tree - White box</th>
<th>Random Forest - Black box</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy 0.5634</td>
<td>0.7253</td>
</tr>
<tr>
<td></td>
<td>Recall 0.4209</td>
<td>0.6165</td>
</tr>
<tr>
<td></td>
<td>Precision 0.4150</td>
<td>0.4762</td>
</tr>
<tr>
<td></td>
<td>F1 0.4105</td>
<td>0.4632</td>
</tr>
</tbody>
</table>

For the Gender prediction, we can observe that RF and VC in black box achieves much better results than the white box as it was expected. Specifically, we observe that the decision tree has ~15% lower efficiency in all metrics (accuracy, recall, precision, and f1 score).

<table>
<thead>
<tr>
<th>Gender</th>
<th>Decision Tree - White box</th>
<th>RF - Black box</th>
<th>Voting Classifier - Black box</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy 0.6358</td>
<td>0.7847</td>
<td>0.7978</td>
</tr>
<tr>
<td></td>
<td>Recall 0.6358</td>
<td>0.7848</td>
<td>0.7979</td>
</tr>
<tr>
<td></td>
<td>Precision 0.6358</td>
<td>0.7847</td>
<td>0.7978</td>
</tr>
<tr>
<td></td>
<td>F1 0.6358</td>
<td>0.7847</td>
<td>0.7978</td>
</tr>
</tbody>
</table>

We should highlight that although the voting classifier achieves the best results (~80%),
when we pass it in LIME, the feature importance of the features is zero and therefore we considered to not use it in our application because that could cost the loss of the explainability functionality, to gain less than 2% of accuracy. The main emphasis in this section should be put to the fact that using models with complex features is difficult to be interpreted by the end user and thus, in many cases no results are returned by LIME because the importance of the features is too small.

Tables 4-17, 4-18 and 4-19 show the results from black box model (Random Forest) and white box models (decision tree and logistic regression). In some cases, we observe that the LR performs better than the RF, such as anger, disgust and surprise while in other aspects the black box models performed better. In each case of emotion, the results differ by ~ 1-3% between LR and RF. Decision tree classifier is performing worse than the other two and, in some cases, (surprise) accuracy can be up to 5% lower.

<table>
<thead>
<tr>
<th>Table 4-17: Random Forest performance metrics for emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Forest – Black box</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>anger</td>
</tr>
<tr>
<td>0.7811</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>0.7200</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
</tr>
<tr>
<td>0.7895</td>
</tr>
<tr>
<td><strong>F1</strong></td>
</tr>
<tr>
<td>0.7344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4-18: Logistic regression performance metrics for emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logistic regression – White box</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>anger</td>
</tr>
<tr>
<td>0.7992</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>0.7607</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
</tr>
<tr>
<td>0.7894</td>
</tr>
<tr>
<td><strong>F1</strong></td>
</tr>
<tr>
<td>0.7706</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4-19: Decision tree performance metrics for emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decision tree – White box</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>anger</td>
</tr>
<tr>
<td>0.7578</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>0.7283</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
</tr>
<tr>
<td>0.7360</td>
</tr>
<tr>
<td><strong>F1</strong></td>
</tr>
<tr>
<td>0.7316</td>
</tr>
</tbody>
</table>

At this point, we have to mention that for sentiment analysis we used a black box model with complex architecture because for explainability results, we present a number of tweets which represents the corresponding category (negative, positive, and neutral). So, a complex black box model architecture was sufficient for our purpose. Note that in the other aspects of profiling it was not possible to present them like this because the tweets were all merged and
given to the model as a single text. Although, we present a white box (LR) and two black box models and it is obvious that VC model overperforms both of the other two cases.

Table 4-20: Logistic regression, Random Forest and a voting schema for sentiment prediction

<table>
<thead>
<tr>
<th></th>
<th>LR – White Box</th>
<th>RF – Black box</th>
<th>VC – Black box</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.5972</td>
<td>0.6295</td>
<td>0.6361</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.5697</td>
<td>0.6491</td>
<td>0.6185</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.5662</td>
<td>0.5432</td>
<td>0.5945</td>
</tr>
<tr>
<td><strong>F1</strong></td>
<td>0.5677</td>
<td>0.5566</td>
<td>0.6039</td>
</tr>
</tbody>
</table>

In Table 4-21 we present our experiments with our context model for Big Five prediction which described in Chapter 3 and those that included by Karanatsiou et al (2020). Our approach consists of a simple architecture including only a number of context features. As explained previously, due to incompatibility of LIME package with text regression predictions we limited only to use context features. However, we observe that models which proposed by Karanatsiou et al, (2020) for Big Five prediction are performing much better. That is something that expected, as we didn’t use text for our predictions which is important for personality prediction. So, we considered to include both approaches because the pretrained were not possible to be explainable and our under-performed model could be explainable.

