ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανίχνευση φύλου στο Twitter μέσω υβριδικού αλγορίθμου μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ορέστης Π. Γιαννακόπουλος-Καρακώντης

Επιβλέπουσα: Ιωάννα Ρουσσάκη
Επίκουρη Καθηγήτρια

Αθήνα, Νοέμβριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανίχνευση φύλου στο Twitter μέσω υβριδικού αλγορίθμου μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ορέστης Π. Γιαννακόπουλος-Καρακώντης

Επιβλέπουσα : Ιωάννα Ρουσσάκη
Επίκουρη Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Νοεμβρίου 2017.

.....	.....	.....
Ιωάννα Ρουσσάκη	Ευστάθιος Συκάς	Συμεών Παπαβασιλείου
Επίκουρη Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2017

.....
Ορέστης Π. Γιαννακόπουλος-Καρακώντης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα μέσα κοινωνικής δικτύωσης παρουσιάζουν ιδιαίτερη αύξηση στη δημοτικότητα τους τα τελευταία χρόνια, με το Twitter να αποτελεί ένα από τα πιο δημοφιλή. Παρά την ελευθερία που παρέχεται από το Twitter για την πρόσβαση σε δεδομένα που παράγονται από τους χρήστες του, δεν υπάρχουν υποχρεωτικά πεδία στα προφίλ των χρηστών που να δηλώνουν τα δημογραφικά τους στοιχεία. Το γεγονός αυτό σε συνδυασμό με την χρησιμότητα των δημογραφικών στοιχείων για ερευνητικούς αλλά και εμπορικούς σκοπούς, έχει οδηγήσει σε πολυάριθμες έρευνες που προτείνουν έμμεσους τρόπους ανίχνευσης διαφόρων δημογραφικών στοιχείων χρηστών που διαθέτουν λογαριασμό στο Twitter. Ειδικότερα για την ανίχνευση του φύλου, έχουν προταθεί ποικίλες μεθοδολογίες που βασίζονται στη χρήση αλγορίθμων μηχανικής μάθησης. Οι περισσότερες από αυτές τις μεθοδολογίες εξαρτώνται από την γλώσσα των χρηστών και χρησιμοποιούν πολυάριθμα χαρακτηριστικά πραγματοποιώντας την εκπαίδευση των αλγορίθμων μηχανικής μάθησης σε χώρους υψηλών διαστάσεων. Για το λόγο αυτό, τέτοιες προσεγγίσεις περιορίζονται κυρίως σε συγκεκριμένες εθνικότητες χρηστών, είναι ιδιαίτερα χρονοβόρες και παρουσιάζουν υψηλή κατανάλωση υπολογιστικών πόρων με αποτέλεσμα να μην μπορούν να επεκταθούν αποδοτικά σε μεγάλους πληθυσμούς χρηστών του Twitter.

Στην παρούσα εργασία προτείνεται ένας αποδοτικός τρόπος για την ανίχνευση του φύλου χρηστών του Twitter, χρησιμοποιώντας μόνο την φωτογραφία προφίλ, το όνομα και το χρώμα θέματος που είναι διαθέσιμα από τα προφίλ των χρηστών. Ο συνδυασμός αυτών των στοιχείων δεν έχει χρησιμοποιηθεί ξανά σε προηγούμενες εργασίες εν γνώση μας. Κατά τη διάρκεια της μελέτης, πραγματοποιήθηκαν πειράματα με αλγορίθμους Naive Bayes, Μηχανές Διανυσμάτων Υποστήριξης και Πιθανολογικά Νευρωνικά Δίκτυα ως ταξινομητές φύλου επιβλεπόμενης μάθησης. Η υλοποίηση τους έγινε στη γλώσσα Python με χρήση των βιβλιοθηκών scikit-learn και neupy. Επίσης χρησιμοποιήθηκαν οι υπηρεσίες δυο αξιόπιστων εξωτερικών πηγών: του Face++ για την ανάλυση των εικόνων και του Genderize για την ταξινόμηση των ονομάτων κατά φύλο. Στο πρώτο μέρος των πειραμάτων, παρουσιάζονται τρεις διακριτές προσεγγίσεις, η κάθε μια βασισμένη σε ένα από τα τρία προαναφερθέντα πεδία του προφίλ, και αξιολογείται η απόδοση τους. Επίσης εξάγονται συμπεράσματα για τις διαφορετικές συμπεριφορές των δυο φύλων στο Twitter, σύμφωνα με το κάθε πεδίο. Στη συνέχεια, κάθε διακριτή προσέγγιση συνδυάζεται σε έναν υβριδικό αλγόριθμο μηχανικής μάθησης. Χρησιμοποιώντας τρία Πιθανολογικά Νευρωνικά Δίκτυα και μια Μηχανή Διανυσμάτων Υποστήριξης σε διαφορετικά στάδια της διαδικασίας, επιτεύχθει 87.2% accuracy στις προβλέψεις φύλου χρησιμοποιώντας τη μέθοδο 5-fold cross-validation για κάθε μοντέλο επιβλεπόμενης μάθησης. Όλα τα πειράματα πραγματοποιήθηκαν σε δείγμα χρηστών αντιπροσωπευτικό του συνολικού πληθυσμού του Twitter για να γίνει βέβαιο ότι η προτεινόμενη μέθοδος μπορεί να γενικευτεί αξιόπιστα.

Η εργασία αυτή καταδεικνύει ότι χρησιμοποιώντας μόνο έναν πολύ μικρό αριθμό χαρακτηριστικών από τα προφίλ χρηστών στο Twitter, είναι δυνατή η ανίχνευση του φύλου τους πετυχαίνοντας έναν πολύ καλό συνδυασμό κλιμακωσιμότητας (scalability) και ακρίβειας (accuracy).

Λέξεις κλειδιά

κοινωνικά δίκτυα, Twitter, ανίχνευση φύλου, μηχανική μάθηση, υβριδικός αλγόριθμος, εξόρυξη δεδομένων, ανάλυση δεδομένων, χρώμα θέματος, φωτογραφία προφίλ χρήστη, όνομα χρήστη, Μηχανές Διανυσμάτων Υποστήριξης, Πιθανολογικά Νευρωνικά Δίκτυα, Naive Bayes

Abstract

Online social networks have increased their popularity over the last few years, with Twitter being one of the most prominent. Despite the freedom of access granted by Twitter for its user-generated data, there are no obligatory fields in the user profiles that contain their demographics. This fact, along with the usefulness of demographics for scientific and commercial purposes, has led to a vast amount of studies that focus on indirect ways to detect the demographics of Twitter users. Especially for the detection of gender, many methodologies have been suggested based on machine learning algorithms. The majority of these methodologies are language-dependent and make use of a large amount of features in high-dimensional spaces. Due to this fact, they can target only users with specific nationalities and they are particularly time- and resource-consuming, so they can't efficiently scale to large populations of Twitter.

Our approach is simple and efficient, using only the profile picture, the display name and the theme color extracted from profiles of users to detect their gender. This combination of fields has not been utilized in previous works to our knowledge. Throughout this study, we experimented with Naive Bayes, Support Vector Machines and Probabilistic Neural Networks as supervised learning gender classifiers. The classifiers were implemented in Python via the scikit-learn and neupy libraries. We also utilized the services of two reliable external sources: Face++ for image analysis and Genderize for name classification by gender. In the first part of our experiments, we present three distinct approaches based in each of the three aforementioned fields and we evaluate their performance. We also make conclusions about the different behaviors of the two genders in Twitter, in accordance to each field. In the next part, we combine each individual approach in a hybrid machine learning algorithm. Using three Probabilistic Neural Networks and a Support Vector Machine in different stages of the process, we achieve 87.2% accuracy in the prediction of gender using 5-fold cross-validation for every supervised model. All the experiments were made based on a user set that is representative of the entire Twitter population to make sure that our approach can be generalized reliably.

We conclude that by using only a very small amount of features for Twitter users, we can classify them by gender with a very good combination of scalability and accuracy.

Key words

social networks, Twitter, gender detection, machine learning, hybrid algorithm, data mining, data analysis, theme color, profile picture, name, Support Vector Machines, Probabilistic Neural Networks, Naive Bayes

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κυρία Ιωάννα Ρουσσάκη για την εμπιστοσύνη που μου έδειξε και τη δυνατότητα που μου έδωσε να εκπονήσω τη διπλωματική μου εργασία σε ένα τόσο ενδιαφέρον θέμα, καθώς και για την αμέριστη συμπαράσταση και την πολύτιμη καθοδήγηση που μου προσέφερε καθόλη τη διάρκεια της διαδικασίας.

Επίσης θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Νίκο Καλατζή για την βοήθεια του σε πρακτικά κομμάτια της εργασίας.

Ακόμα θα ήθελα να ευχαριστήσω θερμά την μητέρα μου, τον πατέρα μου και την γιαγιά μου για την ανυστερόβουλη αγάπη τους και την έμπρακτη στήριξη και υπομονή τους προς το προσωπό μου σε όλα τα στάδια της ζωής μου.

Τέλος ευχαριστώ όλα τα άτομα εντός και εκτός σχολής που συνέβαλαν το καθένα με διαφορετικό τρόπο στις ευχάριστες στιγμές των φοιτητικών μου χρόνων. Παναγιώτη, Μάκη, Χριστίνα, Αλέξανδρε, Βασίλη, Χάρη, Παύλο, Φοίβη και Δημήτρη σας ευχαριστώ.

Πίνακας Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Μέσα κοινωνικής δικτύωσης και Twitter

Τα μέσα κοινωνικής δικτύωσης (online social networks - OSNs) αποτελούν αναπόσπαστο κομμάτι της καθημερινότητας εκατομμυρίων ανθρώπων σε όλο τον κόσμο οι οποίοι τα χρησιμοποιούν για να επικοινωνήσουν απευθείας μεταξύ τους, αλλά και για να μοιραστούν απόψεις, πληροφορίες και ενδιαφέροντα μέσω αναρτήσεων στα online προφίλ τους. Τα διάφορα OSNs προσφέρουν διαφορετικές δυνατότητες κοινωνικής δικτύωσης στους χρήστες του Διαδικτύου απασχολώντας αρκετό από τον ελεύθερο χρόνο τους. Παρατηρούμε την ανάπτυξη ποικίλων online κοινοτήτων που διαφέρουν ως προς τον αριθμό και την αλληλεπίδραση των μελών τους αλλά και ως προς τα άτυπα μοτίβα συμπεριφορών. Μερικά παραδείγματα πολύ δημοφιλών online κοινωνικών δικτύων είναι το Facebook, το Instagram, το Pinterest και φυσικά το Twitter.

Το Twitter είναι μια online υπηρεσία κοινωνικής δικτύωσης η οποία ιδρύθηκε το Μάρτιο του 2006 από τους Jack Dorsey, Evan Williams, Noah Glass και Biz Stone. Σε αντίθεση με άλλα δημοφιλή OSNs, στα οποία οι χρήστες ως επί το πλείστον δημοσιεύουν προσωπικές φωτογραφίες, μοιράζονται αγαπημένα τους τραγούδια ή επικοινωνούν με ιδιωτικές συνομιλίες, το Twitter είναι προσανατολισμένο στη δημοσίευση σύντομων και συχνών μηνυμάτων. Τα μηνύματα αυτά ονομάζονται tweets και μπορούν να έχουν μήκος το πολύ 140 χαρακτήρες, γεγονός που κατατάσσει το Twitter στην κατηγορία του microblogging. Το περιεχόμενο των tweets μπορεί να εκφράζει μεταξύ άλλων την προσωπική άποψη του χρήστη για κάποιο επίκαιρο πολιτικό, κοινωνικό ή ψυχαγωγικό γεγονός, την αναπαραγωγή κάποιας πρόσφατης είδησης ή την ενημέρωση για κάποιο φλέγον ζήτημα. Συχνά οι χρήστες του Twitter χρησιμοποιούν hashtags στα tweets τους για να τα συσχετίσουν με αντίστοιχα tweets που αναφέρονται στο ίδιο θέμα συζήτησης. Επίσης υπάρχει η δυνατότητα ένα tweet να περιλαμβάνει φωτογραφίες, βίντεο και συνδέσμους. Ο κάθε χρήστης ακολουθεί (Following) τα άτομα για τα οποία θέλει να

ενημερώνεται όταν δημοσιεύουν ένα tweet και αντιστοίχως ακολουθείται και αυτός από άλλους χρήστες (Followers). Οι χρήστες έχουν τη δυνατότητα να απαντήσουν στα tweets άλλων χρηστών (reply) ή να τα αναδημοσιεύσουν (retweet). Τέλος υπάρχει η δυνατότητα για ανταλλαγή απευθείας ιδιωτικών μηνυμάτων μεταξύ των χρηστών (direct messages).

1.2 Δημογραφικά στοιχεία από το Twitter

Η ραγδαία επέκταση των OSNs τα τελευταία χρόνια, έχει οδηγήσει πολλούς ερευνητές διαφορετικών πεδίων να στρέψουν το ενδιαφέρον τους προς αυτά σε μια προσπάθεια ανάλυσης και επεξεργασίας ποικίλων χαρακτηριστικών του παγκόσμιου πληθυσμού.

Ειδικότερα, το Twitter αυξάνει συνεχώς τους χρήστες του με αποκορύφωμα τους 328 εκατομμύρια ενεργούς χρήστες που κατέγραψε το δεύτερο τρίμηνο του 2017 σύμφωνα με το Statista (Σχήμα 1.1[1]). Κατά μέσο όρο οι χρήστες αυτοί δημοσιεύουν 500 εκατομμύρια status updates (tweets) ημερησίως[2]. Για αυτό το λόγο, το Twitter αποτελεί μια τεράστιου μεγέθους άμεσα προσβάσιμη πηγή μετάδοσης επίκαιρων ειδήσεων αλλά και μια ανεκτίμητης αξίας βάση δεδομένων που μπορεί να χρησιμοποιηθεί τόσο από εταιρείες όσο και απο ακαδημαικές κοινότητες για τη διεξαγωγή ερευνών επιστημονικού και εμπορικού ενδιαφέροντος. Πιο συγκεκριμένα, τα δεδομένα που παράγονται από τους χρήστες του Twitter μπορούν να χρησιμοποιηθούν για τον προσδιορισμό των δημογραφικών χαρακτηριστικών του παγκόσμιου πληθυσμού όπως είναι το φύλο, η ηλικία και η εθνικότητα. Τα εξαγόμενα δημογραφικά χαρακτηριστικά έχουν μεγάλη χρησιμότητα σε πολλούς φορείς διαφορετικής ιδιότητας. Ενδεικτικά μπορούν να χρησιμοποιηθούν από:

- Εταιρείες ή πανεπιστήμια που θέλουν να μελετήσουν την κοινή γνώμη (ανάλυση συναισθήματος, πολιτική δραστηριότητα), την εξάπλωση ασθενειών ή ακόμα και για να βελτιώσουν τον χρόνο απόκρισης σε φυσικές καταστροφές [14],[15].

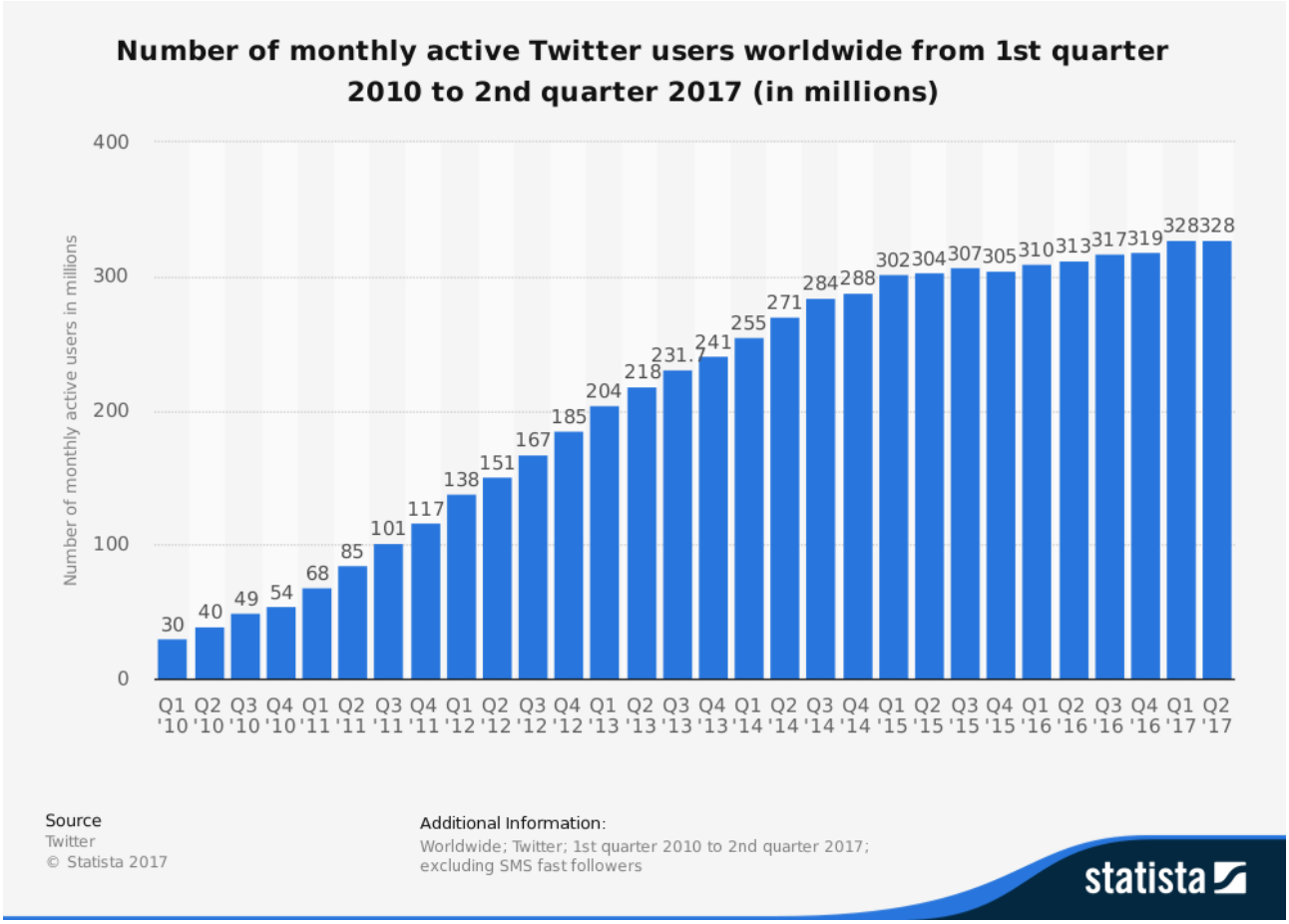
- Κοινωνικούς επιστήμονες που μελετάνε το πώς διαφοροποιείται η συμπεριφορά των ανθρώπων στις online κοινωνίες ανάλογα με τα δημογραφικά τους στοιχεία.

[1] Πηγή: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[2] https://business.twitter.com/

- Εταιρείες που θα μπορούν να προσαρμόσουν καλύτερα τις στοχευμένες διαφημίσεις και το δημόσιο πρόσωπο που δείχνουνε ανάλογα με τα στοιχεία του κοινού στο οποίο απευθύνονται [2].

- Ψυχολογικές έρευνες σχετικά με τα άτομα μιας συγκεκριμένης κοινότητας [16].

- Την αστυνομία που μπορεί να χρησιμοποιήσει δημογραφικά χαρακτηριστικά ως μέρος των ερευνών της [2].



Σχήμα 1.1: Αριθμός των ενεργών χρηστών του Twitter παγκοσμίως από το 2010 έως το 2017

Μερικά από τα πλεονεκτήματα του Twitter για την εξαγωγή δημογραφικών στοιχείων είναι τα εξής:

- Χάρις στον τεράστιο αριθμό των μελών του, προσφέρει μεγάλο όγκο πληροφορίας παραγόμενης από τους χρήστες του, η οποία μπορεί να οδηγήσει σε αξιόπιστα συμπεράσματα.

- Τα δεδομένα που παράγονται από τους χρήστες του ανανεώνονται συνεχώς και έτσι είναι δυνατόν να ανανεωθούν παράλληλα οι εκτιμήσεις των δημογραφικών χαρακτηριστικών που μελετούνται.

- Είναι ένας δωρεάν και εύκολος τρόπος συλλογής μεγάλης ποσότητας δεδομένων. Το περιεχόμενο που παράγεται από τους χρήστες του Twitter και τα σχετικά μεταδεδομένα είναι διαθέσιμα στην κοινότητα και επιπλέον υπάρχουν φιλικά στον χρήστη web interfaces και APIs για την πρόσβαση σε αυτά τα δεδομένα.

- Το Twitter έχει την επιλογή του *geotagging* δηλαδή τη δημοσίευση ενός tweet μαζί με την πληροφορία της τοποθεσίας. Το geotagging χρησιμοποιείται όλο και περισσότερο και έτσι τα συμπεράσματα που βγαίνουν από τις διάφορες έρευνες θα μπορούν στο μέλλον να στοχεύουν σε συγκεκριμένες γεωγραφικές τοποθεσίες.

- Λόγω της πληθώρας παλαιότερων σχετικών ερευνών, υπάρχουν αρκετές εφαρμοσμένες μεθοδολογίες με τα ανάλογα αποτελέσματα και συμπεράσματα στις οποίες μπορεί να βασιστεί και να τις επεκτείνει μια νέα έρευνα.

1.3 Κίνητρο και Προκλήσεις

Παρόλο που τα δημογραφικά στοιχεία αποτελούν πολύ καθοριστικό παράγοντα για την συμπεριφορά των ανθρώπων, οι περισσότερες επιστημονικές έρευνες που επικεντρώνονται στα OSNs δεν τα λαμβάνουν υπόψιν λόγω απουσίας επαρκών στοιχείων.

Το Twitter ειδικότερα, αν και παρέχει εύκολη πρόβαση στα δεδομένα των χρηστών του, δεν περιλαμβάνει υποχρεωτικά πεδία στα οποία ο χρήστης πρέπει να δηλώσει τα δημογραφικά του στοιχεία όπως το φύλο και την ηλικία του. Η πληροφορία των δημογραφικών στοιχείων πρέπει λοιπόν να εξαχθεί με έμμεσους τρόπους. Οι τρόποι

αυτοί ποικίλλουν και μπορεί να περιλαμβάνουν την ανάλυση των στοιχείων που εξάγονται από τα πεδία του προφίλ ενός χρήστη ή την μελέτη των tweets που δημοσιεύει.