Table 4-21: OCEAN model comparison (RMSE) of explainable and non-explainable models

<table>
<thead>
<tr>
<th>Personality - OCEAN</th>
<th>Openness</th>
<th>Conscientious</th>
<th>Extraversion</th>
<th>Agreeableness</th>
<th>Neuroticism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context model RMSE</td>
<td>0.185</td>
<td>0.186</td>
<td>0.216</td>
<td>0.174</td>
<td>0.236</td>
</tr>
<tr>
<td>Pretrained model RMSE (Karanatsiou et al, 2020)</td>
<td>0.158</td>
<td>0.176</td>
<td>0.221</td>
<td>0.221</td>
<td>0.228</td>
</tr>
</tbody>
</table>
In summary, in Figures 4.6-4.7 we observe that black boxes, which by their nature are difficult to interpret and require special techniques such as LIME, in most cases perform better. In addition, there are cases where using this technique it was not possible to give results which is due to the fact that the large number of features combined with a small dataset makes it difficult to explain. Also, complex models are difficult to interpret and render the results to the user in an understandable form. In most cases of the models we have developed, a trade-off has been made for a variety of reasons such as to achieve better explainability, higher efficiency, to optimize the performance of the application as well as to align with the time frames that we had to complete this dissertation. In the following we present some innovative approaches which are the SOTA models for the profiling aspects that included in Mediator.
Comparison with SOTA models for each aspect:

Abdallah et.al. (2020) based on the fact that the massive use of the social media and the huge number of messages that are shared on the internet, create a fertile ground for automatically detecting the age and gender of the people who write these messages, they presented an automate tool with a unique set of features that used to analyze a given text. The features include the unigram, part of speech, and production rules. The accuracy results of the proposed method outperform the existing techniques. The best results achieved by using the production rules features. Their main goal was to predict the age and gender of the writers of deceptive text. A complete model was trained and tested with several classifiers on a data set to determine the gender and age the writers of deceptive text. In order to achieve the highest accuracy possible, gain-based feature selection was implemented to removes irrelevant and redundant features that do not have high impact. The experimental results outperform the existing techniques that deal with open domain text. The best results are achieved for gender prediction using features (CFG) with accuracy of 82.81% via the SVM classifier. The best prediction for age was achieved using CFG with accuracy of 83.2% via SVM classifier.

In a recent study, the sentiments of the tweets or reviews published in the twitter were identified by searching for the particular keyword in tweets and then evaluate the polarity of the tweets as positive and negative (Shamantha et. al., 2020). The sentiments of the tweets that are tweeted on a twitter evaluated based on feature selection of each score words. In order to select the best features Naive Bayes Classifier (NBC) is used for training and testing the features of a words and also evaluating the sentiment polarity of each tweet. Performance evaluation parameters such as accuracy, precision and time is taken into consideration and compared with three machine learning classifiers, namely, Random Forest, Naive Bayes and Support Vector Machine (SVM). Based on SOTA models, the authors achieved accuracy slightly higher than 80%.

The main aim of Ab. Nasir et al, (2020) was to develop text-based emotion recognition and prediction system. Therefore, four supervised machine learning classification algorithms such as Multinomial Naïve Bayes, Support Vector Machine, Decision Trees, and k-Nearest Neighbors were investigated. The model was developed based on Ekman's six basic emotions which are anger, fear, disgust, joy, guilt and sadness. Data pre-processing techniques such as stemming, stop-words, digits and punctuation marks removal, spelling correction, and tokenization were implemented. A benchmark of ISEAR (International Survey on Emotion Antecedents and Reactions) dataset was used to test all models. Multinomial Naïve Bayes classifier resulted the best performance with an average accuracy of 64.08%.

Furthermore, Halim et. al., presented a supervised machine learning-based framework for emotion recognition from email text. The proposed framework was incorporated with three feature selection methods, three classification methods, two clustering techniques, and multiple cluster validity indices. Experiments were performed by extracting 14 features from the datasets. Three feature selection techniques, i.e., Principal Component Analysis (PCA), Mutual Information (MI), and Information Gain (IG) were used to identify the optimum features for classification. Afterwards, three classifiers, namely, Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machines (SVM), were employed to predict the emotions of the emails. Classifiers’ performance was evaluated using F1 score, precision, accuracy, and recall. Based on the feature selection results, experiments were
conducted on all three datasets by vertically partitioning them into all features, top features, and bottom features. During experiments, 83% accuracy was achieved for the proposed work to predict an emotion. The obtained results suggested better performance of the current work.

Last but not least, personality is an important parameter as it differentiates various individuals from one another. **Personality** prediction is an evergreen area of research. Kunte & Panicker, (2019), applied Multinomial Naïve Bayes, AdaBoost and LDA to compare which algorithm has higher relevance for personality prediction. Thus, according to their results it is found that Multinomial Naïve Bayes has highest accuracy of 73.43, precision of 0.7, and recall of 0.71 and F1-score of 0.72.

There is no doubt that already designed and developed SOTA models are able to achieve high accuracy performance results. Nevertheless, one of our main targets was to develop and train the models so that to be able to provide explanations that are understandable by the end users, and thus, the SOTA models were inappropriate in this case and therefore, we had to design and develop appropriate models for that. In our case, we aim to achieve an acceptable tradeoff which means to achieve a balance between ML model’s accuracy and its results explainability that consist two desirable but incompatible features, as described previously in this Section. Many practical applications of machine learning systems call for the ability to explain why certain predictions are made. But such explanations are not always compatible with our machine learning model. In the following we comment on some explanations that have been extracted from our models. Also, lists of most representative words for each prediction are given to evaluate that the predictions are based in features that corresponds to each predicted class.

### 4.3 Explainability - Validating the Results

In this section, the explainability results using LIME technique are discussed. More specifically, we are going to discuss direct results that were extracted from a Twitter account through our platform. Which were the most important features and what was their impact to the predictions? We provide some indicative results of each type of data - context and text. LIME explainer can be used only for local explanations which means that the results are just about a single instance. We will not present every plot that can be exported due to the large number of the models that were included to our platform. Thus, key plots of the classifiers are presented along with tables which present influential features about the ML models. A random Twitter account was selected and the LIME results are presented in the following tables.

<table>
<thead>
<tr>
<th>Sadness – Important words</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>'heartbreaking'</td>
<td>'policing'</td>
</tr>
<tr>
<td>'city'</td>
<td>'tonight'</td>
</tr>
<tr>
<td>'rugby'</td>
<td>'time'</td>
</tr>
<tr>
<td>'drug'</td>
<td>'important'</td>
</tr>
<tr>
<td>'service'</td>
<td>'watch'</td>
</tr>
<tr>
<td>'body'</td>
<td>'anger'</td>
</tr>
<tr>
<td>'really'</td>
<td>'documentary'</td>
</tr>
<tr>
<td>'working'</td>
<td>'monday'</td>
</tr>
</tbody>
</table>
In Table 4-22, prediction results are presented for a specific (random-selected) Twitter account using LIME. The most important words are presented along with the value-importance given by LIME. Most of the words are obviously related to a ‘sadness’ emotion. Especially, “heartbreaking”, “city”, “drug”, “policing” and “anger”, among others. Despite that, other words may be more difficult to explain, like “rugby” or “body” because they are not directly related to sadness or any other emotion. For this reason, in our platform, we represent these feature importance results using word clouds providing an overall explainability. It is obvious that explanations for a single prediction cannot provide an overall view of the model, but only explanations about certain prediction model. We experiment with a few Twitter accounts to verify that the models were trained appropriately and the given results were not ‘random’. As it is expected, in some cases, the models predict wrong emotion class, but the given explanations still make sense because the more impactful words are still the same.

The following table (Table 4-23) present some of the features that have been exported as part of a prediction explanations for each model of emotion. We have chosen the 5 most relevant words for each model as it is obvious that there are words that cannot be easily correlated with the predictions. The number of words that returned by LIME depends by the number of tweets that an account has been posted and even if the number of posts is high that could be decreased by the preprocessing that we apply. For example, tweets that include only a link (URL) or just two words will return zero or two words respectively.

**Table 4-23: Emotion – Top-5 words per prediction**

<table>
<thead>
<tr>
<th>Emotion prediction – Important words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anger</strong></td>
</tr>
<tr>
<td>Rape</td>
</tr>
<tr>
<td>Brutal</td>
</tr>
<tr>
<td>Poor</td>
</tr>
<tr>
<td>Dislike</td>
</tr>
<tr>
<td>Unsuccessfull</td>
</tr>
</tbody>
</table>

Furthermore, context model explainability results for gender are presented in Table 4-24. Specifically, most important features per class were extracted for the second model of gender. Additional, for the textual model the most representative words for a specific prediction (account predicted as ‘male’) are ‘migrant’, ‘doctor’, ‘husband’, ‘patriots’ and ‘crowd’. Although the prediction is correct the explanations may seem not sufficient. This occurs because in the content of tweets words contained that are either not known to the model or a part of them is deleted in preprocessing.

**Table 4-24: Gender Explainability – Male prediction**

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average likes</td>
<td>Avg sentiment pos</td>
</tr>
<tr>
<td>Number of urls</td>
<td>Text model</td>
</tr>
<tr>
<td>flooding</td>
<td>number of stopwords</td>
</tr>
<tr>
<td>Words average length</td>
<td>post length</td>
</tr>
<tr>
<td>Number of emojis</td>
<td>Number of hashtags</td>
</tr>
</tbody>
</table>
The main explainability tool that is used in Mediator is LIME which provides the explanations to the end users using the already trained models. We used also Shapley values (Lundberg et al., 2020) technique for each model to validate that the model prediction results do not include bias, discrimination or overfit. At this point, it is worthwhile to mention that explainability techniques are also used to debug, adjust, and improve our developed models. For instance, extracting high results could be driven by wrong features, or we may detect that a model was trained with character-based features which was not desired, and thus take proper actions. In the following, Figure 4.8, we present features importance selectively showing the results for the personality aspect. Features with negatives importance affect the prediction close to zero while positives to maximum value. We have to remind the reader that the predicted values are zero to five. Due to space limitations, we present the results which extracted by Shap only for Openness model as indicative example in Figure 4.8.