Με βάση τα παραπάνω μπορούμε να εξάγουμε το συμπέρασμα ότι υπάρχει ένα χάσμα μεταξύ της χρησιμότητας των δημογραφικών χαρακτηριστικών από χρήστες του Twitter και της ευκολίας με την οποία μπορούν αυτά να αποκτηθούν. Από το σύνολο των δημογραφικών στοιχείων, το φύλο είναι ένα από τα θεμελιώδη και διακριτά, με πληθώρα μελετών να επικεντρώνονται στις διαφορές των αντρών και των γυναικών και πώς αυτές επηρεάζουν το σχετικό πεδίο έρευνας.

Στην παρούσα διπλωματική εργασία επιχειρούμε να συμβάλλουμε στο δημοφιλές αυτό και γεμάτο προκλήσεις πεδίο έρευνας για την ανίχνευση του φύλου των χρηστών του Twitter.

1.4 Αντικείμενο της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανίχνευση του φύλου των χρηστών του Twitter. Η ανίχνευση του φύλου γίνεται μέσω αλγορίθμων μηχανικής μάθησης οι οποίοι εκπαιδεύονται με βάση κάποια χαρακτηριστικά που εξάγονται από τα προφίλ των χρηστών. Τα χαρακτηριστικά αυτά είναι ανεξάρτητα της γλώσσας οπότε η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί για τον συνολικό πληθυσμό του Twitter χωρίς να χρειάζεται περιορισμός σε συγκεκριμένες εθνικότητες. Επιπλέον λόγω της απλότητας και του πολύ μικρού αριθμού των χαρακτηριστικών που λαμβάνονται για κάθε χρήστη, αλλά και της διαδικασίας που ακολουθήθηκε για την επεξεργασία τους, η προσέγγιση αυτή είναι πολύ φιλική ως προς την κατανάλωση χρόνου και υπολογιστικών πόρων και μπορεί να κλιμακώσει για την ανίχνευση του φύλου σε βάσεις δεδομένων που περιέχουν μεγάλο αριθμό χρηστών.

Οι στόχοι της παρούσας εργασίας είναι οι ακόλουθοι:

- Επίτευξη μιας ανταγωνιστικής απόδοσης στην ανίχνευση φύλου των χρηστών του Twitter, ακολουθώντας μια απλή και χαμηλής πολυπλοκότητας προσέγγιση η οποία εφαρμόζεται σε ένα δείγμα χρηστών αντιπροσωπευτικό του συνολικού πληθυσμού του Twitter.

- Εξέταση ενός συνδυασμού στοιχείων από τους χρήστες του Twitter που δεν έχει πραγματοποιηθεί ξανά σε παρόμοιες μελέτες και εξαγωγή συμπερασμάτων για την χρησιμότητα του καθενός ξεχωριστά.

- Παρατήρηση της επίδρασης των αλλαγών στη δομή του Twitter στην δυνατότητα ανίχνευσης φύλου των χρηστών του.

1.5 Οργάνωση κειμένου

Το υπόλοιπο της παρούσας διπλωματικής εργασίας είναι οργανωμένο ως εξής:

Στο Κεφάλαιο 2 περιγράφεται η δομή ενός προφίλ στο Twitter, καθώς και κάποιες σημαντικές αλλαγές σε αυτή τη δομή σε σχέση με παλαιότερα. Στη συνέχεια κάνουμε μια σύντομη εισαγωγή στο αντικείμενο της Μηχανικής Μάθησης εστιάζοντας στους αλγορίθμους επιβλεπόμενης μάθησης και το πρόβλημα της Ταξινόμησης. Παρουσιάζουμε το θεωρητικό υπόβαθρο για τους τρεις αλγορίθμους επιβλεπόμενης μάθησης που χρησιμοποιήσαμε και αναφέρουμε μετρικές και μεθόδους για την αξιολόγηση τους. Τέλος, αναφέρουμε τη διαδικάσια που ακολουθήσαμε για την επικοινωνία με το API του Twitter μέσω της βιβλιοθήκης Tweepy και περιγράφουμε σύντομα την χρησιμοποιήση των βιβλιοθηκών Scikit-learn και Neupy για την υλοποιήση των αλγορίθμων επιβλεπόμενης μάθησης, παραθέτοντας μικρά δείγματα κώδικα.

Στο Κεφάλαιο 3 προσφέρουμε μια σφαιρική εικόνα για το πρόβλημα της ανίχνευσης του φύλου ενός ατόμου αναφέροντας σχετικές μελέτες. Στη συνέχεια γίνεται μια εκτενής αναφορά σε μελέτες με αντικείμενο την ανίχνευση φύλου συγκεκριμένα σε χρήστες του Twitter.

Στο Κεφάλαιο 4 περιγράφουμε τις αμιγείς προσεγγίσεις μας για την ανίχνευση φύλου με χρήση του χρώματος θέματος, της εικόνας προφίλ και του ονόματος των χρηστών του Twitter. Για κάθε προσέγγιση γίνεται μια σύντομη ανάλυση του αντίστοιχου πεδίου και παρουσιάζεται η εξωτερική πηγή υπηρεσιών ή οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν μαζί με αναλυτική περιγραφή της όλης διαδικασίας. Τέλος, για κάθε προσέγγιση παρουσιάζουμε τα αποτελέσματα μας και διατυπώνουμε τα συμπεράσματα μας σχετικά με τη διαφορετική συμπεριφορά των δυο φύλων σύμφωνα με το κάθε εξεταζόμενο πεδίο.

Στο Κεφάλαιο 5 περιγράφουμε την υβριδική προσέγγιση που προτείναμε για την ανίχνευση φύλου των χρηστών του Twitter. Αρχικά αναφέρουμε τα πλεονεκτήματα της προσέγγισης αυτής. Στη συνέχεια, εξηγούμε την διαδικασία εισαγωγής αλγορίθμων μηχανικής μάθησης στις διαδικασίες χρησιμοποίησης της εικόνας προφίλ και του ονόματος, με σκοπό την ομοιομορφία τους σε σχέση με την χρησιμοποίηση του χρώματος. Στη συνέχεια, περιγράφουμε την υλοποίηση του υβριδικού αλγορίθμου και παραθέτουμε τα αποτελέσματα που πετύχαμε με την χρησιμοποίηση του για διαφορετικούς συνδυασμούς μοντέλων επιβλεπόμενης μάθησης. Τέλος, εξετάζουμε τον περιορισμό της ανίχνευσης φύλου σε υποσύνολα χρηστών για τα οποία έχουμε μεγαλύτερη πεποίθηση ορθότητας και καταγράφουμε τα σχετικά ευρήματα.

Στο Κεφάλαιο 6 παρουσιάζονται τα τελικά συμπεράσματα της εργασίας και προτείνονται μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Υπόβαθρο

2.1 Δομή του προφίλ χρηστών Twitter

2.1.1 Τρέχουσα έκδοση προφίλ χρηστών Twitter

Κατά τη δημιουργία ενός προφίλ στο Twitter, τα μόνα υποχρεωτικά πεδία που πρέπει να συμπληρώσει ο χρήστης και φαίνονται δημόσια στο προφίλ του χρήστη είναι το username και το display name. Οι υπόλοιπες πληροφορίες του προφίλ είναι προαιρετικές. Μέσω των προαιρετικών πεδίων, ο χρήστης έχει τη δυνατότητα να διαμορφώσει το προφίλ του συμπληρώνοντας προσωπικά του στοιχεία και ανεβάζοντας φωτογραφίες. Επιπλέον, ο χρήστης μπορεί να επιλέξει για το προφίλ του ένα χρώμα θέματος (theme color) της αρεσκείας του.

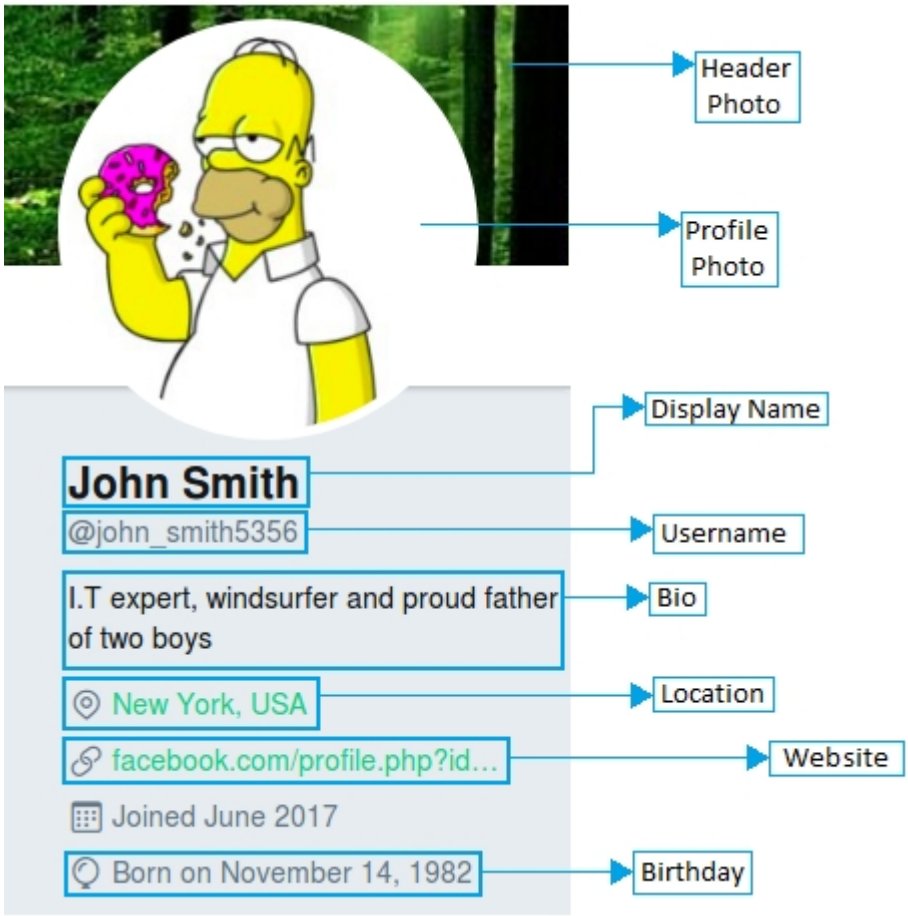
Τα πεδία που φαίνονται στο προφίλ ενός χρήστη και τα οποία μπορεί να επεξεργαστεί είναι τα εξής[3]:

- Φωτογραφία κεφαλίδας-Header photo
- Φωτογραφία προφίλ-Profile photo
- Username
- Όνομα-(Display) Name
- Τοποθεσία-Location
- Ιστοσελίδα-Website
- Χρώμα θέματος-Theme Color
- Βιογραφικό-Bio (maximum 160 characters)
- Γενέθλια-Birthday

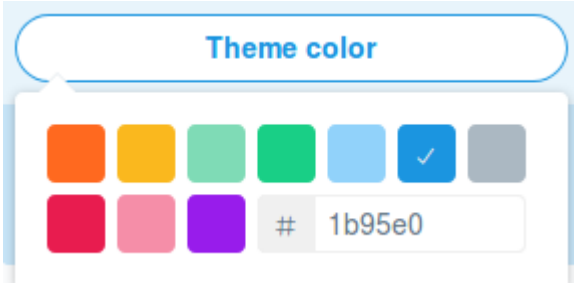
Από τα παραπάνω πεδία μόνο το username πρέπει να είναι μοναδικό για κάθε χρήστη.

[3] https://help.twitter.com/en/managing-your-account/how-to-customize-your-profile

Στο Σχήμα 2.1 φαίνεται η μορφή που έχουν τα πεδία που αναφέραμε πιο πάνω. Στο Σχήμα 2.2 φαίνεται ο τρόπος επιλογής του χρώματος θέματος.



Σχήμα 2.1: Δομή προφίλ χρήστη Twitter



Σχήμα 2.2: Επιλογή χρώματος θέματος στο Twitter

2.1.2 Παλαιότερες εκδόσεις προφίλ χρηστών Twitter

Το Twitter πραγματοποιεί ανά περιόδους κάποιες αλλαγές στη δομή του προφίλ των χρηστών του. Μερικές από αυτές τις αλλαγές είναι ικανές να επηρεάσουν τις σχετικές έρευνες που χρησιμοποιούν το Twitter σε κάποια συγκεκριμένη χρονική περίοδο για την εξαγωγή δημογραφικών στοιχείων.

Μια από αυτές τις αλλαγές αφορά τα χρώματα τα οποία ο χρήστης μπορεί να επιλέξει για να διαμορφώσει το προφίλ του. Με αναζήτηση στο Διαδίκτυο[4] αλλά και με βάση παλαιότερες μελέτες [2] βλέπουμε ότι παλαιότερα οι χρήστες του Twitter είχαν την δυνατότητα να επιλέξουν χρώματα για 5 διαφορετικά σημεία του προφίλ τους. Τα σημεία αυτά ήταν τα εξής: *profile background, profile text, profile sidebar border, profile link, profile sidebar fill*. Κατά τη διάρκεια εκπόνησης της παρούσας εργασίας, ο χρήστης μπορούσε να επιλέξει μόνο ένα theme color για να τροποποιήσει την εμφάνιση του προφίλ του. Χρησιμοποιώντας το Twitter API, το οποίο περιγράφεται στην Ενότητα 2.3.1, ανακαλύψαμε ότι μας επιστρέφονται κάθε χρήστη τα πεδία *profile background color, profile text color, profile sidebar border color, profile link color, profile sidebar fill color* και όχι το *theme color.* Με πειραματισμούς σε ένα προφίλ που φτιάξαμε στο Twitter για τους σκοπούς της εργασίας, είδαμε ότι αλλάζοντας το theme color το μόνο πεδίο χρώματος που επηρεαζόταν από αυτά που μας επιστρέφονταν από το Twitter API ήταν το *profile link color.* Επίσης, παρατηρώντας τα προφίλ άλλων χρηστών είδαμε ότι τα υπόλοιπα 4 πεδία χρώματος τα οποία μας επιστρέφονταν με διάφορες τιμές για τον καθένα, δεν εμφανίζονταν πουθενά στο προφίλ τους. Φαίνεται δηλαδή ότι οι αλλαγές που είχαν κάνει μερικοί χρήστες στο προφίλ τους παλιότερα υπάρχουνε σαν προσβάσιμα δεδομένα, αλλά δεν φαίνονται στο Twitter. Για τους σκοπούς της παρούσας εργασίας, πήραμε τις τιμές του *theme color* για κάθε χρήστη μέσω του *profile link color* που μας επέστρεφε το API του Twitter.

Παρατηρήσαμε επίσης ότι σε παλαιότερες μελέτες ([11],[4]), το πεδίο που χαρακτηρίζει το Twitter ως *(display) name* αναφέρεται ως “full name” ή “user name” ενώ το πεδίο που το Twitter χαρακτηρίζει ως *username*, αναφέρεται ως “screen name”. Δεν γνωρίζουμε αν αυτό προέκυψε από επίσημη αλλαγή της ονομασίας των πεδίων από το Twitter ή από επιλογή των συγγραφέων. Στο υπόλοιπο της παρούσας εργασίας συμπεριλαμβανομένης και της αναφοράς σε παλαιότερες μελέτες, τα πεδία θα αναφέρονται ως *display name* και *username* σε

[4] http://profilerehab.com/twitter-help/change_twitter_colors

συνάφεια με την επίσημη ονομασία που δίνει το Twitter και το Σχήμα 2.1.

Τέλος, βάσει ανακοίνωσης στο blog του Twitter[5], σκοπεύεται να αλλάξει το μέγιστο μέγεθος για τα tweets από 140 χαρακτήρες σε 280. Όπως είναι εύκολο να συμπεράνει κανείς, αυτές οι αλλαγές στη δομή του Twitter θα επηρεάσουν σε μεγάλο βαθμό τις σχετικές έρευνες ανίχνευσης φύλου που βασίζονται στην ανάλυση των πεδίων του προφίλ του χρήστη και των tweets για να εξάγουν τα αποτελέσματα τους.

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένα πεδίο της Επιστήμης Υπολογιστών που δίνει στους υπολογιστές τη δυνατότητα να μαθαίνουν και να κάνουν προβλέψεις πάνω σε δεδομένα, χωρίς να προγραμματίζονται ρητά για αυτό το σκοπό. Το πεδίο της Μηχανικής Μάθησης έχει εφαρμογή μεταξύ άλλων στη Μηχανική Όραση, στην Οπτική Αναγνώριση Χαρακτήρων (OCR) και στο φιλτράρισμα των email.

Οι αλγόριθμοι Μηχανικής Μάθησης χωρίζονται σε 2 κατηγορίες: τους αλγορίθμους επιβλεπόμενης μάθησης (supervised learning) και μη επιβλεπόμενης μάθησης (unsupervised learning). Οι αλγόριθμοι επιβλεπόμενης μάθησης εκπαιδεύονται πάνω σε δείγματα δεδομένων εισόδου γνωρίζοντας την επιθύμητη εξοδο για κάθε δείγμα. Στη συνέχεια χρησιμοποιούν τη γνώση που απέκτησαν σε καινούρια δεδομένα εισόδου, επιχειρώντας να κάνουν προβλέψεις ή εκτιμήσεις για τις αντίστοιχες εξόδους.
Αντίθετα, οι αλγόριθμοι μη επιβλεπόμενης μάθησης δεν εκπαιδεύονται με επιθυμητές εξόδους αλλά προσπαθούν να βρουν τη δομή ή τις σχέσεις μεταξύ των δεδομένων εκπαίδευσης συνήθως χωρίζοντας τα δεδομένα σε ομάδες (clustering). Στη συνέχεια, παίρνοντας ως είσοδο ένα καινούριο δείγμα δεδομένων το τοποθετούν στην κατάλληλη ομάδα.

Το κάθε δείγμα δεδομένων το οποίο επεξεργάζονται οι αλγόριθμοι Μηχανικής Μάθησης έχει κάποια χαρακτηριστικά (features). Ένα χαρακτηριστικό είναι μια μετρήσιμη ιδιότητα των δεδομένων η οποία μπορεί να έχει διακριτές ή συνεχείς τιμές. Παραδείγματος χάριν, για ένα σύνολο δεδομένων όπου το κάθε δείγμα αντιπροσωπεύει ένα κτήριο, μπορεί να υπάρχουν τα χαρακτηριστικά *ύψος* με

[5] https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

συνεχείς τιμές και *χρώμα* με διακριτές τιμές. Τα χαρακτηριστικά των δεδομένων είναι αυτά που δίνουν την απαραίτητη γνώση στους αλγορίθμους Μηχανικής Μάθησης προκειμένου να εκτελέσουν την επιθυμητή εργασία (πρόβλεψη επιθυμητής εξόδου, ομαδοποίηση, κ.λπ.).

2.2.1 Αλγόριθμοι Επιβλεπόμενης Μάθησης - Το πρόβλημα της Ταξινόμησης

Μια από τις κυριότερες λειτουργίες για την οποία χρησιμοποιούνται οι αλγόριθμοι Μηχανικής Μάθησης είναι η ταξινόμηση των δεδομένων (classification). Το πρόβλημα της ταξινόμησης εντάσσεται στην κατηγορία της επιβλεπόμενης μάθησης. Κατά τη διαδικασία της ταξινόμησης, ένας αλγόριθμος επιβλεπόμενης μάθησης καλείται να αναγνωρίσει σε ποια κατηγορία ανήκει ένα δείγμα δεδομένων. Αρχικά ο αλγόριθμος τροφοδοτείται με ένα σύνολο δειγμάτων εκπαίδευσης όπου για κάθε δείγμα λαμβάνει τα σχετικά χαρακτηριστικά του και την κατηγορία στην οποία ανήκει. Όταν τελειώσει το στάδιο της εκπαίδευσης, ο αλγόριθμος είναι σε θέση να ταξινομήσει ανά κατηγορία καινούρια δείγματα δεδομένων.

2.2.2 Θεωρητικό υπόβαθρο αλγορίθμων Επιβλεπόμενης Μάθησης

Στη συνέχεια αναφέρουμε το εν συντομία το θεωρητικό υπόβαθρο τριών αλγορίθμων επιβλεπόμενης μάθησης που χρησιμοποιούνται συχνά σε παρόμοιες μελέτες ως ταξινομητές για την ανίχνευση φύλου (βλ. Κεφάλαιο 3) και τους οποίους χρησιμοποίησαμε στην παρούσα διπλωματική εργασία για την ανίχνευση φύλου των χρηστών του Twitter. Οι αλγόριθμοι αυτοί είναι ο Naive Bayes, οι Μηχανές Διανυσμάτων Υποστήριξης και τα Πιθανολογικά Νευρωνικά Δίκτυα. Οι πηγές πληροφορίας που χρησιμοποιήσαμε είναι οι [40], [41], [42].

2.2.2.1 Naive Bayes

Οι ταξινομητές Naive Bayes έχουν χρησιμοποιηθεί σε πολλές παρόμοιες μελέτες με σκοπό την ανίχνεση φύλου στους χρήστες του Twitter. Είναι αλγόριθμοι επιβλεπόμενης μάθησης βασισμένοι στην εφαρμογή του θεωρήματος του Bayes με την “naive” υπόθεση της ανεξαρτησίας μεταξύ όλων των ζευγών χαρακτηριστικών. Δεδομένης μιας μεταβλητής κλάσης y και ενός εξαρτημένου

διανύσματος χαρακτηριστικών x_1 ως x_n, το θεώρημα του Bayes διατυπώνει την ακόλουθη σχέση:

$$P(y|x_1,\dots,x_n)=\frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)}$$

Χρησιμοποιώντας την "naive" υπόθεση ανεξαρτησίας έχουμε:

$$P(x_i|y,x_1,\dots,x_{(i-1)},x_{(i+1)},\dots,x_n)=P(x_i|y)$$

Για όλα τα i αυτή η σχέση απλοποιείται σε:

$$P(y|x_1,\dots,x_n)=\frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_{1,},\dots,x_n)}$$

Επειδή το $P(x_1,\dots,x_n)$ είναι σταθερό δεδομένου της εισόδου, μπορούμε να χρησιμοποιήσουμε τον ακόλουθο κανόνα ταξινόμησης:

$$P(y|x_1,\dots,x_n)\propto P(y)\prod_{i=1}^{n}P(x_i|y) \quad \Rightarrow \quad \hat{y}=\underset{y}{arg\,max}\,P(y)\prod_{i=1}^{n}P(x_i|y)$$

και μπορούμε να χρησιμοποιήσουμε Maximum A Posteriori (MAP) εκτίμηση για να υπολογίσουμε τα P(y) και $P(x_i|y)$, όπου το πρώτο είναι η σχετική συχνότητα της κλάσης y στο σύνολο εκπαίδευσης.

Οι διαφορετικοί Naïve Bayes ταξινομητές διαφέρουν κυρίως στις υποθέσεις που κάνουν σχετικά με την κατανομή του $P(x_i|y)$.