To sum up, in this chapter, we provided a comprehensive report of all the stages that we passed through during the experimentation for every aspect, as well as the models that we developed. In particular, we presented the stages from the phase of the data collection of ground truth datasets to the evaluation of performance, describing also in detail the performance metrics that have been used to evaluate the process of the proposed methodologies and model architectures. This process is separated in the following steps: 1) Gather the datasets that were used for the training of our models and their performance evaluation, as well as the modifications that were applied to them in order to fit the problems and our use case; 2) Perform hyperparameter optimization for the models of every aspect to identify the multiple parameters that needed to be configured and optimized; 3) Extract and present the results of the conducted experiments and the metrics that were used to highlight performance and the effectiveness of the developed models; and finally 4) discuss the explainability findings and the validation of our results.
In conclusion, it is worthwhile to highlight that this dissertation’s main contribution is twofold: a) the development of a multi-aspect user profiling platform that can be used to analyze personality traits about Twitter users, and b) the provision of explainability over the given results to prevent bias, and to be easier for the model to be validated and adjusted, building user trust and understanding. Therefore, as described previously in this section, the models are not SOTA and were trained in order to be able to provide explanations that are understandable by the end users. Another consideration and requirement for the development of our platform is that an online platform must provide results in an acceptable time manner (i.e., matter of milliseconds). In the next chapter, a comprehensive discussion on the development of Mediator platform is provided along with the description and analysis of its main components. A real case scenario and screenshots from the front-end provided to make Mediator more understandable to the reader.
Chapter 5

5. Building an online platform for user profiling – Mediator

In this chapter, we describe the architecture of Mediator platform that has been designed and developed in the context of this work taking into consideration the challenges described in Chapter 1 and the main purposes of this work. Namely, to develop not only a novel platform, named Mediator, which is online and open source creating a complete user profile including a wide spectrum of valuable psycho-demographics aspects identified in the literature, but also an explainable user profiling approach using machine learning algorithms. The developed application consists of two main individual components: the frontend and the backend, as typical for a web application architecture. In the following, the proposed and developed architecture is described in detail, as well as the most important libraries that have been used are presented. Finally, a presentation of the user interface is given, including the process to insert the username of the Twitter account that a participant desires to analyze, and the presentation of the final results to the end user in a graphic user interface.

5.1 Use Case Scenario

The main aim of Mediator is to build an online app that predicts profiling aspects of a Twitter user, such as age, gender, emotion, sentiment, personality, ethnicity and location creating a complete user profile including psycho-demographic aspects. The main advantage of Mediator against other user profiling tools that are popular and available online is that the platform incorporates explainability capabilities over the final results using advanced and sophisticated techniques and thus the end user has the ability to interpret the results.

Once the user visits the page, there is a search bar where he can put the username of the user, he wants to further analyze creating a user profile. Afterwards, he can select what specific features he would like to analyze from the Twitter profile. Then pressing the enter, the analysis starts. This may take a few minutes depending on how many traits the user has chosen to analyze. After that, the results are displayed in the page. Pictures from Mediator have been placed in the following sections. The display page might present all the results (depending on the selections) and the user needs to scroll down to see each aspect. Last but not least, Mediator provides explainability over the final results, and thus the user can understand how these results were exported providing transparency over the results.
5.2 System Requirements and Architecture

The main goal of this research work is to build an online app that predicts profiling aspects of a Twitter user, such as age, gender, emotion, sentiment, personality, ethnicity and location. In general, the developed platform aims to create a complete user profile including psycho-demographic aspects. Another crucial feature of the platform is the provision of explainability capabilities over the final results using advanced and sophisticated techniques. That means that the predictions made by the machine learning models have to be explainable and understandable by the end users. For instance, the result of an individual’s age has to be explained. Specifically, “25 y.o.” age prediction about a Twitter user should also provide the reasons for this result, for example, “large number of twitter followers” or “high ratio of emoticons in text” could be two possible considered factors.

On top of that, the end user must be able, in a user-friendly interface, to search for a Twitter account and get some insights about that account. Those insights are expected to be the results of each aspect alongside with their explanations, as mentioned previously. The results consist of the labels that are described in the previous chapter, and each profiling aspect has different labels as results. The platform should be able to provide proper explanations in each prediction. These explanations are based on feature importance values, which indicate the importance of the selected features. It is obvious that some of the features are not as important as others and therefore, they should not be included in the final results.

In order to establish a reliable application, we have defined specific requirements that the system must meet in order to satisfy the user at each level. Generally, it is essential that the platform is easy in use and does not require special handling or particular skills from the users. On the contrary, we should assume that user could be anyone who has access to the internet and a browser; from experts to beginners. In particular, the main requirements that our developed platform should meet are presented in the following. The user profiling platform should:

- **Provide a user-friendly UI:** This term includes the “easiness or difficulty” of a user to navigate to the platform’s UI. A friendly interface can provide easy guidance to users that are not familiarized with the platform. In other words, this requirement assures that users will be able to understand and use the platform without any specific guidance or tutorial.
- **Predict accurate insights:** The machine learning models should perform effectively and efficiently providing accurate and correct results to the user.
- **Provide explainable results:** This requirement ensures that predictions and decisions made by the model’s agents have to be explainable. Explanations for each decision provide clarity and the users can trust the ml model, and thus the output results.
- **Provide fast response:** An online platform must have a short response time. Nowadays, internet users tend to expect quick responses (i.e., in milliseconds) and possible delays may lead to mistrust and abandonment of the platform.
- **Enable user feedback:** The predictions may be inaccurate. Therefore, user’s feedback is desired and essential in order to correct the wrong predictions and improve the results over time.
- **Be open source:** Projects developed as open source, collaboratively can be significantly expanded and optimized, as well as failures are much easier to be detected and improved.

- **Be available at any time a user requests platform’s service:** Generally, web applications are designed to be available anytime and by any device that can be connected to the Internet. This constitutes the main motivation and desired objective to design and develop a web application.

Due to the complexity of the system, a suitable architecture design is required to facilitate easy system development. In our system, several distinct modules have been designed performing different functions required for the final results. In addition to the initial development of the UI, the maintainability that can be achieved later is also important to consider while developing our system. For example, for the introduction of a new classifier for the predictions, it should be easier for the platform to be developed and plugged into the already in-production platform. The system architecture is presented in Figure 5-1 that follows.

![Figure 5.1: System Architecture](image-url)

As depicted in Figure 5-1, the platform consists of four basic elements for its proper operation. The frontend, the backend, the communication with Twitter social media platform, and the machine learning models. The first two elements are main components of the web application. While the other two can be considered part of the back-end element. Both of them are essential for the interface and server communication. Furthermore, the
backend executes the processes that are required for results. As extensively explained in the previous chapters, the machine learning models are responsible for making the predictions and returning the final results. Finally, the communication with Twitter social media platform is done in order to receive all the necessary account data in order to be channeled, afterwards, to the machine learning section for the final results.

5.3 Platform Implementation

5.3.1 Twitter API

Twitter social media platform provides an API for connecting and interacting with external applications (Twitter API Documentation, 2021). In order to access this tool, a Twitter account is required and upon request on Twitter itself, the social media user may be provided with access tokens, which, however, have some limitations in terms of the amount of data the subsequent user can receive. In the design of the proposed user profiling platform, this limitation has been taken into account and therefore, it is impossible to lose the connection between this tool and our developed user profiling platform, but a slight delay in the results might be considered.

In our application, we have used two endpoints, which concern, the first regarding the user download and the second about the tweets that an individual has published. The tweets that are received for each individual user are limited to the first 100 in their profile, because firstly, Twitter social media platform poses this limitation and secondly, so that our application not to be overloaded, as afterwards the models will have to provide explanations for their results using this amount of data. As we discuss in the following, some machine learning models only use the data from the user profile, while others use just the content from the tweets of the user, and finally, some models make use of the combination of these two types of input data. In all cases, they return to the JSON format output, which is now one of the most commonly used for a wide spectrum of applications, while it is readable even by humans.

5.3.2 Frameworks and libraries

Application development is often done by using a software framework. The latter provides pre-installed libraries and tools that the developer can use to implement their own platform. In this way, the required development time is reduced, and the developer has more time to focus on the design of his application as well as its optimization in terms of performance evaluation and usability. In addition, an important element is the segmentation offered by this type of software frameworks.

Therefore, the use of a Framework was necessary for the implementation of our user profiling tool. The application of this research work was developed using Flask software framework and python programming language. In addition, several libraries supported by
this programming language are used for implementing the functions of the online platform, the training of machine learning models and the communication with Twitter API. The most important are described in the following:

- **Flask**: According to the Wikipedia, Flask “is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries to be implemented”. It is able to support many crucial application features which are useful for rapid development and easy maintenance. The key is that Flask is written in Python which provides compatibility with many libraries for data science (Welcome to Flask — Flask Documentation, 2021).

- **Sklearn**: It is a well-known library that contains ready-made efficient tools for software development including machine learning. The libraries supported and provided by Sklearn cover every stage in such an application (i.e., our user profiling application). On top of that, the way of its implementation is extremely pleasant and easy to understand and use, because it follows specific standards and thus, it can be applied in a wide spectrum of applications with no or minimal modifications (scikit-learn: machine learning in Python, 2021).

- **Tweepy**: It is a python library that enables communication between an external app with Twitter social media platform. It supports many operations functions to get or post data and can be used by AI bots. Also, it is able to set a time limit between requests to avoid restrictions of Twitter application (Tweepy, 2021).

- **Pandas**: It is a commonly used library for data management and offers a wide variety of tools for representing and viewing statistics. The most important element of this library is the Dataframe tool. This element is supported by libraries used for developing machine learning models and therefore, it is an asset to any data scientist willing to develop such projects (Pandas - Python Data Analysis Library, 2021).