Ο Gaussian Naive Bayes (GNB) υποθέτει ότι η πιθανότητα των χαρακτηριστικών είναι Γκαουσιανή, δηλαδή:

$P(x_i|y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(\frac{-(x_i-\mu_y)^2}{2\sigma_y^2}\right)$, όπου οι παράμετροι σ_y, μ_y υπολογίζονται χρησιμοποιώντας εκτίμηση μέγιστης πιθανοφάνειας.

Ο Multinomial Naive Bayes (MNB) είναι η έκδοση του αλγορίθμου για πολυωνυμική κατανομή δεδομένων. Η κατανομή παραμετροποιείται από τα διανύσματα $\theta_y = (\theta_{y1},\dots,\theta_{yn})$ για κάθε κλάση y, όπου n είναι ο αριθμός των χαρακτηριστικών και θ_{yi} είναι η πιθανότητα $P(x_i|y)$ του χαρακτηριστικού i να εμφανίζεται σε ένα δείγμα που ανήκει στην κλάση y. Οι παράμετροι θ_y υπολογίζονται με βάση τη σχετική

συχνότητα:

$\widehat{\theta_{yi}} = \frac{N_{yi} + a}{N_y + an}$, όπου $N_{yi} = \sum_{x \in T} x_i$ είναι ο αριθμός των φορών που το χαρακτηριστικό i εμφανίζεται σε ένα δείγμα της κλάσης y στο training set T και

$N_y = \sum_{i=1}^{|T|} N_{yi}$ είναι ο συνολικός αριθμός όλων των features για την κλάση y.

Παρά τις υπεραπλουστευμένες υποθέσεις τους, οι Naïve Bayes ταξινομητές έχουν αρκετά καλή απόδοση σε ποικίλα πραγματικά προβλήματα, όπως η ταξινόμηση εγγράφων και το φιλτράρισμα spam. Απαιτούν μικρό αριθμό από δείγματα εκπαίδευσης για να υπολογίσουν τις απαραίτητες παραμέτρους. Επιπλέον είναι εξαιρετικά γρήγοροι σε σύγκριση με πιο εκλεπτυσμένες μεθόδους. Ακόμα, λόγω της φύσης τους βοηθούν στην ελαχιστοποίηση των προβλημάτων που προκύπτουν από το “curse of dimensionality”, το οποίο αναφέρεται σε διάφορα φαινόμενα που εμφανίζονται με την ανάλυση και οργάνωση των δεδομένων σε χώρους υψηλών διαστάσεων (συχνά με εκατοντάδες ή χιλιάδες διαστάσεις).

Αν και οι Naive Bayes ταξινομητές έχουν ικανοποιητική απόδοση, είναι γνωστοί ως κακοί εκτιμητές της πιθανότητας με την οποία μια πρόβλεψη τους είναι σωστή.

2.2.2.2 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM) είναι μια κατηγορία δικτύων πρόσθιας τροφοδότησης[6]. Ανήκουν στην οικογένεια των αλγορίθμων επιβλεπόμενης μάθησης και δεδομένου ενός επισημασμένου (labelled) συνόλου δεδομένων εκπαίδευσης κατασκευάζουν ένα βέλτιστο διαχωριστικό υπερεπίπεδο που κατηγοριοποιεί τα καινούρια δείγματα.

Η λειτουργία του αλγορίθμου SVM βασίζεται στην εύρεση του υπερεπιπέδου που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων. Τα σημεία αυτά είναι τα πλέον δύσκολα να ταξινομηθούν και ονομάζονται *διανύσματα υποστήριξης (support vectors)*. Το διπλάσιο αυτής της απόστασης ονομάζεται *περιθώριο διαχωρισμού*. Το βέλτιστο υπερεπίπεδο είναι αυτό που μεγιστοποιεί το περιθώριο διαχωρισμού με αποτέλεσμα να μεγιστοποιεί την απόσταση μεταξύ των

[6] Στα δίκτυα πρόσθιας τροφοδότησης η πληροφορία κινείται μόνο προς τα εμπρός από τους κόμβους εισόδου προς τους κόμβους εξόδου, δηλαδή δεν υπάρχουν βρόχοι ανάδρασης στο δίκτυο.

σημείων δεδομένων των 2 διαφορετικών κλάσεων όπως φαίνεται και στο Σχήμα 2.2.

Μια θεμελιακή ιδέα για την ανάπτυξη του αλγορίθμου μάθησης ενός SVM είναι ο πυρήνας εσωτερικού γινομένου (inner product kernel) μεταξύ ενός διανύσματος υποστήριξης $\mathbf{x_i}$ και ενός διανύσματος **x** το οποίο αντλείται από το χώρο δεδομένων εισόδου. Ο αλγόριθμος μάθησης αναφέρεται και ως μέθοδος πυρήνα (kernel method), λόγω της σημαντικής ιδιότητας ότι τα διανύσματα υποστήριξης αποτελούνται από ένα μικρό υποσύνολο σημείων δεδομένων τα οποία εξάγει ο αλγόριθμος μάθησης από το ίδιο το δείγμα εκπαίδευσης. Οι πιο διαδεδομένες μέθοδοι πυρήνα είναι η γραμμική, η πολυωνυμική, η σιγμοειδής και η radial basis function (RBF). Οι 3 τελευταίες μέθοδοι πυρήνα μπορούν να χρησιμοποιηθούν για την ταξινόμηση μη γραμμικά διαχωρίσιμων δεδομένων.

Ακολουθεί μια συνοπτική περιγραφή για τις μαθηματικές σχέσεις που διέπουν το υπερεπίπεδο ενός SVM και το περιθώριο διαχωρισμού. Η περιγραφή αυτή αναφέρεται στην περίπτωση 2 γραμμικά διαχωρίσιμων κλάσεων για λόγους απλότητας:

Θεωρούμε το δείγμα εκπαίδευσης $\{(x_i, d_i)\}_{i=1}^{N}$, όπου $\mathbf{x_i}$ είναι το i-οστό διάνυσμα εισόδου και $\mathbf{d_i}$ είναι η αντίστοιχη επιθυμητή έξοδος ($d_i = \pm 1$ για τις 2 δυνατές κλάσεις). Η εξίσωση μιας επιφάνειας απόφασης με τη μορφή ενός υπερεπιπέδου που εκτελεί το διαχωρισμό είναι $w^T x + b = 0$, όπου **x** είναι ένα διάνυσμα εισόδου, **w** ένα προσαρμόσιμο διάνυσμα βαρών και **b** μια πόλωση. Οπότε μπορούμε να ορίσουμε $w^T x_i + b \geq 0$ για $d_i = +1$ και $w^T x_i + b < 0$ για $d_i = -1$. Θεωρούμε το βέλτιστο υπερεπίπεδο με $\mathbf{w = w_o}$ και $\mathbf{b = b_o}$. Η απόσταση ενός σημείου δεδομένων **x** από το βέλτιστο υπερεπίπεδο ορίζεται ως $\frac{w_0^T x + b_0}{\|w_0\|}$. Έχοντας θεωρήσει ένα κανονικοποιημένο βέλτιστο υπερεπίπεδο που ικανοποιεί την ιδιότητα $\min_{x_i \in X} |w_0^T x_i + b_0| = 1$, το περιθώριο διαχωρισμού είναι ίσο με $\frac{2}{\|w_0\|}$. Από την εξίσωση αυτή παρατηρούμε ότι η μεγιστοποίηση του περιθωρίου διαχωρισμού μεταξύ δυο κλάσεων είναι ισοδύναμη με την ελαχιστοποίηση της Ευκλείδιας νόρμας του διανύσματος βαρών **w** από την οποία προκύπτει το βέλτιστο διάνυσμα βαρών $\mathbf{w_o}$.

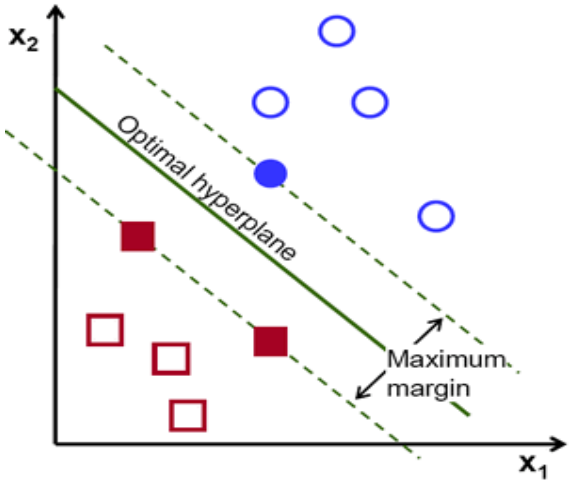
Μερικά από τα πλεονεκτήματα των Μηχανών Διανυσμάτων Υποστήριξης είναι τα εξής :

1. Είναι αποτελεσματικές σε χώρους υψηλών διαστάσεων.

2. Παραμένουν αποτελεσματικές σε περιπτώσεις που ο αριθμός των διαστάσεων είναι μεγαλύτερος από των αριθμό των δειγμάτων δεδομένων.

3. Είναι ευέλικτες: Μπορούν να οριστούν διαφορετικές μέθοδοι πυρήνα για τη συνάρτηση απόφασης.

Τέλος, αναφέρουμε επίσης ότι οι Μηχανές Διανυσμάτων Υποστήριξης χρησιμοποιούνται και σε προβλήματα παλινδρόμησης (regression).



Σχήμα 2.3: Βέλτιστο υπερεπιπεδο ενός SVM

2.2.2.3 Πιθανολογικά Νευρωνικά Δίκτυα

Ένα Πιθανολογικό Νευρωνικό Δίκτυο (Probabilistic Neural Network-PNN) είναι ένα είδος νευρωνικού δικτύου πρόσθιας τροφοδότησης που έχει ευρεία χρήση σε προβλήματα ταξινόμησης και αναγνώρισης προτύπων. Περιέχει 4 επίπεδα κόμβων:

- *Input Layer*: Στο επίπεδο αυτό υπάρχουν N κόμβοι εισόδου, όπου N είναι ο αριθμός των χαρακτηριστικών των δειγμάτων που εξετάζονται. Ο κάθε κόμβος τροφοδοτεί την τιμή του αντίστοιχου χαρακτηριστικού εισόδου σε κάθε νευρώνα του Hidden/Pattern Layer.

- *Hidden/Pattern Layer*: Αυτό το επίπεδο περιέχει έναν νευρώνα για κάθε δείγμα (διάνυσμα χαρακτηριστικών) εκπαίδευσης. Οι νευρώνες χωρίζονται σε K ομάδες, όπου K ο αριθμός των διαφορετικών κλάσεων των δεδομένων. Θεωρούμε ότι η κλάση k έχει P διανύσματα χαρακτηριστικών στο σύνολο εκπαίδευσης $\left\{x^{(p)}: p=1,\ldots,P\right\}$. Με είσοδο ένα διάνυσμα χαρακτηριστικών **x** ο κάθε νευρώνας στην ομάδα k υπολογίζει μια Γκαουσιανή συνάρτηση κεντραρισμένη στα χαρακτηριστικά αυτού του δείγματος:

  $g_k(x)=\frac{1}{\sqrt{\left(2\pi\sigma^2\right)^N}}\exp\left(\frac{-\left\|x-x^{(p)}\right\|}{2\sigma^2}\right)$, όπου **x(p)** το διάνυσμα εκπαίδευσης που αντιστοιχεί στον συγκεκριμένο νευρώνα, **N** η διάσταση των διανυσμάτων χαρακτηριστικών και **σ** η τυπική απόκλιση. Όλοι οι νευρώνες της ομάδας k τροφοδοτούν τις τιμές που υπολογίσανε, στον κόμβο του Summation Layer που αντιστοιχεί σε αυτή την κλάση.

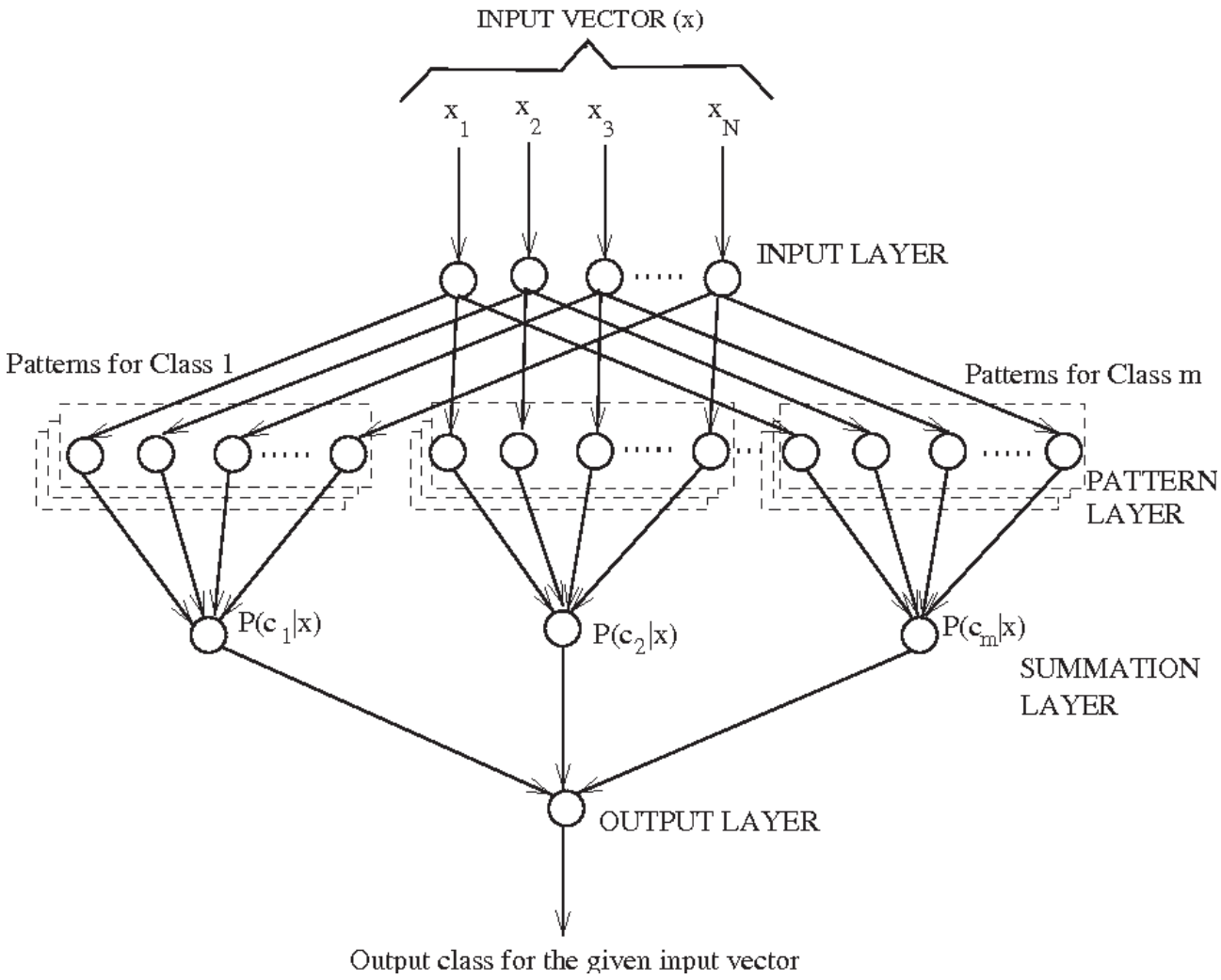
- *Summation Layer*: Αποτελείται από αριθμό κόμβων ίσο με τον αριθμό των κλάσεων προς ταξινόμηση K. Ο κόμβος που σχετίζεται με την κλάση k, παίρνει τις υπολογισμένες τιμές από τους νευρώνες του Hidden/Pattern Layer που ανήκουν στην αντίστοιχη κλάση και υπολογίζει με τη μέθοδο Parzen window τη συνάρτηση πυκνότητας πιθανότητας (probability density function – pdf) που αντιστοιχεί σε αυτή την κλάση:

  $f_k(x)=\frac{1}{\sqrt{\left(2\pi\sigma^2\right)^N}}\frac{1}{P}\sum_{p=1}^{P}\exp\left(\frac{-\left\|x-x^{(p)}\right\|}{2\sigma^2}\right)$.

- *Output Layer*: Αποτελείται από έναν νευρώνα ο οποίος παίρνει τις τιμές των συναρτήσεων $f_k(\mathbf{x})$ για όλες τις κλάσεις k από τους νευρώνες του προηγούμενου επιπέδου και δίνει σαν έξοδο την κλάση με τη μεγαλύτερη τιμή (maximum a posteriori value).

Στο Σχήμα 2.3[7] φαίνεται η αρχιτεκτονική ενός Πιθανολογικού Νευρωνικού Δικτύου που περιγράφηκε παραπάνω.

[7] Πηγή: https://www.semanticscholar.org/paper/Setting-up-a-Probabilistic-Neural-Network-for-Clas-Selekwa-Kwigizile/c25fbbe1e74bb78dbe207aa580b1769ff2ecb6d9



Σχήμα 2.4: Αρχιτεκτονική ενός PNN

Μερικα πλεονεκτήματα της χρήσης ενός PNN είναι τα εξής:

1. Η εκπαίδευση του δικτύου είναι πολύ γρήγορη συγκριτικά με τα perceptrons πολλαπλών επιπέδων.

2. Έχει καλή ανοχή σε ακραία σημεία δεδομένων.

3. Παράγει πιθανότητες για την κλάση που ανήκει κάποιο δείγμα με καλή αξιοπιστία.

4. Οι επιφάνειες απόφασης του προσεγγίζουν τον Bayes Classifier (ο Bayes

Classifier ελαχιστοποιεί την πιθανότητα λάθους ταξινόμησης).

2.2.3 Αξιολόγηση αλγορίθμων Επιβλεπόμενης Μάθησης

Τα μοντέλα επιβλεπόμενης μάθησης παρουσιάζουν μεγάλες διαφορές στην απόδοση τους ανάλογα με τον αλγόριθμο υλοποίησης τους, τα χαρακτηριστικά των δεδομένων εισόδου, τις επιθυμητές εξόδους και τη φύση του προβλήματος που καλούνται να επιλύσουν. Για να προσδιοριστεί η απόδοση ενός μοντέλου χρησιμοποιούνται διάφορες μέθοδοι και μετρικές. Μια μετρική σχετίζεται με τον δείκτη απόδοσης που λαμβάνεται υπόψη για την ποσοτικοποίηση της επιτυχίας ενός μοντέλου, ενώ οι μέθοδοι καθορίζουν τη διαδικάσια με την οποία εξάγεται αυτή η μετρική. Παρακάτω αναφέρουμε τις μετρικές και τις μεθόδους που χρησιμοποιούνται συχνά για την αξιολόγηση των δυαδικών ταξινομητών και εξηγούμε ποιες επιλέξαμε στην παρούσα εργασία και για ποιους λόγους.

2.2.3.1 Μετρικές Αξιολόγησης

Οι μετρικές που χρησιμοποιούνται συνήθως για την αξιολόγηση της απόδοσης ενός δυαδικού ταξινομητή αναφέρονται ακολούθως μαζί με τους αντίστοιχους τύπους:

- accuracy : $\frac{\#\,correct\ predictions}{\#\,all\ predictions}$

- precision : $\frac{T_p}{T_p+F_p}*100\%$, όπου T_p: ο αριθμός των true positives (δείγματα που ορθώς ταξινομήθηκαν ως θετικά), F_p: ο αριθμός των false positives (δείγματα που λανθασμένα ταξινομήθηκαν ως θετικά)

- recall : $\frac{T_p}{T_p+F_n}$, όπου T_p: ο αριθμός των true positives, F_n: ο αριθμός των false negatives (δείγματα που λανθασμένα ταξινομήθηκαν ως αρνητικά)

- F1 score : $\frac{2*precision*recall}{precision+recall}$

Στην παρούσα μελέτη χρησιμοποιήσαμε ως μετρική των μοντέλων επιβλεπόμενης μάθησης το *accuracy*. Το accuracy επιλέχθηκε για 2 λόγους. Πρώτον, είναι μια μετρική που χρησιμοποιείται ευρέως σε παρόμοιες έρευνες ανίχνευσης φύλου (βλ. Κεφάλαιο 3) και αποτελεί την μετρική που τονίζεται κυρίως για τη σύγκριση των αποτελεσμάτων μεταξύ των ερευνών αυτών. Δεύτερον, μέσω του accuracy μπορούμε να πάρουμε αρκετά αξιόπιστη πληροφορία για την απόδοση των δυαδικών ταξινομητών που χρησιμοποιούμε στην παρούσα μελέτη για την ανίχνευση φύλου σε χρήστες του Twitter. Ο λόγος είναι ότι η πρόβλεψη ότι ένας χρήστης είναι άντρας ή γυναίκα έχει την ίδια βαρύτητα για τα αποτελέσματα μας κατά την αξιολόγηση της ορθότητας της. Με άλλα λόγια, δεν υπάρχει κάποια μεροληψία ως προς το ποιο φύλο θα πρέπει να προβλέπεται τις περισσότερες φορές σωστά.

Σημειώνεται ότι σε άλλες περιπτώσεις το accuracy δεν είναι επαρκές και απαιτείται η χρήση και άλλων μετρικών. Για παράδειγμα, ένα μοντέλο-ταξινομητής το οποίο αποφασίζει με βάση κάποια δεδομένα αν ένας άνθρωπος πάσχει από μια αρρώστια ή όχι, χρειάζεται μετρικές που θα λαμβάνουν υπόψην τα false positives (άνθρωποι που διαγνώσθηκαν ότι πάσχουν από την ασθένεια λανθασμένα) και τα false negatives (άνθρωποι που πάσχουν από την ασθένεια και θεωρήθηκαν υγιείς), δίνοντας τους τον κατάλληλο τρόπο ποσοτικοποίησης. Σε αυτή την περίπτωση πρέπει να χρησιμοποιηθούν και τα precision και recall.

Στη συνέχεια της παρούσας εργασίας το *accuracy* θα αναφέρεται και ως “ακρίβεια” και δεν θα πρέπει να μπερδεύεται με το *precision* που έχει την ίδια ελληνική μετάφραση.

2.2.3.2 Μέθοδοι Αξιολόγησης

Για την πραγματοποίηση πειραμάτων με έναν ταξινομητή επιβλεπόμενης μάθησης υπάρχουνε διαφορετικές μέθοδοι σχετικά με τον τρόπο που γίνεται η εκπαίδευση και η αξιολόγηση του συστήματος. Μερικές από τις πιο διαδεδομένες μεθόδους είναι οι Holdout, k-fold cross-validation και leave-one-out cross-validation.

Κατά τη μέθοδο Holdout, το σύνολο δεδομένων χωρίζεται σε δυο αμοιβαίως αποκλειόμενα υποσύνολα όπου το ένα λειτουργεί σαν training set και το άλλο σαν test set (ένας συνήθης διαχωρισμός είναι το 70% του συνόλου δεδομένων να λειτουργεί σαν training set και το 30% σαν test set). Το μοντέλο επιβλεπόμενης

μάθησης εκπαιδεύεται πάνω στο training set και στη συνέχεια αξιολογείται η απόδοση του με βάση τις προβλέψεις του για τα δεδομένα του test set.

Κατά τη διαδικάσια του k-fold cross-validation τα δεδομένα χωρίζονται σε k υποσύνολα. Σε κάθε έναν από τους k γύρους της μεθόδου αυτής, τα k-1 υποσύνολα χρησιμοποιούνται ως training set για τον αλγόριθμο και το υποσύνολο που απομένει χρησιμοποείται ως test set με βάση το οποίο εξάγονται οι επιθυμητές μετρικές. Στο τέλος όλης της διαδικασίας υπολογίζεται ο μέσος όρος για κάθε μετρική και οι προκύπτοντες αριθμοί αποτελούν τον δείκτη απόδοσης του εκάστοτε αλγορίθμου μηχανικής μάθησης.

Τα πλεονεκτήματα της μεθόδου k-fold cross-validation είναι ότι υπάρχει μικρή μεροληψία (bias) στη διαδικασία μάθησης του αλγορίθμου καθώς ένα μεγάλο ποσοστό των δεδομένων χρησιμοποιούνται για την εκπαιδευση του αλγορίθμου και επιπλέον μειώνεται η διακύμανση (variance) καθώς όλα τα δεδομένα περνούν από το test set. Με άλλα λόγια, το k-fold cross-validation προσφέρει αντικειμενικά αποτελέσματα για την επίδοση ενός μοντέλου αποφεύγοντας το overfitting[8] του αλγορίθμου στα δεδομένα μέσω της χρήσης των διαφορετικών folds, και προσφέρει έναν τρόπο εξαγωγής συμπερασμάτων για το πόσο καλά γενικεύει αυτό το μοντέλο και πόσο μπορεί δυνητικά να αποδώσει καλά σε καινούρια δεδομένα.

Μια εξαντλητική έκδοση του cross-validation είναι το leave-one-out. Σε αυτό τον τύπο cross-validation, ο αριθμός των folds είναι ίσος με τον αριθμό των δειγμάτων στο σύνολο δεδομένων. Με άλλα λόγια, σε κάθε γύρο της μεθόδου χρησιμοποιείται ένα μόνο δείγμα δεδομένων ως test set ενώ όλα τα υπόλοιπα λειτουργούν ως training set. Στο τέλος της διαδικασίας υπολογίζονται οι μέσοι όροι των μετρικών που εξήχθησαν σε κάθε γύρο.

Στην παρούσα διπλωματική εργασία επιλέξαμε τη μέθοδο k-fold cross-validation για να υπολογίσουμε το accuracy των αλγορίθμων επιβλεπόμενης μάθησης που δοκιμάσαμε. Ο λόγος που επιλέξαμε αυτή τη μέθοδο είναι ότι προσφέρει ένα καλό συνδυασμό κατανάλωσης υπολογιστικών πόρων και αξιόπιστων αποτελεσμάτων. Πιο συγκεκριμένα, η μέθοδος Holdout δεν προτιμήθηκε επειδή δεν έχουμε στη διάθεση μας αρκετά δεδομένα ώστε να κάνουμε κάποιον διαχωρισμό του συνόλου δεδομένων μας χωρίς να χάσουμε σημαντική πληροφορία εκπαίδευσης ή ικανότητα

[8] Overfitting ονομάζεται το φαινόμενο κατά το οποίο ένας αλγόριθμος μηχανικής μάθησης έχει προσαρμοστεί πολύ στενά σε ένα σύνολο δεδομένων με αποτέλεσμα να αποτυγχάνει να αποδώσει καλά σε καινούρια δεδομένα (στην περίπτωση ενός ταξινομητή να μπορεί να προβλέψει σωστά την κλάση των καινούριων δεδομένων).

αξιόπιστου testing. Από την άλλη, η μέθοδος leave-one-out είναι πολύ απαιτητική υπολογιστικά και καταναλώνει υπερβολικό χρόνο.

Πιο συγκεκριμένα, για όλα τα πειράματα με αλγορίθμους επιβλεπόμενης μάθησης χρησιμοποιήσαμε stratified 5-fold cross-validation. Στην “stratified” εκδοχή της μεθόδου, σε κάθε fold περιέχεται περίπου το ίδιο ποσοστό δειγμάτων ανά κλάση όσο και σε ολόκληρο το σύνολο δεδομένων. Η εκδοχή αυτή γενικά είναι καλύτερη όσον αφορά τη μεροληψία και τη διακύμανση σε σύγκριση με το κανονικό cross-validation [17].

2.2.3.3 Αναζήτηση Πλέγματος

Κατά την διεξαγωγή πειραμάτων με έναν αλγόριθμο επιβλεπόμενης μάθησης, χρειάζεται πολύ συχνά η εύρεση των υπερπαραμέτρων του αλγορίθμου που οδηγούν στην βέλτιστη απόδοση του. Οι υπερπαράμετροι ορίζονται πριν τη διαδικάσια μάθησης και καθορίζουν ιδιότητες του μοντέλου όπως η πολυπλοκότητα του, η μορφή της επιφάνειας απόφασης κ.α. Μια δημοφιλής τεχνική που χρησιμοποιείται για αυτό το σκοπό είναι η αναζήτηση πλέγματος (grid search) [38]. Κατά την τεχνική αυτή πραγματοποιείται εξαντλητικός έλεγχος σε συνδυασμούς υπερπαραμέτρων με προκαθορισμένο εύρος τιμών προκειμένου να βρεθεί ο συνδυασμός που έχει την καλύτερη απόδοση.

2.3 Υλοποίηση

Για το πρακτικό κομμάτι της παρούσας μελέτης χρησιμοποιήσαμε την γλώσσα προγραμματισμού Python. Η Python είναι μια γλώσσα υψηλού επιπέδου η οποία χρησιμοποιείται συχνά σε μελέτες διαφόρων επιστημονικών πεδίων. Την επιλέξαμε κυρίως για δυο λόγος. Πρώτον, δίνει τη δυνατότητα στον προγραμματιστή να πετύχει το επιθυμητό αποτέλεσμα με λίγες γραμμές κώδικα και δεύτερον υπάρχουνε πολλές διαθέσιμες βιβλιοθήκες με έτοιμες υλοποιήσεις πολύ χρήσιμων εργαλείων που μπορούν να χρησιμοποιηθούν σαν βάση από τον προγραμματιστή ανάλογα με τον σκοπό του. Τέλος, δίνει έμφαση στην αναγνωσιμότητα του κώδικα επιβάλλοντας συγκεκριμένους τρόπους για τη δόμηση του.

Στη συνέχεια περιγράφουμε την επικοινωνία με το API του Twitter και την χρησιμοποίηση αλγορίθμων επιβλεπόμενης μάθησης μέσω της γλώσσας Python με

χρήση των ανάλογων βιβλιοθηκών.

2.3.1 Twitter API – Tweepy

Για τους σκοπούς της παρούσας μελέτης χρειάστηκε να επικοινωνήσουμε με το API που προσφέρει το Twitter[9]. Για την επικοινωνία με το API του Twitter χρησιμοποιήσαμε το Tweepy. Το Tweepy είναι μια open-source βιβλιοθήκη της Python[10] που λειτουργεί σαν wrapper για το API του Twitter παρέχοντας στον προγραμματιστή εύκολη πρόσβαση στις υπηρεσίες του. Τα βήματα που ακολουθήσαμε για την πρόσβαση στα δεδομένα των χρηστών του Twitter αναλύεται παρακάτω:

1. Το Twitter API προυποθέτει *OAuth Authentication*[11] πριν την πραγματοποίηση κάποιου request. Για το σκοπό αυτό αρχικά δημιουργήσαμε ένα Twitter application[12] και αποκτήσαμε πρόσβαση σε ένα *API Key* και ένα *API Secret* καθώς και σε ένα *access token* τα οποία χρησιμοποιήθηκαν για τη δημιουργία ενός *authentication handler* με χρήση του Tweepy:

```
auth = tweepy.OAuthHandler(API_Key, API_Secret)
auth.set_access_token(key, secret)
```

2. Στη συνέχεια δημιουργήσαμε ένα instance του Twitter API μέσω της μεθόδου *tweepy.api* δίνοντας ως παραμέτρους τον *authentication handler* που κατασκευάστηκε προηγουμένως και τις *wait_on_rate_limit = True, wait_on_rate_limit_notify = True*:

```
api = tweepy.API(auth, wait_on_rate_limit=True,
  wait_on_rate_limit_notify=True)
```

   Οι 2 τελευταίες παράμετροι σχετίζονται με το rate-limit που επιβάλλει το Twitter API για την αποστολή request. Πιο συγκεκριμένα, τα rate limits διαιρούνται σε διαστήματα 15 λεπτών με τη δυνατότητα πραγματοποίησης 15 ή 180 GET requests ανά διάστημα. Όταν ο αριθμός των requests υπερβεί το rate limit, το API επιστρέφει το response code *HTTP 429 "Too Many Requests"* και απαιτείται αναμονή πριν την αποστολή νέου request.

[9] https://developer.twitter.com/en/docs
[10] https://github.com/tweepy/tweepy
[11] https://oauth.net/articles/authentication/
[12] https://apps.twitter.com/

3. Μέσω του Twitter API instance που δημιουργήσαμε με το Tweepy καλέσαμε τη μέθοδο *api.user_timeline* για να πάρουμε τα δεδομένα κάποιου χρήστη δίνοντας ως παράμετρο το *user ID*. Η μέθοδος αυτή επιστρέφει τα 20 πιο πρόσφατα tweets για τον χρήστη με το συγκεκριμένο *user ID* (αν ο χρήστης αυτός είναι προσβάσιμος) καθώς και όλα τα δεδομένα που προέρχονται από τα πεδία του προφίλ του. Κατασκευάσαμε ένα αντικείμενο *tweepy.Cursor* δίνοντας στον constructor του αντικειμένου τη μέθοδο *api.user_timeline* και την παράμετρο της μεθόδου *user_id*. Στη συνέχεια, πήραμε για τον επιθυμητό χρήστη ένα αντικείμενο *Status.* Ένα αντικείμενο Status αντιστοιχεί σε κάθε tweet που έχει κάνει ο χρήστης και περιέχει εκτός από τις πληροφορίες του tweet και τα δεδομένα των πεδίων του προφίλ του χρήστη. Με τη συνάρτηση *process_status* απομονώσαμε μόνο τα δεδομένα που μας ενδιέφεραν για την παρούσα μελέτη και τα αποθηκεύσαμε στη βάση δεδομένων μας (περισσότερες πληροφορίες στην Ενότητα 4.1). Κατόπιν, σταματήσαμε τη διαδικασία γιατί δεν μας χρειάστηκαν επιπλέον αντικείμενα Status. Για περισσότερο έλεγχο σχετικά με τα rate limits του Twitter χρησιμοποιήσαμε τη συνάρτηση *limit_handled* περιμένοντας 15 λεπτά κάθε φορά που ξεπερνούσαμε το όριο των requests στο Twitter API. Ο σχετικός κώδικας για την απόκτηση των δεδομένων ενός χρήστη παρατίθεται παρακάτω:

```
def limit_handled(cursor):
    while True:
        try:
            yield cursor.next()
        except tweepy.RateLimitError:
            time.sleep(15 * 60)

cursor = tweepy.Cursor(api.user_timeline,user_id=user_id)
for status in limit_handled(cursor.items()):
    process_status(status)
    break
```

Για πληρότητα αναφέρουμε ότι το Twitter παρέχει και ένα Streaming API μέσω του οποίου μπορεί κάποιος να αποκτήσει πρόσβαση στα tweets που δημιουργούνται σε πραγματικό χρόνο. Το Streaming API είναι χρήσιμο για την απόκτηση μεγάλου αριθμού από tweets και οι περισσότεροι ερεύνητες το χρησιμοποιούν για να κατασκευάσουν τις βάσεις δεδομένων τους, φιλτράροντας τα αποκτηθέντα tweets ανάλογα με τις προδιαγραφές που επιθυμούν. Για παράδειγμα, μπορούν να επιλέξουν να κρατήσουν μόνο τους χρήστες που έχουν ένα συγκεκριμένο αριθμό από followers ή έχουν δημιουργήσει το προφίλ τους μετά από κάποια συγκεκριμένη

ημερομηνία. Επίσης το streaming μπορεί να περιοριστεί σε tweets που περιέχουν συγκεκριμένες λέξεις ή hashtags προκειμένου να αναλυθεί μια συγκεκριμένη μερίδα του παγκόσμιου πληθυσμού του Twitter. Το Tweepy παρέχει τις κλάσεις *StreamListener* και *Stream* οι οποίες χρησιμοποιούνται για το φιλτράρισμα των εισερχόμενων tweets και για την ανάλυση τους με σκοπό να κρατηθούν τα επιθυμητά στοιχεία.

Στην παρούσα μελέτη δεν χρησιμοποιήθηκε το Streaming API καθώς δεν κατασκευάσαμε δικιά μας βάση δεδομένων από χρήστες του Twitter (βλ. Ενότητα 4.1).

2.3.2 Υλοποιήσεις αλγορίθμων Επιβλεπόμενης Μάθησης

Για την ανίχνευση φύλου των χρηστών του Twitter, χρησιμοποιήσαμε τους ταξινομητές επιβλεπόμενης μάθησης: Naive Bayes, SVM και PNN. Παρακάτω περιγράφουμε την υλοποιήση τους με χρήση των βιβλιοθηκών της Python: Scikit-learn και Neupy.

2.3.2.1 Scikit-learn και Neupy

Οι υλοποιήσεις των αλγορίθμων Naive Bayes και Support Vector Machine πραγματοποιήθηκαν μέσω της βιβλιοθήκης Scikit-learn [6]. Το Scikit-learn είναι μια open-source (διανέμεται με BSD license) βιβλιοθήκη της Python για μηχανική μάθηση βασισμένη στις βιβλιοθήκες NumPy, SciPy, και matplotlib. Παρέχει απλά και αποδοτικά εργαλεία για εξόρυξη και ανάλυση δεδομένων με πολλές υλοποιήσεις αλγορίθμων μηχανικής μάθησης για προβλήματα ταξινόμησης (classification), ομαδοποίησης (clustering) και παλινδρόμησης (regression). Επίσης μέσω των μεθόδων που περιλαμβάνει μπορεί να γίνει εύκολη αξιολόγηση της απόδοσης των αλγορίθμων που δοκιμάζονται αλλά και να πραγματοποιηθεί προεπεξεργασία των δεδομένων εισόδου στους αλγορίθμους προκειμένου να αυξηθεί η αποδοτικότητα τους.

Χρησιμοποιήσαμε 2 ταξινομητές βασισμένους στον αλγόριθμο του Naive Bayes: τον Gaussian Naive Bayes και τον Multinomial Naive Bayes. Η υλοποίηση τους έγινε μέσω των κλάσεων *GaussianNB* και *MultinomialNB*:

```
clf1 = GaussianNB()
clf2 = MultinomialNB()
```

Για την υλοποίηση των SVMs με πυρήνα: RBF, πολυωνυμικό ή σιγμοειδή χρησιμοποιήσαμε την κλάση *SVC* η υλοποίηση της οποίας είναι βασισμένη στην στη βιβλιοθήκη *libsvm*[13]. Για SVM με γραμμικό πυρήνα κάναμε χρήση της κλάσης *LinearSVC* η οποία είναι υλοποιημένη με την βιβλιοθήκη *liblinear*[14]. Ο λόγος που χρησιμοποιήσαμε την κλάση *LinearSVC* για τον γραμμικό πυρήνα είναι ότι λόγω της υλοποιήσης με την *liblinear* κλιμακώνει καλύτερα στον μέγεθος των δειγμάτων εκπαίδευσης που χρησιμοποιήσαμε σε σχέση με την υλοποιήση με *libsvm* με την οποία η εκπαίδευση του αλγορίθμου κατανάλωνε πολύ χρόνο.

Μέσω της κλάσης *SVC* είχαμε τη δυνατότητα να επιλέξουμε τις υπερπαραμέτρους *C* και *gamma* του αλγορίθμου. Η υπερπαράμετρος C προσδιορίζει το trade-off μεταξύ της λάθος ταξινόμησης των δειγμάτων εκπαίδευσης και της απλότητας της επιφάνειας απόφασης. Μια μικρή τιμή για το C εξομαλύνει την επιφάνεια απόφασης, ενώ μια υψηλή τιμή για το C στοχεύει στην σωστή ταξινόμηση όλων των παραδειγμάτων του συνόλου εκπαίδευσης δίνοντας στο μοντέλο ελευθερία να διαλέξει περισσότερα δείγματα ως διανύσματα υποστήριξης. Η υπερπαράμετρος gamma καθορίζει πόσο μεγάλη επιρροή θα έχει ένα διάνυσμα εκπαίδευσης. Οι μικρές τιμές του gamma δίνουν 'μακρινή' ενώ οι μεγάλες 'κοντινή' επιρροή. Οι παράμετροι gamma μπορούν να θεωρηθούν ως το αντίστροφο της ακτίνας επιρροής των διανυσμάτων υποστήριξης. Επίσης είχαμε τη δυνατότητα να επιλέξουμε το βαθμό του πολυωνύμου για τον πολυωνυμικό πυρήνα μέσω της παραμέτρου *degree*.

Όσον αφορά το SVM με γραμμικό πυρήνα, μπορέσαμε μέσω της κλάσης *LinearSVC* να καθορίσουμε την συνάρτηση απώλειας διαλέγοντας την βασιζόμενη σε συντελεστή εξάρτησης (hinge) ή τη βασιζόμενη στο τετράγωνο του συντελεστή εξάρτησης (squared hinge) αλλά και την παράμετρο C.

Η κατασκευή των SVM ταξινομητών έγινε όπως φαίνεται στο παρακάτω code snippet:

```
clf2 = svm.SVC(kernel='rbf',C=C_val,gamma=gamma_val)
clf3 = svm.SVC(kernel='poly',C=C_val,gamma=gamma_val,degree=degree_val)
clf4 = svm.SVC(kernel='sigmoid',C=C_val,gamma=gamma_val)
clf5 = svm.LinearSVC(loss=loss_val,C=C_val)
```

Για την υλοποίηση των Πιθανολογικών Νευρωνικών Δικτύων χρησιμοποιήσαμε την Neupy. Η Neupy είναι μια βιβλιοθήκη της Python για Νευρωνικά Δίκτυα που

[13] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[14] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

υποστηρίζει πολλούς τύπους νευρωνικών δικτύων από απλά perceptrons μέχρι μοντέλα Βαθιάς Μάθησης (Deep Learning). Παρέχει αρκετή συμβατότητα με τη βιβλιοθήκη Scikit-learn, γεγονός που κάνει εύκολη τη χρησιμοποιήση των δυο βιβλιοθηκών ταυτόχρονα.

Για την κατασκευή ενός PNN χρησιμοποιήθηκε η κλάση *PNN* της Neupy με δυνατότητα επιλογής της παραμέτρου *τυπική απόκλιση (std)*:

```
clf6 = algorithms.PNN(std=std)
```

Για τη διαδικασία εκπαίδευσης ενός ταξινομητή χρησιμοποιούμε τη μέθοδο *fit* δίνοντας ως παραμέτρους τα διανύσματα χαρακτηριστικών και τις επιθυμητές εξόδους για κάθε δείγμα στο training set. Στη συνέχεια χρησιμοποιούμε τη μέθοδο predict για να κάνουμε πρόβλεψη για τις κλάσεις που ανήκουν τα δεδομένα του test set. Επιπλέον με χρήση της μεθόδου *predict_proba* μπορούμε να πάρουμε τις πιθανότητες τις οποίες δίνει ο ταξινομητής στο να ανήκει ένα δείγμα δεδομένων σε κάθε κλάση. Θεωρώντας ως X_train το σύνολο των διανυσμάτων εκπαίδευσης, ως Y το σύνολο των επιθυμητών εξόδων για τα διανύσματα εκπαίδευσης και ως X_test το σύνολο διανυσμάτων του test set παραθέτουμε το παρακάτω σχετικό code snippet:

```
clf.fit(X_train,Y)
clf.predict(X_test)
clf.predict_proba(X_test)
```

Αναφέρουμε ότι για την εξαγωγή πιθανοτήτων μέσω της κλάσης *SVC*, πρέπει να δοθεί η παράμετρος *probability=True.* Το πιθανοτικό μοντέλο SVM κατασκευάζεται με εφαρμογή cross-validation στο σύνολο εκπαίδευσης οπότε η κλάση που έχει την μεγαλύτερη πιθανότητα με βάση το *predict_proba* μπορεί να διαφέρει από την κλάση που προβλέπεται από το *predict.*

Όπως αναφέραμε στην Ενότητα 2.2.3.2 χρησιμοποίησαμε stratified 5-fold cross-validation για τον προσδιορισμό του accuracy ενός μοντέλου-ταξινομητή. Η διαδικασία αυτή φαίνεται παρακάτω για έναν ταξινομητή SVM με RBF πυρήνα και υπερπαραμέτρους C=100, gamma=0.001:

```
kf = StratifiedKFold(n_splits=5)
kf.get_n_splits(X,Y)
clf = svm.SVC(kernel='rbf',C=100,gamma=0.001)
acc_sum = 0


for train_index, test_index in kf.split(X,Y):
 X_train, X_test = X[train_index], X[test_index]
 y_train, y_test = Y[train_index], Y[test_index]
 clf.fit(X_train, y_train)
 predicted = clf.predict(X_test)
 acc_sum += metrics.accuracy_score(y_test, predicted)
```

Σημειώνουμε ότι τα παραπάνω ήταν μια πολύ σύντομη περιγραφή του προγραμματιστικού κομματιού της παρούσας εργασίας με μικρά ενδεικτικά δείγματα κώδικα. Το σύνολο του κώδικα είναι διαθέσιμο μέσω του ηλεκτρονικού αποθετηρίου κώδικα github.

Κεφάλαιο 3

Συναφής Βιβλιογραφία

Το πρόβλημα της ανίχνευσης του φύλου ενός ανθρώπου αποτελεί ένα πολύ ενδιαφέρον και πολυμελετημένο πεδίο έρευνας. Στο Κεφάλαιο αυτό αναφέρουμε σχετικές μελέτες που έχουν πραγματοποιηθεί προς αυτό το σκοπό. Στην Ενότητα 3.1 δίνουνε μια σφαιρική εικόνα για το πρόβλημα της ανίχνευσης φύλου. Στην Ενότητα 3.2 επικεντρωνόμαστε στις μελέτες που έχουν γίνει για την ανίχνευση του φύλου των χρηστών του Twitter.