### 5.3.3 Technologies

Figure 5.2 depicts all the technologies used to develop the application of our user profiling platform. As it can be seen, there are two main parts of the platform:

- Back-end
- Front-end

The **backend**, also referred to as the data access layer of the application, is implemented in python programming language. Flask framework is chosen for backend implementation due to its suitable capabilities for our application, the easiness of its installation and its rapid development. Also, an important factor is that Flask framework is developed in python which supports a wide spectrum of data science algorithms and functions including machine learning and artificial intelligence. Also, there is compatibility with Django and the selection of the framework depends on the familiarity of the developer of each of these two frameworks. Finally, it is separated in many different components that perform various functions suitable for user profiling applications. **Frontend**, also referred to as presentation layer, is the interface that a user is able to interact with the developed application. The technologies that have been used for frontend are HTML, CSS and JavaScript. Also, Flask supports Jinja2 which provides additional capabilities towards the communication between
the application’s backend and frontend parts. On top of that, a significant integrated feature of Flask is its capability to implement some main functions allowing parameters to be transmitted from the backend to the frontend directly with the implementation of simple commands.

![Diagram of technologies](image.png)

*Figure 5.2: Technologies of the platform.*

The most important part of the developed application is, undoubtedly, the development and implementation of the machine learning models, while this is also the main purpose of creating this user profiling tool. The machine learning models are part of the rest of the application, but it is not mandatory to use Flask for its operation. On the contrary, it is feasible to use the developed platform and simply plug it to another technology/framework for developing and implementing the machine learning models.

### 5.4 Backend - architecture

The backend of the application is implemented in python programming language. Also, Flask software framework is implemented in Python. As discussed previously, the main role of backend in an application is to further process the request received from the frontend and respond respectively. In particular, backend establishes the logic and enables the computations and actions that required in order to provide the proper results back to frontend according to the relevant request.

More precisely, in the context of our work, firstly, the backend receives a request from the frontend part of the application and forwards this request to the Profile Builder (PB) section of the backend. PB is the organizer of the whole User Profiling service/process in our developed platform. Therefore, this is the responsible element to take all the necessary steps towards the accessing and gathering of the data from Twitter social media platform, and then, to forward those data to the module that is in charge of performing the machine learning algorithms and the respective explainability processes. The pre-trained existing models return the result separately to PB. After receiving all the results - separately from each model - PB merges them and generates the final/total results. After that, the final/total
results are forwarded back to UI, and the end user is able to explore them getting valuable insights out of them.

The feature extraction constitutes a separate section/module in the developed user profiling application and can be used by any of the developed models that our platform includes. A class, namely FeatureExtractor, has been implemented in this module, which includes all the features that can be extracted from the text in user’s tweets. In particular, this class is used by individual classes that relate to each different aspect, i.e., each characteristic of the individual traits (e.g., age, ethnicity, etc.) includes different features. So, the individual classes use the FeatureExtractor method to collect the necessary features. This module receives and returns data in pandas Dataframe format.

5.5 Explainable Component

As mentioned in the previous chapters, the explainability is one of the most important parts of this research work. Many libraries have been widely used for this purpose. In our work, we are using the LIME library (Ribeiro et. al., 2016), which is available online, as well as it is implemented in python as the rest of the developed components. LIME library implements techniques for enabling explainability over AI results. In particular, the importance of the predictions is returned by LIME in the form of a list: each profiling trait has a list of the most important features. These importance values indicate that they have positively or negatively affected the prediction of each class for classification and regression ML problems. Then, the results returned in the user interface are only a small part of the whole list as they are not all equally important, and they do not contribute all equally to the final prediction results. We have set that a number of approximately 10 values for each feature will be presented. Especially, in the case of models that contain text, the number of features exported by the BoW can be a few thousand, and such a size would not be easily manageable and understandable by the user.

It is worthwhile to highlight that the explanations for the two different types of features, i.e., text and behavioral feature, are differentiate. In particular, text features are represented using wordclouds, while behavioral features are presented using charts. A wordcloud is more understandable and a graphic representation can provide the importance of the words that mostly affect the model’s predictions (either positively or negatively). On the other hand, the importance of behavioral features in charts may be either negative or positive. Therefore, visual representations of their state (i.e., positive, negative) along with their importance can be achieved by bar charts, radar plots etc.

5.6 User Interface - Frontend

In this subsection, we present the system interface, that is the frontend part of the application. This is the interaction between the end user and the platform. It is designed to be as simple as possible so that it does not require much effort or many actions from the users, and thus to be user-friendly and accessible to a large public. The process that a user should follow to use our user profiling platform is the following:
1. Firstly, the user has to enter the username of the account which intends to analyze.
2. Secondly, the user has to press the search button.
3. Then, the backend part of the application performs the required actions, in the background, and after a few seconds, the results are displayed on the page.
4. Finally, the results are presented together with their explanations. In addition, we provide some additional information about each feature to clarify and highlight how this can be read and explained by the user.