3.1 Ανίχνευση φύλου χρηστών διαδικτυακών συστημάτων/υπηρεσιών

Πολλές μελέτες έχουν πραγματοποιηθεί επικεντρώνοντας το ενδιαφέρον τους στην ανάλυση κειμένων με σκοπό την ανίχνευση του φύλου του συγγραφέα με βάση τον τρόπο γραφής και τα γλωσσολογικά σχήματα που διαχωρίζουν τα δυο φύλα. Το πρόβλημα αυτό υπόκειται στην κατηγορία των προβλημάτων Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing-NLP).

Η μελέτη της σχέσης μεταξύ φύλου και χρήσης της γλώσσας είναι εκτενής (για μια σφαιρική εικόνα βλέπε [44], [10]). Έχουν εκδοθεί έρευνες οι οποίες υποστηρίζουν ότι αναλύοντας τα γλωσσολογικά χαρακτηριστικά που σχετίζονται με την αντρική και γυναικεία χρήση της γλώσσας, είναι δυνατό να ανιχνευθεί το φύλο ενός ατόμου ([45], [46], [47]).

Οι Koppel, Argamon, Simoni (2002) [20], χρησιμοποίησαν αυτοματοποιημένες τεχνικές κατηγοριοποίησης κειμένου και εκμεταλλευόμενοι συνδυασμούς από απλά λεξιλογικά και συντακτικά χαρακτηριστικά για την ανίχνευση του φύλου πέτυχαν περίπου 80% ακρίβεια στις προβλέψεις τους.

Σε μια μεταγενέστερη έρευνα οι Argamon, Koppel, Pennebaker, Schler (2009) [21] προσπάθησαν να συμπεράνουν το φύλο, την ηλικία, τη μητρική γλώσσα και την προσωπικότητα του συγγραφέα σε μια συλλογή από ανώνυμα κείμενα.

Χρησιμοποιώντας έναν συνδυασμό από χαρακτηριστικά εξαγόμενα από το περιεχόμενο του κειμένου αλλά και από το στυλ γραφής, πέτυχαν 76.1% ακρίβεια στην ταξινόμηση των δυο φύλων. Οι Goswami, Sarkar, Rustagi (2009) [22] εργαζόμενοι στην ίδια συλλογή κειμένων άυξησαν την ακρίβεια της ταξινόμησης σε 89.2% χρησιμοποιώντας το μέσο μήκος πρότασης, τη χρήση αργκό (slang) και τη χρήση λέξεων εκτός λεξικού.

Οι Cheng, Chandramouli, Subbalakshmi (2011) [23] χρησιμοποίησαν χαρακτηριστικά βασισμένα σε λέξεις και χαρακτήρες, χαρακτηριστικά βασισμένα στη δομή και το συντακτικό καθώς και "function words" σε μια μεγάλη συλλογή κειμένων από το Reuters News και σε ένα σύνολο από email της Enron. Πειραματίστηκαν με την χρήση Support vector machines, Bayesian logistic regression και AdaBoost decision trees για την ταξινόμηση των δυο φύλων με βάση τα παραπάνω χαρακτηριστικά. Το καλύτερο accuracy που πέτυχαν ήταν 85.1% με χρήση SVMs και τον συνδυασμό όλων των χαρακτηριστικών.

Οι Peersman, Daelemans, Van Vaerenbergh (2011) [24] επικεντρώθηκαν στο Βέλγικο κοινωνικό δίκτυο Netlog[15] επιχειρώντας να προσδιορίσουν το φύλο και την ηλικία των χρηστών σε μια συλλογή από κείμενα μηνυμάτων συνομιλίας. Τα χαρακτηριστικά που χρησιμοποίησαν για να εκπαιδεύσουν ένα Support Vector Machine ήταν: unigrams, bigrams και trigrams[16] λέξεων καθώς και bigrams, trigrams και tetragrams χαρακτήρων. Πέτυχαν 88.8% ακρίβεια στην ταξινόμηση.

Οι Αραβαντινού, Σιμάκη, Μπόρας και Μεγαλοοικονόμου (2015) [25] επιχείρησαν στο έργο τους να κάνουν ταξινόμηση κατά φύλο σε συγγραφείς διαδικτυακών blog. Χρησιμοποίησαν τους αλγορίθμους μηχανικής μάθησης Support Vector Machines, decision trees και lazy-learning αλγορίθμους όπου τους εκπαίδευσαν με στατιστικά, POS-tagging και language model χαρακτηριστικά. Με τη μέθοδο random forests πέτυχαν accuracy 70.5%. Κατέληξαν στο συμπέρασμα ότι η επιλογή των κατάλληλων γλωσσικών χαρακτηριστικών για την ανίχνευση του φύλου αυξάνει σημαντικά την ακρίβεια των προβλέψεων ανεξαρτήτως του αλγορίθμου μηχανικής μάθησης που χρησιμοποιείται.

Οι Baumann, Krasnova, Veltri και Ye (2015) [26] παρουσιάσανε μια σφαιρική εικόνα για την υπάρχουσα έρευνα που έχει διεξαχθεί για τις διαφορές των δυο φύλων στη

[15] http://nl.netlog.com/

[16] Το *n*-gram είναι μια ακολουθία *n* συνεχόμενων αντικειμένων από μια δεδομένη ακολουθία κειμένου. Τα αντικείμενα μπορεί να είναι λέξεις, χαρακτήρες, συλλαβές κ.α.

χρησιμοποίηση των microblogs.

'Αλλες μέθοδοι που έχουν χρησιμοποιηθεί για την ανίχνευση φύλου στο Διαδίκτυο περιλαμβάνουν την συγκέντρωση και επεξεργασία ονομάτων και φωτογραφιών από χρήστες κοινωνικών δικτύων.

Οι Tang, Ross, Saxena, Chen (2011) [18] επικέντρωσαν την έρευνα τους στο Facebook με σκοπό να μελετήσουν τον προσδιορισμό του φύλου με μια προσέγγιση βασισμένη στα ονόματα των χρηστών. Σαρώνοντας τα δημόσια προφίλ 1.67M χρηστών με τοποθεσία τη Νέα Υόρκη, σύλλεξαν μια ευρεία λίστα ονομάτων και το κάθε όνομα αντιστοιχήθηκε σε μια εκτίμηση δημοτικότητας και μια πιθανότητα φύλου. Χρησιμοποιώντας ένα υποσύνολο της βάσης τους που αποτελούσε το 95.1% όλων των χρηστών και περιείχε χρήστες που δεν είχαν διφορούμενο όνομα, είχαν αρκετό αριθμό φίλων και παρείχαν επιπλέον πληροφορίες στα προφίλ τους, κατάφεραν 95.2% ακρίβεια προβλέψεων με χρήση αλγορίθμων μηχανικής μάθησης. Έχοντας προσδιορίσει το φύλο των περισσότερων χρηστών της βάσης δεδομένων τους, ανέλυσαν τη διαφορετική συμπεριφορά των 2 φύλων στο Facebook. Παρατήρησαν ότι οι γυναίκες φαίνεται να είναι πιο προσεκτικές σχετικά με την ιδιωτικότητα τους σε σχέση με τους άντρες κρύβοντας σε μεγαλύτερο ποσοστό στοιχεία από το προφίλ τους που αφορούν μεταξύ άλλων το φύλο, την ηλικία και τις σεξουαλικές τους προτιμήσεις.

Οι You, Bhatia, Sun, Luo (2014) [19] μελέτησαν στο Pinterest, τη σχέση μεταξύ του φύλου ενός χρήστη και των φωτογραφιών που ανεβάζει, για ένα σύνολο 1500 χρηστών. Για κάθε χρήστη, εξετάζοντας τα "*pinboards*" που είχε δημιουργήσει, μέτρησαν το ποσοστό των φωτογραφιών του που αντιστοιχούσε σε κάθε μια από τις 33 κατηγορίες που προσέφερε το Pinterest. Επιπλέον, για να αναλύσουν το περιεχόμενο των φωτογραφιών, χρησιμοποίησαν ένα μοντέλο αναγνώρισης προτύπων το οποίο αντιπροσωπεύει κάθε εικόνα ως ένα διάνυσμα "οπτικών λέξεων". Παίρνοντας ως χαρακτηριστικά για τους χρήστες το περιεχόμενο των φωτογραφιών τους και την κατανομή αυτών ως προς τις 33 κατηγορίες, πέτυχαν 71.9% accuracy στις προβλέψεις τους για το φύλο. Παρατήρησαν ότι το περιεχόμενο των φωτογραφιών έδινε πιο χρήσιμη πληροφορία σε σχέση με τις κατηγορίες στις οποίες ανήκαν, συμπεραίνοντας ότι η ανάλυση του περιεχομένου των φωτογραφιών στα OSNs μπορεί να βοηθήσει στην ανίχνευση φύλου.

3.2 Ανίχνευση φύλου χρηστών Twitter

Ειδικότερα στο Twitter, υπάρχουν δυο προσεγγίσεις για την ανίχνευση του φύλου ενός χρήστη: i)μέσω ανάλυσης του κειμένου των tweets των χρηστών με μεθοδολογίες NLP, ii)μέσω ανάλυσης των στοιχείων που παρέχει ο χρήστης στο προφίλ του (φωτογραφία προφίλ, όνομα, χρώμα κλπ.). Η πρώτη προσέγγιση χρειάζεται μια σεβαστού μεγέθους συλλογή από tweets ενός χρήστη με την προϋπόθεση ότι ο χρήστης είναι αρκετά ενεργός ώστε να παρέχει αυτή την πληροφορία. Έχει όμως το πλεονέκτημα ότι ένας χρήστης που επιχειρεί να μην φανερώσει το φύλο του μέσω των στοιχείων του προφίλ του μπορεί να ταυτοποιηθεί από τον τρόπο γραφής του χωρίς να το θέλει. Από την άλλη, η δεύτερη προσέγγιση παρέχει έναν πιο γρήγορο και εύκολο τρόπο προσδιορισμού του φύλου αλλά έχει το προαπαιτούμενο ο χρήστης να έχει συμπεριλάβει ηθελημένα στο προφίλ του, στοιχεία που βοηθούν σε αυτό τον προσδιορισμό.

Η πρώτη μελέτη για τον προσδιορισμό του φύλου στους χρήστες του Twitter έγινε από τους Rao, Yarowsky, Shreevats και Gupta (2010) [1]. Εκτός από το φύλο, οι συγγραφείς επιχείρησαν να συμπεράνουν και την ηλικία, την καταγωγή και τις πολιτικές πεποιθήσεις των χρηστών. Για τη μελέτη του φύλου κατασκεύασαν ένα ground truth dataset, χρησιμοποιώντας σαν πηγές αντρικές και γυναικείες πανεπιστημιακές αδελφοτήτες καθώς και αντρικά και γυναικεία προϊόντα υγιεινής. Κατέληξαν με ένα σύνολο ταυτοποιημένων χρηστών που περιείχε περίπου 500 άντρες και 500 γυναίκες. Αρχικά, εξέτασαν σαν πιθανά χαρακτηριστικά τη δομή του δικτύου ενός χρήστη (followers, following) και τη συμπεριφορά του στην επικοινωνία (respone/retweet/tweet frequency) και κατέληξαν ότι δεν υπάρχει κάποια αξιοσημείωτη διαφορά ανάμεσα στα δυο φύλα για αυτά τα δυο χαρακτηριστικά. Τα συμπεράσματα σχετικά με την μη διαφοροποίηση των δυο φύλων στην επικοινωνιακή συμπεριφορά έρχονται σε αντίθεση με τα αποτελέσματα παλιότερης έρευνας πάνω σε τηλεφωνικές συνομιλίες [27]. Στην συνέχεια πειραματίστηκαν με τρια SVM μοντέλα. Το πρώτο έπαιρνε σαν είσοδο τα κοινωνιογλωσσικά (sociolinguistic) χαρακτηριστικά του τρόπου γραφής των χρηστών, το δεύτερο το περιεχόμενο του κειμένου των tweets (unigrams,bigrams) και το τρίτο, τις προβλέψεις των δυο προηγούμενων μαζί με τα αντίστοιχα βάρη τους. Πέτυχαν 71.8% ακρίβεια στην ταξινόμηση βάσει φύλου των χρηστών χρησιμοποιώντας τα κοινωνιογλωσσικά χαρακτηριστικά, 68.7% ακρίβεια με τα *n*-grams και 72.3% ακρίβεια με το συνδυαστικό μοντέλο. Τελος, παρατήρησαν ότι η χρήση emoticons, αποσιωπητικών και η επανάληψη ίδιων γραμμάτων σε μια λέξη

για έμφαση, υποδεικνύουν ότι ο χρήστης μάλλον είναι γυναίκα και ότι οι λέξεις που ακολουθούν την κτητική ανωνυμία “my” είναι πολύτιμα στοιχεία για τον προσδιορισμό του φύλου.

Οι Burger, Henderson, Kim και Zarrella (2011) [11] στην μελέτη τους χρησιμοποίησαν μια τεράστια βάση δεδομένων συγκριτικά με την δουλειά των [1] που χρησιμοποίησαν μόλις 1000 χρήστες. Συγκέντρωσαν μια συλλογή με 213M tweets από 18.3M χρήστες του Twitter. Από αυτό το σύνολο ταυτοποίησαν το φύλο 184k χρηστών που είχαν δημοσιεύσει 4.1M tweets. Η ταυτοποίηση έγινε μέσω εξωτερικών blog sites που οι χρήστες είχαν παραθέσει στο πεδίο w*ebsite* του προφίλ τους, τα οποία περιείχαν το φύλο των χρηστών. Επαλήθευσαν χειροκίνητα το φύλο 1000 τυχαίων χρηστών για να βεβαιωθούν για την αξιοπιστία της βάσης τους, κοιτώντας το *bio* στο προφίλ των χρηστών για προφανή στοιχεία που μαρτυρούσαν το φύλο τους. Μόνο 150 χρήστες (15% του δείγματος) είχαν στα *bios* τους ανάλογα στοιχεία. 136 χρήστες από αυτούς είχαν προφίλ σε εξωτερικό blog site με πεδίο επιλεγμένου φύλου, και σε όλους το φύλο αυτό ταυτιζόταν με αυτό που εξήχθει από το *bio.* Συνεπώς κατέληξαν ότι τα δεδομένα τους ήταν υψηλής ποιότητας. Για τα πειράματά τους χώρισαν το σύνολο των ταυτοποιημένων χρηστών σε υποσύνολα training, development και testing. Το σύνολο χρηστών περιείχε 55% γυναίκες και 45% άντρες το οποίο συμβαδίζει με το ποσοστό των φύλων στο Twitter στο οποίο κατέληξε μια παλιότερη έρευνα [28]. Επίσης, το σύνολο είχε γλωσσική ποικιλομορφία με το 66.7% των χρηστών να δημοσιεύσουν tweets στα Αγγλικά ενώ στους υπόλοιπους χρήστες οι επικρατέστερες γλώσσες ήταν τα Πορτογαλικά με 14.4% και τα Ισπανικά με 6%. Η αναλογία αυτή διαφέρει με την έρευνα του Wauters (2010) [29] που ανέφερε ως τις τρεις επικρατέστερες γλώσσες των tweets τα Αγγλικα με 50%, τα Ιαπωνικά με 14% και τα Πορτογαλικά με 9%. Η προσέγγιση τους για την ανίχνευση του φύλου ήταν ανεξάρτητη της γλώσσας. Τα χαρακτηριστικά που χρησιμοποίησαν ήταν *n*-grams λέξεων και γραμμάτων προερχόμενα από τα tweets, το *display name*, το *username* και το *bio*. Τα χαρακτηριστικά ήταν δυαδικής λογικής δηλώνοντας ύπαρξη ή απουσία ενός *n*-gram σε κάθε περίπτωση. Δεν έλαβαν υπόψην *n*-grams που εμφανίζονταν σε λιγότερους από 3 χρήστες. Για την εξαγωγή συμπερασμάτων χρησιμοποίησαν Support Vector Machines, Naive Bayes και Balanced Winnow2 αλγορίθμους τροφοδοτώντας τους με συνδυασμούς των παραπάνω χαρακτηριστικών. Την καλύτερη απόδοση πέτυχε μια δική τους υλοποίηση του Winnow αλγορίθμου η οποία μείωσε σημαντικά τις απαιτήσεις σε μνήμη για την εκπαίδευση. Για να επιλέξουν τις παραμέτρους μάθησης του αλγορίθμου πραγματοποίησαν αναζήτηση

πλέγματος. Η ακρίβεια του αλγορίθμου αυξανόταν με τον αριθμό των tweets του κάθε χρήστη που λάμβαναν υπόψην αλλά και με τον συνολικό αριθμό των χρηστών που χρησιμοποιούταν για την εκπαίδευση. Συμπέραναν ότι το περιεχόμενο των tweets περιέχει περισσότερη gender-specific πληροφορία σε σχέση με το *bio* του χρήστη. Παρατήρησαν επίσης ότι τα πιο gender-specific χαρακτηριστικά από κείμενα tweet ανήκαν στον γυναικείο πληθυσμό. Μερικά από αυτά τα χαρακτηριστικά ήταν emoticons σε συμφωνία με τις παρατηρήσεις των [1]. Χρησιμοποιώντας χαρακτηριστικά εξαγόμενα μόνο από το *display name* πέτυχαν 89.1% ακρίβεια ενώ με τον συνδυασμό των υπόλοιπων τριών είχαν 84.3% ακρίβεια, πράγμα που δείχνει την πολυτιμότητα του *display name* για τον προσδιορισμό του φύλου ενός χρήστη. Με τον συνδυασμό *username* και tweets ο οποίος αντιπροσωπεύει μια συνήθη περίπτωση πληροφορίας σε διαφορετικά κοινωνικά δίκτυα και chat rooms, πέτυχαν 81.4% ακρίβεια. Τέλος, εξάγοντας *n*-grams και από τα τέσσερα πεδία πέτυχαν ανίχνευση φύλου με 92% ακρίβεια. Για να συγκρίνουν την απόδοση του ταξινομητή τους με την αντίστοιχη απόδοση για τον προσδιορισμό φύλου από ανθρώπους, χρησιμοποιήσαν το Amazon Mechanical Turk (AMT)[17]. Κάθε εργάτης μελετούσε όλα τα tweets ενός χρήστη και καλούταν να αποφανθεί για τον φύλο του. Για κάθε χρήστη αποφάσιζαν πέντε εργάτες του AMT. Για την τελική απόφαση πειραματίστηκαν παίρνοντας είτε την γνώμη της πλειοψηφίας πετυχαίνοντας 65.7% ακρίβεια είτε χρησιμοποιώντας έναν αλγόριθμο μεγιστοποίησης προσδοκίας (EM) πετυχαίνοντας 67.3% ακρίβεια. Η απόδοση του Winnow αλγορίθμου με εκμετάλλευση όλων των tweets ήταν σημαντικά καλύτερη με 75.5% ακρίβεια.

Οι Al Zamal, Liu, Ruths (2012) [12] χρησιμοποίησαν την αρχή της ομοφυλίας για να προσδιορίσουν το φύλο, την ηλικία και τις πολιτικές πεποιθήσεις (Δημοκρατικός/Ρεπουμπλικάνος) 400 χρηστών του Twitter. Ως ομοφυλία ορίζεται η τάση των ανθρώπων να σχετίζονται και να σχηματίζουν δεσμούς με όμοιους τους[18]. Σύμφωνα με παλιότερες έρευνες αυτή η τάση υπάρχει και στα κοινωνικά δίκτυα [30]. Το φύλο των χρηστών ταυτοποιήθηκε με βάση το όνομα που είχε βάλει ο κάθε χρήστης στο προφίλ του. Προϋπόθεση ήταν το όνομα να είναι ανάμεσα στα 100 πιο κοινά ονόματα για μωρά γεννημένα στις ΗΠΑ, σύμφωνα με το U.S. Social Security Administration. Συγκέντρωσαν τα τελευταία 1000 tweets από τους ταυτοποιημένους χρήστες αλλά και από όλους τους χρήστες που αυτοί ακολουθούσαν. Τα χαρακτηριστικά που χρησιμοποίησαν ήταν k-top λέξεις (οι k πιο

[17] https://www.mturk.com/
[18] https://en.wikipedia.org/wiki/Homophily

διαφοροποιητικές λέξεις που χρησιμοποιούνταν από κάθε φύλο), k-top stems[19], k-top bigrams/trigrams, k-top hashtags, k-top co-stems, τάση για retweeting, στατιστικά στοιχεία συχνότητας (tweets, mentions, hashtags, links, και retweets ανα μέρα) και αναλογία followers/following. Πραγματοποίησαν τα πειράματα τους χρησιμοποιώντας σαν ταξινομητή ένα Support Vector Machine και πραγματοποιώντας 10-fold cross-validation για την εξαγωγή της ακρίβειας των προβλέψεων. Για το φύλο χρησιμοποιώντας χαρακτηριστικά και από τους "γείτονες" του κάθε χρήστη πέτυχαν 80.2% ακρίβεια, ενώ μόνο με δεδομένα από το προφίλ του χρήστη πέτυχαν ακρίβεια 79.5%.

Οι Bamman, Eisenstein, Schnoebelen (2012) [31] μελέτησαν τη σχέση μεταξύ φύλου, γλωσσικού στυλ και σχέσεων στα κοινωνικά δίκτυα. Για τους χρήστες που επεξεργάστηκαν υπήρχε προαπαιτούμενο να έχουν χρησιμοποίησει τουλάχιστον 50 από τις 1000 πιο κοινές λέξεις στα Αγγλικά. Η ταυτοποίηση του φύλου των χρηστών έγινε με βάση τα στοιχεία απογραφής από το US Social Security administration χωρίς να αναφέρεται κάποια επαλήθευση. Η τελική βάση περιείχε περίπου 14.4k χρήστες και 9.2M tweets. Η μελέτη τους διαχωρίζεται από προηγούμενες παρόμοιες δουλειές σε δυο βασικά σημεία.
Πρώτον, παίρνουν υπόψην τους θεωρητικά επιχειρήματα και ποιοτικές αποδείξεις ότι το φύλο μπορεί να εκφραστεί μέσα από μια ποικιλία εκφράσεων και συμπεριφορών. Ομαδοποιώντας τους συγγραφείς των tweets, αναγνώρισαν ένα εύρος από διαφορετικά στυλ. Πολλές από αυτές τις ομάδες είχαν ισχυρά στοιχεία φύλου, αλλά ο τρόπος που χρησιμοποιούσαν την γλώσσα αντικρουόταν με τα συγκεντρωτικά στατιστικά για τη σχέση γλώσσας-φύλου. Αυτό δείχνει ότι υπάρχουν γλωσσικά στυλ που σχετίζονται στενά με κάποιο φύλο αλλά έρχονται σε αντίθεση με τις γλωσσικές τάσεις που είχαν προηγούμενως αποδοθεί σε άντρες ή γυναίκες ως αδιαφοροποιήτες κοινωνικές ομάδες. Δεύτερον, για να μελετήσουν τα άτομα που διαφοροποιούνταν από τα γενικά μοτίβα γλώσσας-φύλου, φτιάξανε έναν logistic regression ταξινομητή φύλου με unigrams λέξεων ως χαρακτηριστικά, ο οποίος είχε 88% ακρίβεια με 10-fold cross-validation. Παρατήρησαν ότι οι χρήστες που τους αποδόθηκε λάθος φύλο από τον ταξινομητή, είχαν κοινωνικές συνδέσεις με πολύ λιγότερη ομοφυλία σε σχέση με τους υπόλοιπους. Κατέληξαν ότι η ομοφυλία και η χρήση mainstream γλωσσικών στοιχείων σχετικών με το φύλο είναι στενά συνδεδεμένες πράγμα που παρουσιάζει ένα βασικό πρόβλημα στη σχέση του ατόμου με τις κοινότυπες νόρμες και ρόλους των φύλων. Για το λόγο αυτό, επιλέγουν να δουν τους χρήστες ως άτομα που πράττουν ως κάποιο φύλο, με έναν

[19] https://en.wikipedia.org/wiki/Word_stem

τρόπο που επηρεάζει τις γλωσσικές τους επιλογές και την κοινωνική τους συμπεριφορά.

Οι Deitrick, Miller, Valyou, Dickinson, Munson και Hu (2012) [32] πρότειναν την χρήση νευρωνικών δικτύων για τον προσδιορισμό του φύλου. Το dataset τους περιείχε 3031 tweets επισημασμένα (labelled) με το φύλο του χρήστη. Πειραματίστηκαν με Balanced Winnow και Modified Balanced Winnow μοντέλα. Χρησιμοποιώντας το Modified Balanced Winnow μοντέλο με 53 επιλεγμένα *n*-gram χαρακτηριστικά πέτυχαν 98.5% ακρίβεια. Σε μετέπειτα μελέτη οι Miller, Dickinson και Hu (2012) [33] συγκέντρωσαν 3000 tweets επισημασμένα με φύλο και έκαναν χρήση stream αλγορίθμων με Percepton και Naive Bayes με *n*-grams λέξεων και χαρακτήρων. Ανέφεραν 99.3% accuracy με την ταξινόμηση των Perceptron όταν το μήκος των tweets ήταν τουλάχιστον 75 λέξεις.

Οι Liu & Ruths (2013) [3] στην δουλειά τους συμπεριέλαβαν το όνομα του χρήστη για τον προσδιορισμό του φύλου με τρόπο που όπως αναφέρουν δεν είχε ξαναγίνει σε προηγούμενες μελέτες. Πιο συγκεκριμένα χρησιμοποίησαν ένα gender-association score σύμφωνα με την φόρμουλα $\left(M\left(x\right)-F\left(x\right)\right)/\left(M\left(x\right)+F\left(x\right)\right)$, όπου το M(x) αντιπροσωπεύει πόσες φορές εμφανίζεται το όνομα x σε άντρες σύμφωνα με τα δεδομένα από το US Census του 1990 και το F(x) αντίστοιχα σε γυναίκες. Η βάση πάνω στην οποία πραγματοποίησαν τα πειράματά τους κατασκευάστηκε από εργάτες του AMT οι οποίοι σε τριάδες παρατήρησαν τις φωτογραφίες προφίλ από χρήστες του Twitter και τους κατέταξαν σε άντρες, γυναίκες ή άγνωστο. Έγιναν δεκτοί μόνο οι χρήστες για τους οποίους και οι τρεις εργάτες συμφώνησαν για το φύλο τους. Η διαδικασία αυτή έγινε σε 50k τυχαίους χρήστες οι οποίοι είχαν δημοσιεύσει τουλάχιστον 1000 tweets συνολικά. Από αυτούς περίπου το 25% γεγονός το οποίο τους οδήγησε στην παρατήρηση ότι περίπου 1 στους 4 χρήστες του Twitter έχει φωτογραφία χαρακτηριστική του φύλου του. Το τελικό dataset το οποίο κάνανε δημόσια διαθέσιμο περιείχε 4449 άντρες και 8232 γυναίκες. Το ποσοστό 35% των αντρών ήρθε σε αντίθεση με την έρευνα των [34] όπου είχε εκτιμήσει ότι οι άντρες αποτελούν το 45% των χρηστών του Twitter. Από αυτό εξήχθει το συμπέρασμα ότι μάλλον οι γυναίκες είναι πιο πιθανόν να έχουν φωτογραφία που προδίδει το φύλο τους. Αναφέρουν ότι η βάση τους είναι πιο αντιπροσωπευτική του παγκόσμιου πληθυσμού του Twitter από αυτή των [11], και ότι επειδή ο τρόπος κατασκευής της βάσης δεν σχετίζεται με τα χαρακτηριστικά που χρησιμοποίησαν για τα πειράματά τους υπάρχει μεγαλύτερη αντικειμενικότητα στα αποτελέσματά τους. Για τα πειράματά τους χρησιμοποίησαν 4000 χρήστες από

κάθε φύλο και 1000 tweets για τον καθένα. Δοκίμασαν τρεις classifiers: έναν baseline SVM classifier με χαρακτηριστικά αυτά που χρησιμοποιήθηκαν και στη μελέτη [12] πετυχαίνοντας 83.3% ακρίβεια, έναν integrated SVM classifier στον οποίο προστέθηκε το gender-name association score σαν χαρακτηριστικό με 85.2% ακρίβεια και έναν threshold classifier ο οποίος επέλεγε για την ταξινόμηση το gender-name association score αν ήταν μεγαλύτερο από το επιλεγμένο threshold, ενώ σε αντίθετη περίπτωση τον integrated classifier, καταλήγοντας σε 87.1% ακρίβεια για threshold=0.85.

Οι Alowibdi, Buy και Yu (2013) [2] στην μελέτη τους χρησιμοποίησαν χαρακτηριστικά βασισμένα στα 5 πεδία επιλογής χρωμάτων τα οποία παρείχε το Twitter σε κάθε χρήστη για να διαμορφώσει το προφίλ του. Πιο συγκεκριμένα, οι 5 επιλογές αφορούσαν τα *background color, text color, sidebar border color, link color, sidebar fill color.* Όπως αναφέρουν, η προσέγγιση αυτή έχει το πλεονεκτήμα ότι σε αντίθεση με μεθόδους που ακολουθήθηκαν σε παρόμοιες μελέτες είναι ανεξάρτητη της γλώσσας. Επιπλέον, δεν χρησιμοποιείται χώρος υψηλών διαστάσεων για τη εκπαίδευση των αλγορίθμων μηχανικής μάθησης, όπως γίνεται όταν επιλέγονται ως χαρακτηριστικά *n*-grams εξαγόμενα από tweets, ονόματα και βιογραφικά. Αυτό μειώνει την υπολογιστική πολυπλοκότητα των πειραμάτων και προσφέρει έναν αποδοτικό και scalable τρόπο ανίχνευσης του φύλου των χρηστών. Για την κατασκευή της βάσης χρηστών, συγκέντρωσαν Twitter profiles σε διάστημα δυο μηνών και κράτησαν μόνο τους χρήστες για τους οποίους μπόρεσαν να ταυτοποιήσουν το φύλο τους από κάποια εξωτερική πηγή με βάση τα URLs που είχαν παραθέσει στα στοιχεία τους. Η τελική βάση περιείχε περίπου 53k χρήστες χωρισμένους σε 31k άντρες και 22k γυναίκες. Έκαναν δοκιμές με 4 σύνολα χρηστών: το T1 περιείχε όλους τους χρήστες, στο T2 αφαιρέθηκαν όσοι είχαν στο προφίλ τους το default design του Twitter, στο T3 αφαιρέθηκαν οι χρήστες που είχαν επιλέξει ένα από τα 19 predefined designs (συμπεριλαμβανομένου του default) και στο T4 αφαιρέθηκαν όσοι είχαν κάποιο predefined design ή είχαν επιλέξει ως background color το άσπρο ή το μαύρο. Από αυτό το διαχωρισμό των χρηστών βγήκε το συμπέρασμα ότι οι γυναίκες είναι πιο πιθανό να διαλέξουν δικά τους χρώματα ενώ οι άντρες τείνουν να επιλέξουν τα προκαθορισμένα. Για τα πειράματά τους δοκίμασαν μια τεχνική color reduction & quantization. Μειώνοντας την αναπαράσταση των Red, Green & Blue τιμών ενός χρώματος από 8 bit σε 3 bit, ελάττωσαν τον αριθμό των διαφορετικών χρωμάτων από 16M σε 512. Στη συνέχεια ταξινόμησαν τα χρώματα και έδωσαν σε όμοια χρώματα (γειτονικά κατά την ταξινόμηση) διαδοχικές ακέραιες τιμές τις οποίες εισήγαγαν στον ταξινομητή.

Οι ταξινομητές που δοκιμάστηκαν ήταν οι Naive Bayes, Decision Tree, Probabilistic Neural Network και Naïve Bayes/Decision-Tree Hybrid. Πραγματοποίησαν 10-fold cross validation στα 4 σύνολα χρηστών για κάθε ταξινομητή και συγκρίνανε τα αποτελέσματα σχετικά με τον αριθμό των χρωμάτων (1 έως 5) που χρησιμοποιήσανε για την ταξινόμηση και με το αν χρησιμοποιήσανε προεπεξεργασία των χρωμάτων. Πραγματοποιώντας πειράματα στο σύνολο T3, παρατήρησαν ότι χωρίς color quantization & sorting, η χρησιμοποίηση τριών χρωμάτων καταλήγει σε περίπου ίδια απόδοση με τη χρησιμοποίηση τεσσάρων ή πέντε χρωμάτων. Αντίθετα, έχοντας πραγματοποιήσει προεπεξεργασία των χρωμάτων, η ακρίβεια προβλέψεων αυξάνεται με τη χρησιμοποίηση όλων των χρωμάτων. Από το σύνολο των αποτελεσμάτων για την ανίχνεση φύλου των χρηστών του T3, κατέληξαν ότι το color quantization & sorting καταλήγει σε σημαντική αύξηση της απόδοσης. Το καλύτερο accuracy που πετύχανε ήταν 71.6% χρησιμοποιώντας PNN σε ολόκληρη την βάση τους (T1) με color quantization & sorting. Τέλος, ερευνήσανε την επίδραση των διαφορετικών μεγεθών συνόλου δεδομένων εκπαίδευσης, καταλήγοντας ότι δεν θα υπήρχε σημαντικό κέρδος με μεγαλύτερο σύνολο δεδομένων.

Οι συγγραφείς της μελέτης [2], σε μετέπειτα έρευνα τους το 2014 [5] επιχείρησαν να ανιχνεύσουν προφίλ του Twitter που περιέχουν παραπλανητικές πληροφορίες σχετικά με το φύλο του χρήστη. Όπως αναφέρεται είναι η πρώτη έρευνα που έγινε για την ανίχνευση απάτης βασισμένη στην εύρεση αντικρουόμενων πληροφοριών στο προφίλ ενός χρήστη στα OSNs. Για το σκοπό αυτό συλλέξανε ένα σύνολο δεδομένων με 174k χρήστες, το φύλο των οποίων ταυτοποιήθηκε μέσω του Facebook profile τους το οποίο παρείχανε στα στοιχεία τους. Χρησιμοποιώντας έναν Bayesian ταξινομητή με χαρακτηριστικά από τα *display name, username, background color, text color, sidebar border color, link color, sidebar fill color* χώρισαν τους χρήστες σε 5 κατηγορίες: έντονα αντρικά στοιχεία, έντονα γυναικεία στοιχεία, ασθενή αντρικά στοιχεία, ασθενή γυναικεία στοιχεία και ουδέτερο. Στη συνέχεια επιχείρησαν να ανιχνεύσουν χρήστες του Twitter που τα χαρακτηριστικά του προφίλ τους έρχονταν σε σύγκρουση με το φύλο που είχαν επιλέξει στο Facebook. Τα προφίλ που εμφάνιζαν ισχυρά στοιχεία φύλου το οποίο ερχότανε σε αντίθεση με το δηλωμένο φύλο στο Facebook κρίθηκαν ως “πιθανά παραπλανητικά”, ενώ τα προφίλ που εμφάνιζαν ασθενή στοιχεία φύλου αντίθετου με του Facebook κρίθηκαν ως “ενδεχομένως παραπλανητικά”. Με χειροκίνητη επιθεώρηση βρήκαν ότι το 42.8% των “πιθανών παραπλανητικών” και το 8.7% ενός στατιστικά-σημαντικού δείγματος των “ενδεχομένως παραπλανητικών” ήταν

όντως παραπλανητικά. Σημειώνεται ότι έγινε προεπεξεργασία των χρωμάτων με την τεχνική color quantization & sorting που χρησιμοποιήθηκε και στο [2], ενώ στα *display names* και *usernames* έγινε φωνητική ανάλυση μετατρέποντας τα σε ακολουθίες φωνημάτων ανεξάρτητες της γλώσσας. Με αυτό τον τρόπο η προσέγγιση ήταν εξ ολοκλήρου ανεξάρτητη της γλώσσας του κάθε Twitter προφίλ που εξετάστηκε.

Οι Nguyen, Trieschnigg, Doğruöz, Gravel, Theune, Meder (2014) [8] υπέδειξαν τη δυσκολία της ανίχνευσης του φύλου και της ηλικίας από τα κείμενα των tweets. Για να υποστηρίξουν τη θέση τους πραγματοποίησαν ένα crowdsourcing πείραμα με τη μορφή online παιχνιδιού βάζοντας τους παίχτες να μαντέψουν το φύλο και την ηλικία ενός συνόλου χρηστών του Twitter διαβάζοντας έναν αριθμό από τα tweets τους. Οι προβλέψεις των παιχτών είχαν 84% ακρίβεια όταν πάρθηκε σαν απόφαση για το φύλο κάθε χρήστη η άποψη της πλειοψηφίας. Μεμονωμένα, οι παίχτες που είχαν κάνει 7 ή παραπάνω συνολικά προβλέψεις είχαν 71% ακρίβεια κατά μέσο όρο. Παρατήρησαν ότι οι γυναίκες μεγαλύτερης ηλικίας έτειναν να μην εμφανίζουν στοιχεία στα κείμενά τους που να τονίζουν το φύλο τους (σύμφωνα με τις εκτιμήσεις των παιχτών). Παράλληλα δοκίμασαν ένα αυτόματο σύστημα για τον προσδιορισμό του φύλου. Πιο συγκεκριμένα εφάρμοσαν τον αλγόριθμο logistic regression χρησιμοποιώντας unigram features από τα κείμενα πετυχαίνοντας 69% ακρίβεια. Το χαμηλό accuracy του αυτόματου συστήματος αποδόθηκε στο γεγονός ότι χρησιμοποιήθηκαν λίγα tweets για κάθε χρήστη (20-40). Το πείραμα έγινε σε μια βάση 3000 Ολλανδών χρηστών του Twitter. Οι 200 από αυτούς επιλέχθηκαν τυχαία ως test set για το online παιχνίδι και τη μέτρηση της απόδοσης του αυτόματου συστήματος. Τονίστηκε ότι ο τρόπος κατασκευής της βάσης χρηστών στην οποία πραγματοποιούνται τα πειράματα είναι υψίστης σημασίας με σκοπό να είναι όσο το δυνατόν πιο αντιπροσωπευτική του γενικότερου πληθυσμού. Χαρακτήρισαν τη βάση της μελέτης [1] η οποία περιείχε μέλη από αδελφότητες πανεπιστημίων μεροληπτική, επειδή οι συγκεκριμένοι χρήστες είναι πιο πιθανό να δείχνουν μια ισχυρή ταυτότητα φύλου. Τα γενικά συμπεράσματα που εξήγαγαν οι συγγραφείς είναι ότι η συνήθης προσέγγιση με αλγορίθμους επιβλεπόμενης μάθησης, μπορεί να προβλέπει το φύλο για τους περισσότερους χρήστες αλλά το κάνει μαθαίνοντας μια στερεοτυπική συμπεριφορά και παρέχοντας μια απλοική ματία για τον ορισμό του φύλου. Για αυτό το λόγο, πρότειναν το φύλο να αντιμετωπίζεται σε παρόμοιες έρευνες ως κοινωνική μεταβλητή και όχι ως μια στατική βιολογική μεταβλητή.

Οι Vicente, Batista, Carvalho (2015) [4] διερεύνησαν μια μέθοδο ανίχνευσης του φύλου με 192 χαρακτηριστικά βασισμένα στα ονόματα εξαγόμενα από το *username* και το *display name* των χρηστών (για παράδειγμα δυο από τα χαρακτηριστικά δυαδικής λογικής που χρησιμοποιήθηκαν ήταν i)υπάρχει κάποιο όνομα μέσα στο πεδίο, ii)υπάρχει κάποιο όνομα και βρίσκεται στην αρχή). Για να συσχετίσουν τα ονόματα που βρέθηκαν στο προφίλ των χρηστών με το αντίστοιχο φύλο, χρησιμοποιήσαν ένα λεξικό 8k ονομάτων βασισμένο στα δεδομένα του Social Security Administration των ΗΠΑ στο οποίο κρατήθηκαν μόνο τα ονόματα που ήταν αποκλειστικά αντρικά ή γυναικεία και επιπλέον είχαν τουλάχιστον 1000 εγγραφές. Για την εξαγωγή των χαρακτηριστικών, έγινε όπου ήταν απαραίτητο επεξεργασία των δυο πεδίων για την αφαίρεση των επαναλαμβανόμενων φωνηέντων και την αντικατάσταση "leet speak" χαρακτήρων με τους ισοδύναμους τους (π.χ. $3 \rightarrow e$). Συλλέξανε tweets μέσω του Twitter streaming API και περιορίσανε τα δεδομένα τους σε Αγγλικά tweets που είχαν γεωγραφική προέλευση τις ΗΠΑ ή την Αγγλία. Οι χρήστες περιορίστηκαν περαιτέρω σε αυτούς που οι πληροφορίες του προφίλ τους κάνανε match με τουλάχιστον ένα από τα χαρακτηριστικά φύλου. Αυτοί που υπήρχε υποψία ότι μπορεί να είναι bot αφαιρέθηκαν. Ένα τυχαίο δείγμα χρηστών από την βάση επιλέχθηκε για ανίχνευση φύλου με εξέταση των *username* και *display name*, παρατήρηση της φωτογραφίας προφίλ και ανάλυση εξωτερικών blog sites συνδεδεμένων με το προφίλ τους. Με αυτή τη διαδικασία δημιουργήθηκε ένα υποσύνολο 748 χειροκίνητα επισημασμένων χρηστών. Με βάση αυτό το υποσύνολο, παρατήρησαν ότι οι γυναίκες ήταν πιο πιθανό να χρησιμοποιήσουν επανάληψη φωνήεντων και "leet speak" και ότι τα περισσότερα χαρακτηριστικά εξήχθησαν από το *display name* (63% από το *display name* και 37% από το *username*). Διεξήγαγαν πειράματα με αλγόριθμους επιβλεπόμενης μάθησης στο υποσύνολο των 748 ταυτοποιημένων χρηστών χρησιμοποιώντας παραλλαγές του Naive Bayes, Logistic Regression και Support Vector Machines. Την καλύτερη ακρίβεια στην ταξινόμηση βάσει φύλου πέτυχε το Multinomial Naive Bayes με 97.2% χρησιμοποιώντας 5-fold cross-validation. Δοκιμάστηκαν επίσης δυο προσεγγίσεις μη επιβλεπόμενης μάθησης χρησιμοποιώντας τους αλγορίθμους fuzzy c-Means και K-means. Στην πρώτη προσέγγιση χρησιμοποιήθηκε μόνο το επισημασμένο υποσύνολο για training και evaluation ενώ στη δεύτερη χρησιμοποιήθηκαν όλοι οι μη επισημασμένοι χρήστες για το training και οι επισημασμένοι για το evaluation. Η καλύτερη απόδοση επιτεύχθηκε από το fuzzy c-means κατά την δεύτερη προσέγγιση με 96% ακρίβεια. Με βάση αυτό το αποτέλεσμα συμπεράθηκε ότι ο fuzzy c-means αλγόριθμος είναι μια ιδανική επιλογή για ανίχνευση φύλου στο Twitter καθώς δεν χρειάζεται χρήστες επισημασμένους με το φύλο τους, αυξάνει

την απόδοση του με την ύπαρξη περισσότερων δεδομένων και πετυχαίνει ακρίβεια πολύ παρόμοια με την καλύτερη επιβλεπόμενη μέθοδο (διαφορά ακρίβειας 1%).

Οι An, Weber (2016) [7] ερεύνησαν την χρησιμοποίηση των hashtags από χρήστες του Twitter σε σχέση με τα δημογραφικά τους στοιχεία. Επικεντρώθηκαν στους χρήστες που ζούσαν στη Νέα Υόρκη χρησιμοποιώντας μια online υπηρεσία αναζήτησης και φιλτράροντας τα αποτελέσματα με την απαίτηση το πεδίο *location* των χρηστών να περιέχει λέξεις-κλειδιά όπως “ny”, “nyc”, “brooklyn”. Από το προκύπτον σύνολο χρηστών κρατήθηκαν αυτοί οι οποίοι ήταν ενεργοί με περισσότερα από 10 tweets στο ιστορικό τους και ήταν εγγεγραμένοι στο Twitter για περισσότερους από 3 μήνες. Για να ταυτοποιήσουν το φύλο, την ηλικία και την εθνικότητα των ενεργών χρηστών χρησιμοποίησαν το Face++ το οποίο κατάφερε να ανιχνεύσει τα δημογραφικά στοιχεία για το 49% των χρηστών. Για το ταυτοποιημένο σύνολο ενεργών χρηστών που περιείχε περίπου 346k χρήστες, σύλλεξαν τα tweets που είχανε δημοσιεύσει σε διάστημα ενός χρόνου με σκοπό την ανάλυση των hashtags που χρησιμοποιήθηκαν. Το συμπέρασμα που έβγαλαν είναι ότι τα πιο δημοφιλή hashtags είναι σε μεγάλο βαθμό παρόμοια για όλα τις δημογραφικές ομάδες, γεγονός που δείχνει ότι μια ανάλυση σχετικά με τη δημοτικότητα των hashtags και των τάσεων στο Twitter είναι πιθανό να μη συμπεριλάβει συμπεριφορές σχετικές με τα δημογραφικά στοιχεία του κάθε χρήστη. Θεωρούμε ότι αυτό το συμπέρασμα αφορά και την περίπτωση του φύλου των χρηστών ξεχωριστά, καθώς δεν αναφέρεται κάπου ρητά.