Nevertheless, the results may take a while to be displayed and the user may leave the page considering that there is a problem. To avoid such issues and to keep the user in the page, we have included a loading page. It is worthwhile to highlight that the model predictions along with their explanations are displayed in different formats depending on the type of features. In Figure 5.3, the search page section of the app is presented. As an input, a user is needs to select the desired aspect which wants to analyze and type the username of the Twitter account for analysis. Then by pressing the ‘Check Account’ button the analysis will began previewing a loading page to ensure the user that the page is still working.

![Figure 5.3: Mediator - Home screen](image-url)
Figure 5.4: Mediator - Age classification

The word-cloud below highlights the most representative words for this age group category used by the user.

Figure 5.5: Mediator - Gender classification

Contextual model explanations

The explanations above are extracted from contextual data. Word-cloud contains representative words for the predicted class Male - Female.

Representative words - Explanations from Text model
In Figures 5.4 and Figure 5.5, age prediction and gender prediction are presented. The explainability results are given in word-clouds. The word-clouds are presenting the most representative words for each class that was predicted. More specifically, for age prediction, the percentages of each class are given and the one with the highest score is the final prediction. The word-cloud are presenting the words only for the higher class (predicted class). Furthermore, gender prediction is shown on top of the page and there are two types of explanations, context and textual. The first one provides some features that may affect each class while the second one is the most representative words for the predicted class. The final prediction for gender is given by the text model as we described in Chapter 3.

For sentiment analysis the percentage of each class is presented at the top of the screen using with half gauge diagrams (see Figure 5.6). The predictions for sentiment, as mentioned in the previous chapters, made for each tweet separately and a total percentage of each class calculated for the analysis. On top of that, a number of tweets predicted for each class is given for explainability. In case of a non-predicted class then a message is appeared under the appropriate section to mention that there were no tweets predicted with the corresponding class.
Emotion

Chart presents the percentage of each emotion detected in the tweets of the user. The word-clouds below highlights the most representative words for each emotion used by the user.

Figure 5.7: Mediator - Emotion prediction

Anger: 20.0%

Disgust: 30.0%

Sadness: 9.0%

Joy: 29.0%

Figure 5.8: Mediator - Emotion prediction word-clouds
Figure 5.7 shows the overall results for the emotion on a radar plot along with some information about the user. The results are also presented below (see Figure 5.8) with a percentage for each category and there are word-clouds which presents the most representative words for all of them. We have to mention again, that each category is independent from the others and each percentage can be from zero to one hundred.

These charts presents the most valuable features for each of the aspects of OCEAN model. Features with values greater than zero affect the prediction to high. Values lower than zero affect each prediction to the lower values. Value range for each prediction: [0-5]

Figure 5.9: Mediator - Personality prediction

The predictions for the personality of the individual are presented in Figure 5.9. At the top of the screen, a radar plot presents the final predictions for each aspect of Big Five in an interactive chart. Above the radar plot, we can see that there are five buttons which contain the letters that come from the acronym OCEAN (Big Five). Clicking on each of them displays features that could reveal a personality trait. The explanations are presented in the form of a bar chart and extracted by our context model which has no impact to the final predictions. There are negative or positive numbers for each feature. Each of them, negative and positive, affects the prediction closer to zero or closer to 5 respectively.

Topics and interests are presented in Figures 5.10 and 5.11. We have chosen to export the
most common words, bigrams and hashtags used by users as we consider them to be representative of their interests. Verifying the account, we have used for this aspect we noticed that this account belongs to a user whose profession is journalism. In addition, we added PyLDAvis which is popular to export topics using LDA technique (Sharma, 2021).

Figure 5.10: Mediator - Topics 1 prediction

Figure 5.11: Mediator - Topics 2 LDA
In this chapter we presented the architecture of the system that we have developed, Mediator, as part of this dissertation. We presented the two main parts of the system front-end and back-end as well as the explainable component. Furthermore, we analyzed the main python libraries that our platform based. A use case scenario from a user perspective is given and the page structure described in details and screenshots provided for the reader. In the next chapter, comments and discussion about our developed platform and its main contribution are summarized. On top of that, future work on further improvement of the platform is discussed.
Chapter 6

6. Conclusions & Future Work

The contribution of this work in the direction of an efficient and robust user profiling solution is twofold. Firstly, we designed and implemented a Mediator platform, which is online and open source creating a complete user profile including a wide spectrum of valuable psycho-demographics aspects identified in the literature. In particular, the psycho-demographics aspects which are supported by our developed platform are categorized as demographics and psychographics. More specifically, it includes insights about social media user’s age, gender, sentiment, emotion, personality traits that exist in Big Five model, location, ethnicity, and interests. The data used to export the individual’s insights is obtained from their Twitter account. To the best of our knowledge, it is the first work identified in the literature that enables user profiling based on numerous valuable insights about the user. Previous works used only a limited number of insights about the social media user. Therefore, this work contributes towards a novel solution for a more complete user profile creation, and thus an efficient and effective user profiling.