Σε μια πρόσφατη μελέτη, οι Cesare, Grant και Nsoesie (2017) [9] παρουσιάζουν μια ανασκόπηση των υπαρχουσών προσεγγίσεων για την αυτόματη ανίχνευση των δημογραφικών χαρακτηριστικών (μεταξύ των οποίων και του φύλου), για χρήστες κοινωνικών δικτύων. Συγκέντρωσαν σχετικές έρευνες από το Google Scholar οι οποίες είχαν έτος δημοσίευσης μεταξύ από το 2006 έως το 2016. Αναφέρουν τις μεγαλύτερες προκλήσεις που αντιμετωπίζουν αυτές οι έρευνες και πιο συγκεκριμένα την απόκτηση ground truth δεδομένων και την σύγκριση μεταξύ τους λόγω των διαφορετικών μετρικών απόδοσης που χρησιμοποιούν. Επίσης, παρέχουν μια σφαιρική εικόνα από έρευνες που δίνουν έμφαση στην επεκτασιμότητα (scalability) και στην αποδοτικότητα (efficiency) της απόκτησης και επεξεργασίας δεδομένων, και συνιστούν τις καλύτερες πρακτικές. Τέλος, συγκεντρώνουν σε έναν πίνακα όλες τις μελέτες που εξέτασαν, όπου ο ενδιαφερόμενος αναγνώστης μπορεί να βρει για την κάθε μελέτη τον τίτλο και το έτος δημοσίευσης, την πλατφόρμα στην οποία επικεντρώθηκε η μελέτη, μια σύντομη περιγραφή των μεθόδων που

χρησιμοποιήθηκαν, το δημογραφικό στοιχείο που ανιχνεύθηκε και ποια μεταδεδομένα χρησιμοποιήθηκαν για την ανίχνευση του. Στον πίνακα αυτό περιλαμβάνονται πολλές μελέτες για την ανίχνευση φύλου στο Twitter.

Κεφάλαιο 4

Προτεινόμενες αμιγείς προσεγγίσεις και Αξιολόγηση

Στην παρούσα διπλωματική εργασία πραγματοποιήσαμε ανίχνευση του φύλου των χρηστών του Twitter χρησιμοποιώντας δεδομένα που προέρχονται από τα πεδία του προφίλ των χρηστών και πιο συγκεκριμένα την *profile picture*, το *display name* και το *theme color*. Γίνεται χρήση δυο διαδικτυακών υπηρεσιών μέσω των APIs τους: του Face++ για την αναγνώριση φύλου του χρήστη μέσω της *profile picture* του και του Genderize για την αναγνώριση φύλου από το *display name* που έχει επιλέξει ο χρήστης. Για την εκμετάλλευση του *theme color*, κάνουμε χρήση αλγορίθμων επιβλεπόμενης μάθησης υλοποιώντας τους σε γλώσσα Python μέσω των βιβλιοθηκών scikit-learn και neupy.

Οι προσεγγίσεις ανίχνευσης του φύλου με βάση τα προαναφερθέντα στοιχεία του χρήστη έχουν το πλεονέκτημα ότι είναι ανεξάρτητες της γλώσσας. Από τα τρία αυτά χαρακτηριστικά μόνο το όνομα που εξάγεται από το *display name* επηρεάζεται από την χώρα καταγωγής του χρήστη και αυτό αντιμετωπίζεται με την χρησιμοποίηση της παγκόσμιας εμβέλειας βάσης δεδομένων του Genderize. Επιπλέον, οι προσεγγίσεις αυτές προσφέρουν πολύ καλή ταχύτητα και κάνουν δυνατή την επεκτασιμότητα σε μεγάλα σύνολα χρηστών του Twitter. Για την αξιοποίηση της *profile picture* και του *display name,* πραγματοποιείται ένα request για κάθε χρήστη στα APIs του Face++ και του Genderize οπότε ο χρόνος ολοκλήρωσης της διαδικασίας αυξάνεται μόνο γραμμικά με το πλήθος των χρηστών. Κατά την προσέγγιση με βάση το *theme color*, χρησιμοποιείται ελάχιστος αριθμός χαρακτηριστικών για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης.

Στο παρόν Κεφάλαιο, περιγράφουμε αρχικά τη συλλογή δεδομένων για την κατασκευή της βάσης δεδομένων που χρησιμοποιήσαμε για όλα τα πειράματα μας. Στη συνέχεια, παρουσιάζουμε ξεχωριστά τις τρεις μεθόδους ανίχνευσης του φύλου που εφαρμόσαμε με αναλυτική περιγραφή για τη διαδικασία που ακολουθήσαμε στην κάθε μια και σχολιασμό των επιμέρους αποτελεσμάτων. Σημειώνεται ότι στο

παρόν Κεφάλαιο περιγράφονται τα αποτελέσματα χρησιμοποίησης της εικόνας προφίλ και του ονόματος χωρίς χρήση αλγορίθμων μηχανικής μάθησης για να κάνουμε κάποιες ενδιαφέρουσες παρατηρήσεις σχετικά με την διαφορετική συμπεριφορά των δύο φύλων ως προς τα δύο αυτά πεδία. Η εισαγωγή μηχανικής μάθησης στις προσεγγίσεις ανίχνευσης φύλου με βάση την εικόνα και το όνομα, περιγράφεται στο Κεφάλαιο 5.

4.1 Βάση Δεδομένων

Για τα πειράματα που περιγράφονται στην παρούσα εργασία αποκτήσαμε πρόσβαση στο gender-labelled dataset που κάνανε δημόσια διαθέσιμο[20] οι Liu & Ruths [3]. Το dataset αυτό όπως περιγράφηκε και στην Ενότητα 3.2 δημιουργήθηκε με τη χρήση εργατών του Amazon Mechanical Turk οι οποίοι παρατηρώντας τη φωτογραφία προφίλ 50k χρηστών, τους χώρισαν στις κατηγορίες 'Αντρας/Γυναίκα/'Αγνωστο. Με αυτή τη διαδικασία προέκυψε ένα dataset με 4449 χρήστες επισημασμένους ως άντρες και 8232 χρήστες επισημασμένους ως γυναίκες. Όλοι οι χρήστες του dataset αυτού είχαν δημοσιεύσει τουλάχιστον 1000 tweets συνολικά οπότε θεωρούνται αρκετά ενεργοί. Οι Liu & Ruths [3] αναφέρουν ότι διερεύνησαν την ποιότητα των labels φύλου που έδωσαν οι εργάτες του ATM για ένα υποσύνολο χρηστών και κατέληξαν ότι ήταν πολύ αξιόπιστα. Σημειώνουν επίσης ότι το dataset είναι αντιπροσωπευτικό του παγκόσμιου πληθυσμού του Twitter με βάση τη σύγκριση που πραγματοποιήσαν μεταξύ του dataset αυτού και 100k τυχαία επιλεγμένων Άγγλων χρηστών σχετικά μέσω διαφόρων στατιστικών στοιχείων.

Το dataset των Liu & Ruths περιείχε για κάθε έναν από τους 12681 χρήστες μόνο το twitter ID και το φύλο του οπότε κληθήκαμε να συλλέξουμε τα υπόλοιπα στοιχεία των χρηστών τα οποία χρησιμοποιήσαμε στην παρούσα μελέτη. Με βάση το Twitter ID και χρησιμοποιώντας το Twitter API μέσω του Tweepy όπως περιγράφηκε στην Ενότητα 2.3.1, ανακαλύψαμε για τον κάθε χρήστη αν ήταν ακόμα προσβάσιμος μέσω του API (αν δεν είχε διαγράψει το προφίλ του ή δεν είχε περιορίσει την προσβασιμότητα σε “private”). Για όλους τους προσβάσιμους χρήστες αποθηκεύσαμε στη βάση δεδομένων μας τα πεδία *username, display_name, bio, image_url* και *theme_color* όπου μαζί με τα *id* και *gender* του αρχικού dataset που αποκτήσαμε, αποτελούν την πληροφορία που έχουμε για κάθε χρήστη. Η όλη διαδικασία της συλλογής των παραπάνω πεδίων με χρήση του Tweepy ολοκληρώθηκε μέσα σε μια ημέρα.

[20] Πηγή: File “label.json” from http://www.networkdynamics.org/static/datasets/LiuRuthsMicrotext.zip

Η τελική βάση δεδομένων που προέκυψε και την οποία χρησιμοποιήσαμε για τα πειράματα που περιγράφονται στη συνέχεια, αποτελείται από 3283 άντρες και 5176 γυναίκες για ένα σύνολο 8459 χρηστών του Twitter. Οι γυναίκες αποτελούν το 61.19% του συνόλου των χρηστών της βάσης δεδομένων μας σε αντίθεση με το dataset που χρησιμοποίησαν οι Liu & Ruths που το αντίστοιχο ποσοστό των γυναικών ήταν 65%. Αναφέρουμε ότι η αναλογία των 2 φύλων στη βάση δεδομένων μας είναι πιο κοντά στην εκτίμηση της έρευνας [34] που το ποσοστό των γυναικών στο Twitter εκτιμήθηκε στο 55%.

Στο σημείο αυτό αναφέρουμε ότι η επιλογή μας να χρησιμοποιήσουμε για την παρούσα μελέτη τη βάση δεδομένων των Liu & Ruths στηρίζεται στους εξής 3 βασικούς λόγους:

1. Η κατασκευή μιας βάσης δεδομένων που περιέχει χρήστες επισημασμένους με το φύλο τους είναι μια διαδικασία η οποία σε πρώτη φάση είναι αρκετά επίπονη ως προς τον χρόνο και τους πόρους που πρέπει να καταναλωθούν για τη συλλογή ενός μεγάλου συνόλου χρηστών μέσω του Twitter API (οι περισσότερες σχετικές εργασίες αναφέρουν ότι χρειάστηκαν μήνες για τη συλλογή των χρηστών). Στο επόμενο στάδιο, η ταυτοποίηση του φύλου για τους χρήστες στους οποίους είναι εφικτή, μπορεί να γίνει με δυο τρόπους. Ο ένας τρόπος είναι η πληρωμή “εργατών” οι οποίοι θα κάνουν χειροκίνητα την αναγνώριση του φύλου όπως έκαναν οι Liu & Ruths. Ο τρόπος αυτός απορρίφθηκε για οικονομικούς λόγους. Ο δεύτερος τρόπος είναι η χρησιμοποίηση κάποιου πεδίου του προφίλ του Twitter για την εξαγωγή του φύλου με αυτόματα. Ο τρόπος αυτός απορρίφθηκε λόγω χαμηλής αξιοπιστίας.

2. Γενικά οι στρατηγικές κατασκευής των βάσεων από χρήστες του Twitter διαφέρουν στις διάφορες μελέτες, αλλά ακόμα και να χρησιμοποιούνται ίδιες στρατηγικές, τα σύνολα των χρηστών είναι διαφορετικά. Αυτό κάνει αδύνατη την εποικοδομητική σύγκριση των αποτελεσμάτων μεταξύ των διαφορετικών προσεγγίσεων για την ανίχνευση φύλου που χρησιμοποιούνται σε τέτοιες έρευνες. Χρησιμοποιώντας τη βάση των Liu & Ruths έχουμε τη δυνατότητα να συγκρίνουμε την απόδοση της προσέγγισης μας με αυτή της έρευνας [3] στην οποία χρησιμοποιήθηκε μια προσέγγιση που εξαρτάται από τη γλώσσα και περιέχει μεγάλο αριθμό χαρακτηριστικών εξαγόμενων από τα κείμενα των tweets (βλ. Ενότητα 3.2).

3. Οι Liu & Ruths έδειξαν ιδιαίτερη μέριμνα στο να είναι η βάση δεδομένων τους αντιπροσωπευτική του συνολικού πληθυσμού του Twitter. Με αυτό τον τρόπο γνωρίζουμε ότι τα αποτελέσματα που εξάγουμε είναι χρήσιμα για το σύνολο του Twitter.

Για την διεξαγωγή των πειραμάτων μας κάναμε χρήση ολόκληρου του συνόλου των 8459 χρηστών του Twitter. Επειδή τα 2 φύλα δεν είναι ίσα σε αριθμό στη βάση δεδομένων μας, θεωρούμε έναν *baseline gender classifier* ο οποίος προβλέπει ότι όλοι οι χρήστες είναι γυναίκες με ποσοστό επιτυχίας 61.19%. Τα αποτελέσματα των αλγορίθμων που ακολουθούν συγκρίνονται με αυτό το accuracy για την αξιολόγηση της απόδοσης τους.

4.2 Ανίχνευση φύλου μέσω της Εικόνας Προφίλ

4.2.1 Περιγραφή του Face++

Το Face++ Cognitive Services είναι μια πλατφόρμα που προσφέρει τεχνολογίες μηχανικής όρασης. Χρησιμοποιεί ανάλυση φωτογραφίας βασισμένη σε deep-learning μεθόδους που μπορεί να ενσωματοποιηθεί σε custom εφαρμογές. Λανσαρίστηκε το 2012 και έχει χρησιμοποιηθεί από χιλιάδες χρήστες μέσω των facial recognition APIs και SDKs που προσφέρει. Έχει πετύχει επίσης σπουδαία επιτεύγματα στην έρευνα του facial recognition κερδίζοντας παγκόσμιας κλάσης διαγωνισμούς όπως FDDB, 300-W και LFW. Όλα τα APIs και SDKs του Face++ περιέχουν up-to-date αλγορίθμους που παρέχουν κορυφαία αποδοτικότητα, αξιοπιστία και ακρίβεια.

4.2.2 Face Detection API

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε το Face Detection API του Face++ το οποίο προσφέρει ευρωστία (robustness) επειδή έχει υψηλή απόδοση ενάντια σε επιρροές όπως μερικό μπλοκάρισμα προσώπου, ακατάλληλο φωτισμό και πόζα κεφαλιού. Επίσης είναι ακριβές και γρήγορο.

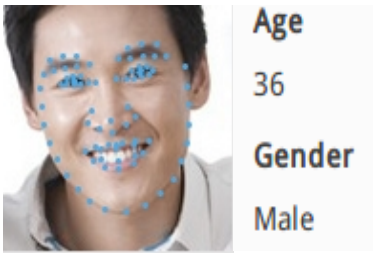
Το Face Detection API διακρίνει όλα τα ανθρώπινα πρόσωπα που περιέχονται σε μια φωτογραφία όπως φαίνεται στο Σχήμα 4.1. Κάθε εντοπισμένο πρόσωπο λαμβάνει ένα *face_token* το οποίο μπορεί να χρησιμοποιηθεί για επακόλουθη ανάλυση. Κατά

τη δημιουργία δωρεάν λογαριασμού στο Face++ αποκτήσαμε ένα Free API Key και ένα Free API Secret τα οποία χρησιμοποιήθηκαν για το authentication των request που κάναμε στο Face Detection API. Με το Free API Key, έχουμε τη δυνατότητα να κάνουμε ανάλυση των 5 μεγαλύτερων προσώπων που αναγνωρίζονται στην εκάστοτε φωτογραφία. Η ανάλυση παρέχει μια ποικιλία χαρακτηριστικών για το κάθε πρόσωπο όπως φύλο, ηλικία, εθνικότητα, εμφάνιση χαμόγελου, πόζα κεφαλιού, συναίσθημα και κατάσταση ματιών (ύπαρξη γυαλιών, ανοιχτό ή κλειστό μάτι κλπ.). Στα Σχήματα 4.2 και 4.3 βλέπουμε την ηλικία και το φύλο που συμπέρανε το Face++ για δυο από τα πρόσωπα του Σχήματος 4.1. Το μόνο χαρακτηριστικό που χρησιμοποιήσαμε στην παρούσα μελέτη είναι το φύλο που αντιστοιχεί στο κάθε αναγνωρισμένο πρόσωπο.

Στη συνέχεια ακολουθεί λεπτομερής περιγραφή του τρόπου με τον οποίο εκμεταλλευτήκαμε τις λειτουργίες του Face++ για την ανίχνεση φύλου των χρηστών του Twitter.

Σχήμα 4.1: Αναγνώριση προσώπων από το Face++

Σχήμα 4.2: Ανάλυση γυναικείου προσώπου από το Face++



Σχήμα 4.3: Ανάλυση αντρικού προσώπου από το Face++

4.2.3 Διεξαγωγή πειραμάτων με το Face++

Για την “offline” ανάλυση της ανίχνευσης φύλου των χρηστών του Twitter με βάση την φωτογραφία προφίλ τους προσθέσαμε στη βάση δεδομένων μας το πεδίο *faceplusplus_gender* για την εκτίμηση που κάνει το Face++ για το φύλο κάθε χρήστη. Η διαδικάσια που ακολουθήσαμε περιγράφεται παρακάτω:

Για κάθε χρήστη στην βάση δεδομένων μας, πήραμε το URL της φωτογραφίας προφίλ του (πεδίο *image_url*). Αν αυτό το URL αντιστοιχούσε στην default profile picture του Twitter[21], δεν κάναμε κλήση στο Face Detection API του Face++ για εξοικονόμηση πόρων, αφού η συγκεκριμένη φωτογραφία δεν θα μας δώσει περισσότερες πληροφορίες σχετικά με το φύλο του χρήστη (αποθηκεύσαμε *faceplusplus_gender = unknown* για την εγγραφή του χρήστη αυτού στη βάση δεδομένων μας). Οι χρήστες της βάσης δεδομένων μας που είχαν την προκαθορισμένη φωτογραφία του Twitter στο προφίλ τους ήταν μόνο 20. Πιο συγκεκριμένα, μόλις το 0.25% του συνόλου των γυναικών και το 0.21% του συνόλου των αντρών δεν είχαν ανεβάσει κάποια φωτογραφία της επιλογής τους. Από αυτό συμπεραίνουμε ότι οι συγκεκριμένοι χρήστες αποτελούν ένα αμελητέο δείγμα που δεν μας επηρέασε στα αποτελέσματα της ανάλυσης εικόνας που διεξάγαμε.

Για τους χρήστες που δεν είχαν την προκαθορισμένη φωτογραφία προφίλ, πραγματοποίησαμε ένα HTTP POST request στο https://api-us.faceplusplus.com/facepp/v3/detect δίνοντας τις ακόλουθες παραμέτρους σύμφωνα με το documentation του Face++[22] :

[21] https://abs.twimg.com/sticky/default_profile_images/default_profile_normal.png
[22] https://console.faceplusplus.com/documents/5679127

- api_key: Το registered Free API Key που έχουμε στη διάθεση μας για να καλέσουμε το API.
- api_secret: Το registered API Secret που έχουμε στη διάθεση μας για να καλέσουμε το API.
- image_url: Το URL της φωτογραφίας προφίλ του εκάστοτε χρήστη που επεξεργαζόμαστε.
- return_attributes: Τα χαρακτηριστικά των προσώπων που θέλουμε να μας επιστρέψει το Face++. Στη συγκεκριμένη περίπτωση “gender”.

Το Face++ επιστρέφει ένα JSON object στο οποίο περιέχονται τα πρόσωπα που αναγνώρισε μαζί με τα επιθυμητά χαρακτηριστικά για το καθένα. Ένα υπόδειγμα απάντησης του Face++ για μια φωτογραφία που περιέχει ένα αντρικό και ένα γυναικείο πρόσωπο φαίνεται παρακάτω:


```
{'time_used': 310,
 'image_id': 'MyfI0BJdzmWeX9D2m3cg7A==',
 'faces':
[{'attributes':
{'gender': {'value': 'Male'}},
'face_token': '974a97092bd305c80963b8b1e19a8ea4',
 'face_rectangle': {'width': 258, 'top': 123, 'height': 258, 'left': 308}},
 {'attributes':
{'gender': {'value': 'Female'}},
 'face_token': 'f7a1ce415e018e7f8d18c605c4356d68', 'face_rectangle': {'width': 232, 'top': 114, 'height':
232, 'left': 746}}],
 'request_id': '1511978034,9399e12e-be41-482b-ae8a-dc13c4bcba67'}
```


Η όλη διαδικασία της αποστολής request στο API του Face++ και της λήψης των απαντήσεων ολοκληρώθηκε στο διάστημα μιας ημέρας.

Σημειώνουμε ότι πριν στείλουμε το URL της εικόνας προφίλ στο Face++ πραγματοποιήσαμε μια προεπεξεργασία. Πιο συγκεκριμένα, επειδή το URL που μας επιστρέφει το Twitter API αντιστοιχεί σε εικόνα χαμηλής ανάλυσης, αντικαταστήσαμε τη λέξη “normal” στη συμβολοσειρά του URL με το “400x400”. Για παράδειγμα το URL:
https://pbs.twimg.com/profile_images/942844760938643458/gHJ7laua_normal.jpg
μετατράπηκε σε:
https://pbs.twimg.com/profile_images/942844760938643458/gHJ7laua_400x400.jpg

Με αυτό τον τρόπο διασφαλίσαμε ότι το Face++ θα έχει την καλύτερη δυνατή απόδοση στην αναγνώριση προσώπων επεξεργάζοντας μια εικόνα υψηλότερης ανάλυσης.

Διαχωρίσαμε τις πιθανές απαντήσεις του Face++ σε υποπεριπτώσεις. Για κάθε υποπερίπτωση παραθέτουμε την αντίστοιχη τιμή που αποθηκεύσαμε στην βάση μας στο πεδίο *faceplusplus_gender* του κάθε χρήστη:

- Το Face++ δεν ανίχνευσε κανένα πρόσωπο στη φωτογραφία: *faceplusplus_gender = unknown*.

- Το Face++ ανίχνευσε έναν η περισσότερους άντρες στη φωτογραφία: *faceplusplus_gender = male*.