Secondly, we developed an explainable and interpretable user profiling approach using machine learning algorithms. User profiling handles personal confidential data, and thus data privacy concerns may arise involving the right of personal privacy, concerning storing, re-purposing, provision to third parties, and displaying of information about oneself via the Internet. As in this work and user profiling works generally, we are looking to deploy artificial intelligence and machine learning using those data, questions of accountability and transparency are particularly important, and if we are unable to properly deliver explainability, in our algorithms, we will seriously be limiting the potential impact of artificial intelligence for such solutions. According to the European Commission “Explainable AI: the basics” report32, there are a few reasons why some form of interpretability / explainability in user profiling solutions might be desirable. These include: (a) giving users confidence and trust in the system, (b) prevent bias, (c) meeting regulatory and policy requirements, (d) improving system design, (e) assessing risk, robustness, and vulnerability, and (f) understanding and verifying the outputs from a system.

For the above reasons, we used machine learning methodologies to extract the insights. On top of that, along with the user’s psycho-demographics aspects, we have included proper and reasonable explanations for each model’s decision to provide transparency and clarity, as well as avoid any kind of bias or discrimination. It is worthwhile to highlight that clarity ensures transparency, proper functioning, and non-discrimination. Finally, reference is made to the problem that arises when researchers are trying to implement efficient machine learning algorithms ensuring the property of interpretability/explainability in their developed solutions. In some cases, it is not possible to use those techniques properly with advanced model structures. Thus, a trade-off between explainability and the accuracy of the models was made.

As already described in the previous chapters, although there are several tools and platforms that make predictions of psycho-demographic aspects, none of the existing approaches enables explainability/interpretability over the predicting values. However, in our case, the models follow specific architectures in order to be able to use techniques to enable explainability/interpretability while achieving comparable prediction accuracy with the state-of-the-art models. As a result, Mediator is an innovative solution that covers all requirements that we identified in the systematic literature review for an accurate, effective and efficient user profiling.

Towards this direction, the main contribution of this work is a web app that predicts profiling aspects of a Twitter user, such as age, gender, emotion, sentiment, personality, ethnicity, topics and location creating a complete user profile including psycho-demographic aspects: Mediator. The main advantage of Mediator against other user profiling tools that are popular and available online is that the platform incorporates explainability capabilities over the final results using advanced and sophisticated techniques and thus the end user has the ability to interpret the results. Once the user visits the page, there is a search bar where he can put the username of the user, he wants to further analyze creating a user profile. Afterwards, he can select what specific feature he would like to analyze from the Twitter profile. Then pressing the enter, the analysis starts. This may take a few minutes depending on the traits the user has chosen to analyze and the data which are available to the Twitter account. After that, the results are displayed in the page. Last but not least, Mediator provides explainability over the final results, and thus the user can understand how these results were exported providing transparency over the results.

There is no doubt that the most important part of the developed application is the development and implementation of the machine learning models, while this is also the main purpose of creating this user profiling tool. For this reason, Chapter 3 and 4 were dedicated to the comprehensive analysis of the machine learning models that were developed for making Mediator an effective and efficient user profiling tool. In particular, in this dissertation, we provided a comprehensive report of all the stages that we passed through during the experimentation for every aspect, as well as the models that we developed. In addition, we presented the stages from the phase of the data collection of ground truth datasets to the evaluation of performance, describing also in detail the performance metrics that have been used to evaluate the process of the proposed methodologies and model architectures. This process is separated in the following steps: 1) Gather the datasets that were used for the training of our models and their performance evaluation, as well as the modifications that were applied to them in order to fit the problems and our use case; 2) Perform hyperparameter optimization for the models of every aspect to identify the multiple parameters that needed to be configured and optimized; 3) Extract and present the results of the conducted experiments and the metrics that were used to highlight performance and the effectiveness of the developed models; and finally 4) discuss the explainability findings and the validation of our results.

Finally, as described previously, there is a trade-off and the models are not SOTA and were trained in order to be able to provide explanations that are understandable by the end users. Another consideration and requirement for the development of our platform is that a web platform must provide results in an acceptable time manner (i.e., matter of milliseconds). In Chapter 5, a comprehensive discussion on the development of Mediator platform is provided along with the description and analysis of its main components. Additionally, a real case scenario and screenshots from the front-end provided to make Mediator more
understandable to the reader. Furthermore, in this chapter, we also presented the architecture of the system that we have developed, Mediator, as part of this dissertation. We presented the two main parts of the system front-end and back-end as well as the explainable component. Furthermore, we analyzed the main python libraries that our platform based. A use case scenario from a user perspective is given and the page structure described in details and screenshots provided for the reader.

To conclude, we would like to discuss future work on further improvement of the platform and its performance in order to be effective, efficient and as much user-friendly as possible. As future work directions, we would like to refer to the integration of additional features in order to be able to predict the individual’s educational background and/or income, as well as further improvement of the trained models and machine learning algorithms used for those predictions. In addition, we would also like to achieve better representation of the explanations offered in each prediction, something that based on the literature, this can be done using LDA to present the words as topics etc. Finally, support for multiple languages for analysis by our ML models also comprises a future research work direction because as of now this is limited only to English language.
References


Bayot, R., Gonçalves, T., & Quresma, P. (2015). Author Profiling of Twitter User. CLEF.