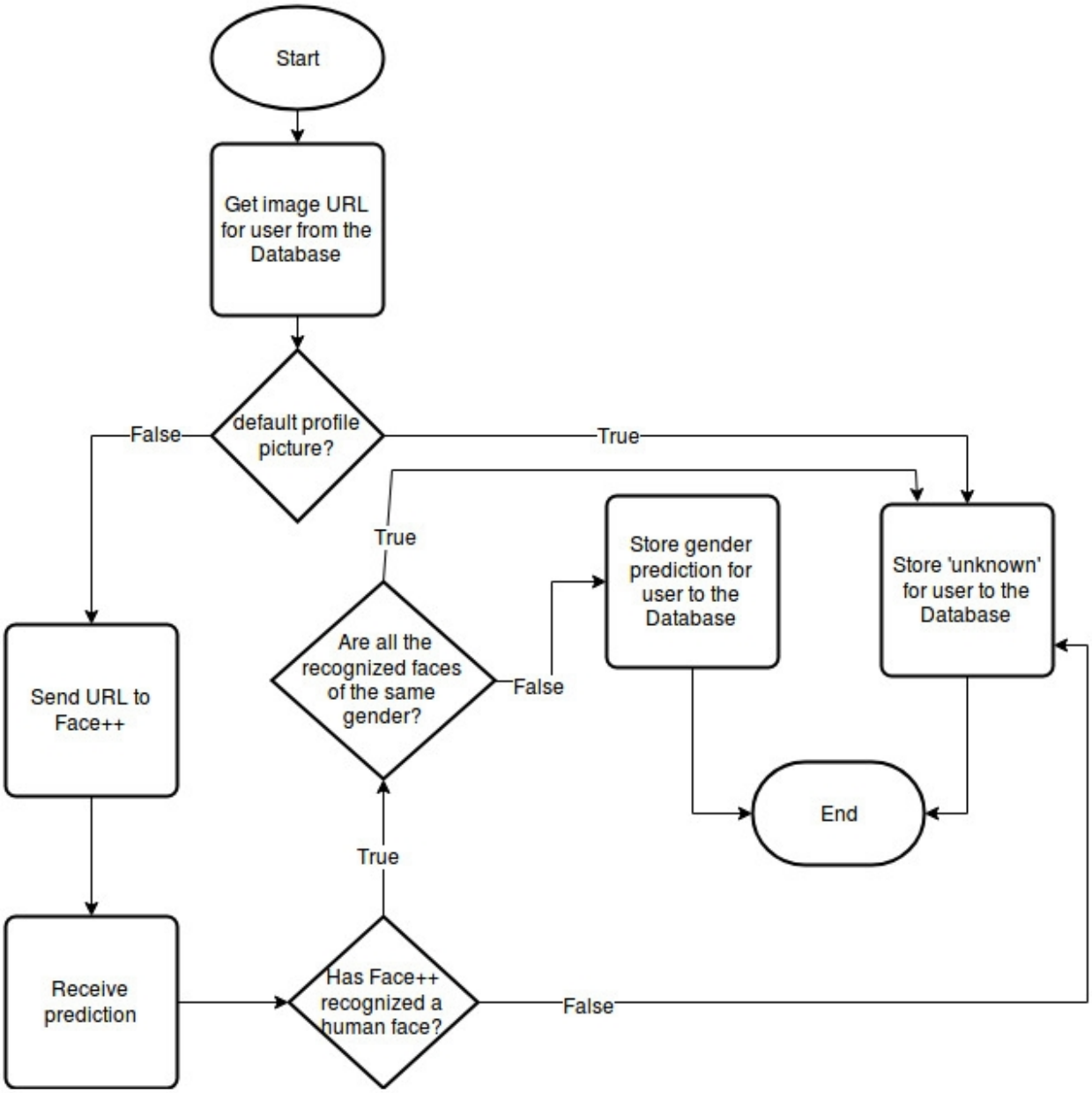
- Το Face++ ανίχνευσε μια η περισσότερες γυναίκες στη φωτογραφία: *faceplusplus_gender = female*.

- Το Face++ ανίχνευσε έναν ή περισσότερους άντρες και μια ή περισσότερες γυναίκες στη φωτογραφία: *faceplusplus_gender = unknown*.

Στην τελευταία περίπτωση όπου αναγνωρίστηκαν και άντρες και γυναίκες στην φωτογραφία προφίλ θα μπορούσαμε να έχουμε αποθηκεύσει το επικρατέστερο φύλο ως προς τον αριθμό των εμφανίσεων ή το φύλο του πρώτου προσώπου που αναγνωρίστηκε. Επιλέξαμε όμως να θεωρήσουμε την ανίχνευση “άκυρη” με σκοπό να έχουμε πιο ακριβείς παρά μεγαλύτερες σε αριθμό εκτιμήσεις και με τον συγκεκριμένο διαχωρισμό περιπτώσεων πετυχαίνουμε αυτό ακριβώς.

Σημειώνεται ότι ο περιορισμός του Face Detection API που επιστρέφει το φύλο μόνο για τα 5 επικρατέστερα πρόσωπα, δεν φαίνεται να επηρεάζει τα αποτελέσματα μας καθώς επιβεβαιώσαμε ότι οι χρήστες που έχουν στην φωτογραφία τους πάνω από 5 πρόσωπα τα οποία επιπλέον είναι του ίδιου φύλου, είναι αμελητέοι.

Στο Σχήμα 4.4 φαίνεται το διάγραμμα ροής για τη διαδικάσια που ακολουθήσαμε για την αναγνώριση φύλου από την εικόνα ενός χρήστη της βάσης δεδομένων μας.



Σχήμα 4.4: Διάγραμμα Ροής για την αναγνώριση φύλου ενός χρήστη μέσω του Face++

4.2.4 Πειραματικά αποτελέσματα βάσει εικόνας προφίλ χρήστη

Με την ολοκλήρωση της ανάλυσης εικόνας προφίλ από το Face++ για όλους τους χρήστες της βάσης δεδομένων μας πήραμε τα ακόλουθα στοιχεία:

- Για τους χρήστες που σημειώθηκε εκτιμώμενο φύλο (δεν υπήρχε *unknown* στο αντίστοιχο πεδίο *faceplusplus_gender* της βάσης μας) επιτεύχθει 88.06% ακρίβεια στο σύνολο των προβλέψεων φύλου. Ειδικότερα, στο σύνολο των αντρών που έγινε εκτίμηση φύλου είχαμε 87.17% ακρίβεια ενώ στο αντίστοιχο σύνολο των γυναικών 88.58% ακρίβεια.

- Οι χρήστες για τους οποίους αναγνωρίστηκε ένα συγκεκριμένο φύλο στην φωτογραφία τους (αναγνωρίστηκε τουλάχιστον ένα πρόσωπο και δεν εμφανίστηκαν πρόσωπα και των δυο φύλων) αποτελούσαν το 74.67% του συνόλου της βάσης. Πιο αναλυτικά, το ποσοστό των αντρών στους οποιους αναγνωρίστηκε ένα συγκεκριμένο φύλο ως προς τον συνολικό αριθμό τους ήταν 70.76% ενώ το αντίστοιχο ποσοστό των γυναικών ήταν 77.14%.

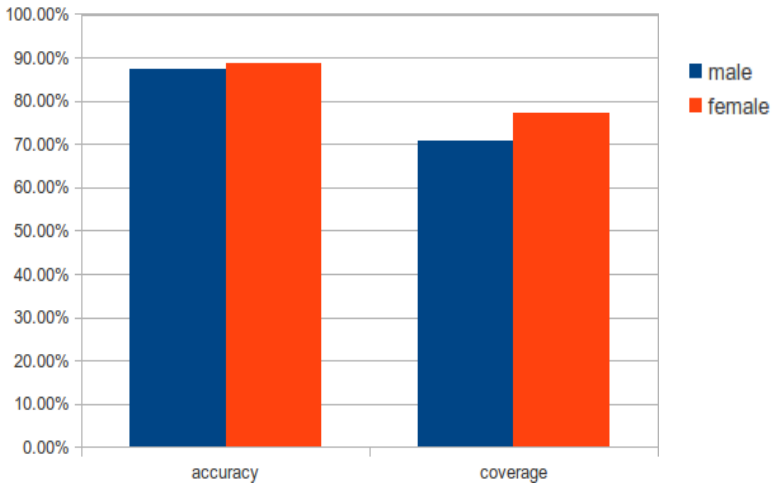
- Το ποσοστό των σωστών προβλέψεων φύλου με βάση το Face++ ως προς το σύνολο όλων των χρηστών της βάσης μας ήταν 65.75% (θεωρώντας όλους τους χρήστες με *faceplusplus_gender* = *unknown* ως λάθος προβλέψεις) .

Από τα παραπάνω στοιχεία μπορούμε να εξάγουμε το συμπέρασμα ότι ο προσδιορισμός φύλου με βάση την φωτογραφία προφίλ ενός χρήστη του Twitter έχει αρκετά καλή επίδοση δεδομένου ότι ο χρήστης έχει ανεβάσει κάποια φωτογραφία που περιέχει ανθρώπινα πρόσωπα και δεν συνυπάρχουνε σε αυτήν πρόσωπα και των δυο φύλων. Αυτό μας δείχνει ότι γενικά οι χρήστες που βάζουν φωτογραφίες προσώπων ενός μοναδικού φύλου στο προφίλ τους, τείνουν να επιλέγουν πρόσωπα του ίδιου φύλου με αυτούς ανεξάρτητα από το αν απεικονίζονται οι ίδιοι στις φωτογραφίες. Αντίθετα, το χαμηλό ποσοστό των σωστών προβλέψεων ως προς τον συνολικό αριθμό των χρηστών, μας δείχνει ότι η χρήση του Face++ για την ανάλυση των φωτογραφιών δεν αρκεί για την ανίχνευση φύλου των χρηστών και προτείνεται να χρησιμοποιηθεί σε συνδυασμό με άλλες μεθόδους ανίχνευσης. Η ακρίβεια των προβλέψεων του φύλου όλων των χρηστών της βάσης μας με χρήση του Face++ είναι μόνο 4.56% καλύτερη από την ακρίβεια 61.19% του *baseline gender classifier* ο οποίος προβλέπει ότι όλοι οι χρήστες είναι γυναίκες.

Όσον αφορά τις διαφορές των 2 φύλων, παρατηρούμε ότι γενικά οι γυναίκες είναι πιο πιθανό να ανεβάσουν στο Twitter προφίλ τους φωτογραφία που να αντιπροσωπεύει το φύλο τους σε σχέση με τους άντρες. Αυτό έρχεται σε συμφωνία με το συμπέρασμα της μελέτης [3] όπου οι συγγραφείς παρατήρησαν ότι το gender-labelled dataset που σύλλεξαν με βάση τις φωτογραφίες προφίλ περιείχε μεγαλύτερο ποσοστό γυναικών σε σχέση με παλαιότερα ευρήματα. Τα αποτελέσματα του Face++ για τα 2 φύλα φαίνονται στο Σχήμα 4.5.[23]

Πρέπει να σημειώσουμε σε αυτό το σημείο ότι η αναγνώριση του φύλου των χρηστών στο αρχικό dataset στο οποίο βασιστήκαμε, έγινε όπως έχουμε αναφέρει με βάση την παρατήρηση της φωτογραφίας προφίλ τους από εργάτες του Amazon Mechanical Turk, οι οποίοι είχαν τη δυνατότητα να αφαιρέσουν χρήστες που η φωτογραφία τους δεν περιείχε ανθρώπινα πρόσωπα ή απεικόνιζε κάποια διασημότητα. Υπάρχει μεγάλη πιθανότητα λοιπόν, τα αποτελέσματα που πήραμε από το Face++ να είναι μεροληπτικά και να μην αντιπροσωπεύουν καθολικά το σύνολο των χρηστών του Twitter. Για παράδειγμα το πολύ χαμηλό ποσοστό των χρηστών με default profile picture που αναφέραμε στην Ενότητα 4.2.3, μπορεί να είναι ένα δείγμα αυτής της μεροληψίας (ωστόσο δεν έχουμε επίσημα στοιχεία για το αντίστοιχο ποσοστό στο σύνολο όλων των χρηστών του Twitter για να συγκρίνουμε). Εντούτοις, η ταυτοποίηση αυτή στο αρχικό dataset έγινε πριν το 2013, και στην παρούσα δουλειά όπως περιγράψαμε και στην Ενότητα 4.1 ξαναπήραμε για όλους τους (ακόμα προσβάσιμους) χρήστες της αρχικής βάσης, τα τωρινά στοιχεία του προφίλ τους μέσω του Twitter API. Μέσα σε αυτά τα στοιχεία ήταν και η φωτογραφία προφίλ την οποία είχαν επιλέξει οι χρήστες κατά το διάστημα εκπόνησης της διπλωματικής αυτής εργασίας. Συμπερασματικά, το μόνο πιθανό μεροληπτικό στοιχείο που παραμένει όσον αφορά τα αποτελέσματα του Face++ είναι το γεγονός ότι οι χρήστες που είχαν επιλέξει αντιπροσωπευτική φωτογραφία του φύλου τους κατά το διάστημα της μελέτης [3], είναι μάλλον πιο πιθανό να είχαν επιλέξει πάλι κάποια gender-specific φωτογραφία κατά την προσπέλαση των στοιχείων τους μέσω του Twitter API, σε σχέση με το γενικότερο σύνολο των χρηστών του Twitter.

[23] Ως coverage (ποσοστό κάλυψης) ορίσαμε το ποσοστό των χρηστών για τους οποίους πραγματοποιήθηκε πρόβλεψη ως προς το σύνολο όλων των χρηστών για κάθε φύλο ξεχωριστά.



Σχήμα 4.5: Ραβδόγραμμα ποσοστών accuracy και coverage για τα 2 φύλα με το Face++

4.3 Ανίχνευση φύλου μέσω του Ονόματος

4.3.1 Περιγραφή του Genderize

Το Genderize.io είναι μια διαδικτυακή υπηρεσία προσδιορισμού του φύλου με βάση το όνομα. Μπορεί να χρησιμοποιηθεί ως δομικό στοιχείο για data analytics, σχεδιασμό στοχευμένων διαφημίσεων ανά φύλο κλπ. Χρησιμοποιεί μεγάλες βάσεις δεδομένων, προερχόμενες από προφίλ χρηστών σε δημοφιλή κοινωνικά δίκτυα και παρέχει αυτά τα δεδομένα μέσω του API του. Αυτή τη στιγμή περιέχει 216286 ξεχωριστά ονόματα από 79 διαφορετικές χώρες και 89 διαφορετικές γλώσσες. Για κάθε όνομα επιστρέφει εκτός από το σχετικό φύλο, και έναν παράγοντα βεβαιότητας με τη μορφή πιθανότητας καθώς και τον αντίστοιχο αριθμό εγγραφών του ονόματος αυτού στις βάσεις δεδομένων.

Παρέχει επιπλέον τη δυνατότητα εφαρμογής “localization” φίλτρων για την επιστροφή αποτελεσμάτων βασισμένων σε δεδομένα μιας συγκεκριμένης χώρας ή γλώσσας. Αυτό είναι αρκετά σημαντικό γιατί οι ονομασίες μπορεί να βασίζονται σε μεγάλο βαθμό στην καταγωγή των ατόμων. Στην παρούσα μελέτη δεν χρησιμοποιήθηκε το localization για λόγους που θα εξηγηθούν παρακάτω.

Το Genderize επιλέχθηκε για τους σκοπούς της παρούσας διπλωματικής εργασίας επειδή αποτελεί μια φθηνή και εύκολη λύση για την ενσωμάτωση της ανάλυσης ονομάτων στις μεθόδους που ακολουθούμε για την ανίχνευση του φύλου. Λόγω της παγκόσμιας εμβέλειας των ονομάτων που περιέχει, δίνει τη δυνατότητα για την ανίχνευση φύλου της συντριπτικής πλειοψηφίας του συνολικού πληθυσμού του Twitter. Επίσης, λόγω της φύσης κατασκευής των βάσεων δεδομένων του, που προέρχονται όπως αναφέραμε από κοινωνικά δίκτυα, είναι κατάλληλο για ανάλυση στοχευμένη στο Twitter.

4.3.2 Διεξαγωγή πειραμάτων με το Genderize

Για να πραγματοποιήσουμε ανίχνευση φύλου με βάση το όνομα του χρήστη με χρήση του Genderize, ακολουθήσαμε τη μέθοδο που αναλύεται παρακάτω.

Αρχικά, κληθήκαμε να αποφασίσουμε αν θα χρησιμοποιήσουμε το *display name* ή/και το *username* των χρηστών για την εξαγωγή του ονόματος το οποίο θα δοθεί στο Genderize για τον προσδιορισμό του φύλου. Το *display name* όπως είδαμε και στην Ενότητα 2.1 αποτελεί ένα προσωπικό αναγνωριστικό για τον κάθε χρήστη χωρίς τον περιορισμό να είναι μοναδικό σε όλο το Twitter. Ως αποτέλεσμα, πολλοί χρήστες επιλέγουν να καταχωρήσουν το ονοματεπώνυμό τους στο συγκεκριμένο πεδίο [4], [36], [37]. Αντίθετα, το *username* επειδή πρέπει να είναι μια μοναδική και χωρίς κενά συμβολοσειρά, αποτελεί λιγότερο χρήσιμη πηγή πραγματικών ονομάτων για δυο λόγους. Πρώτον, για την εξαγωγή ονομάτων από το *username* χρειάζεται πιο σύνθετη ανάλυση σε σχέση με το *display name*. Για παράδειγμα, μια χρήστης του Twitter θα μπορούσε να έχει *display name*: “Jann Tomson” και *username*: “19j4nn_t0mson82”. Από την συμβολοσειρά του *username* θα πρέπει αρχικά να αντικατασταθεί το “leet speak”. Στη συνέχεια κατά τη διαδικάσια εύρεσης πιθανών ονομάτων χρειάζεται ιδιαίτερη προσοχή για να αποκλειστούν άκυρα ονόματα που περιέχονται στην συμβολοσειρά όπως “Ann” και “Tom”. Δεύτερον, ακόμα και μετά από σύνθετη ανάλυση, το *username* φαίνεται να περιέχει λιγότερο αξιόπιστη πληροφορία σε σχέση με το *display name* [4]. Για τους παραπάνω λόγους επιλέξαμε να χρησιμοποιήσουμε μόνο το *display name* στην παρούσα μελέτη, επειδή όπως έχουμε διατυπώσει σκοπός μας είναι να προτείνουμε μια μέθοδο ανίχνευσης του φύλου που συνδυάζει απλότητα και καλή απόδοση. Σημειώνεται ότι και οι Liu & Ruths [3] στη μελέτη τους χρησιμοποίησαν όπως αναφέρουν το “self-reported name” των χρηστών το οποίο αναφέρεται στο *display name* χωρίς να αναφέρουν όμως αναλυτικά τον τρόπο με τον οποίο

επεξεργάστηκαν το πεδίο για να εξαγάγουν τα ονόματα που αναζήτησαν στα δεδομένα του US Census.

Το localization feature του Genderize δεν χρησιμοποιήθηκε γιατί δεν έχουμε πρόσβαση σε κάποια αξιόπιστη πληροφορία σχετικά με την εθνικότητα των χρηστών. Πιο συγκεκριμένα, τα πεδία του προφίλ ενός χρήστη στο Twitter που σχετίζονται με τοποθεσία είναι το προαιρετικό πεδίο *location* το οποίο αναφέρεται στην τοποθεσία του χρήστη και όχι στην καταγωγή του και επιπλέον ο χρήστης μπορεί να το συμπληρώσει αυθαίρετα και το πεδίο *language* που σχετίζεται με την γλώσσα που επιθυμεί ο χρήστης να βλέπει το προφίλ του, όπου πολλοί χρήστες διαφορετικής εθνικότητας επιλέγουν τα Αγγλικά. Επίσης υπάρχει η δυνατότητα της αναγνώρισης της γλώσσας των tweets ενός χρήστη. Όμως πολλοί χρήστες γράφουν tweets στα Αγγλικά ανεξαρτήτως της καταγωγής τους. Συνεπώς, δεν βασιστήκαμε σε κανένα από τα παραπάνω για γεωγραφικό περιορισμό της αναζήτησης των ονομάτων στο Genderize.

Για την εξαγωγή του ονόματος του χρήστη από το *display name* πραγματοποιήσαμε την εξής προεπεξεργασία:

1. Αφαιρέσαμε όλους τους χαρακτήρες τις συμβολοσειράς που δεν ήταν γράμματα της Αγγλικής αλφαβήτου.

2. Μετατρέψαμε τυχόν κεφαλαία γράμματα της συμβολοσειράς σε πεζά.

3. Διαιρέσαμε την προκύπτουσα συμβολοσειρά σε πιθανά ονόματα με βάση τα κενά της. Για παράδειγμα, από τη συμβολοσειρά “tommy lee jones” παίρνουμε τα πιθανά ονόματα: “tommy”, **“**lee”, και **“**jones”.

Δημιουργώντας έναν λογαριασμό στο Genderize με ένα οικονομικό basic subscription, αποκτήσαμε ένα API Key το οποίο χρησιμοποιήθηκε για το authentication των request που κάναμε στο Genderize API. Με το API Key του basic subscription που είχαμε στη διάθεση μας, είχαμε τη δυνατότητα να πραγματοποιήσουμε 100k requests ανά μήνα για την αναγνώριση των ονομάτων, τα οποία ήταν υπεραρκετά για το πλήθος των χρηστών της βάσης δεδομένων μας.

Για τον προσδιορισμό του φύλου που αντιστοιχεί σε κάποιο πιθανό όνομα που προέκυψε από το *display name* ενός χρήστη, πραγματοποιήσαμε ένα HTTP GET resquest στο Genderize API με τη μορφή https://api.genderize.io/?name=example.

Διακρίνουμε τις εξής 2 περιπτώσεις:

- Αν το πιθανό όνομα υπάρχει σαν εγγραφή στις βάσεις δεδομένων του Genderize έστω και μια φορά, μας επιστρέφεται σαν response το εκτιμώμενο φύλο μαζί με τον παράγοντα βεβαιότητας σωστής εκτίμησης (σε μορφή πιθανότητας) και τον συνολικό αριθμό εγγραφών του ονόματος. Για παράδειγμα, ένα GET request με το όνομα “tommy” ως παράμετρο δίνει σαν απάντηση το JSON object:

```
{'count': 979,
 'gender': 'male',
 'name': 'tommy',
 'probability': 0.99}.
```

- Αν το πιθανό όνομα δεν υπάρχει στις βάσεις δεδομένων του Genderize, μας επιστρέφεται σαν φύλο η τιμή “None”. Δίνοντας για παράδειγμα σαν πιθανό όνομα το “abc”, παίρνουμε το JSON object:

```
{'gender': None,
 'name': 'example'}.
```

Για την “offline” ανάλυση της ανίχνευσης φύλου των χρηστών του Twitter με βάση το *display name* τους προσθέσαμε στη βάση δεδομένων μας τα πεδία *genderize_gender, genderize_probability* και *genderize_count*. Η διαδικάσια που ακολουθήσαμε για κάθε χρήστη είναι η εξής:

Στέλνουμε requests στο Genderize για κάθε πιθανό όνομα στη λίστα αρχίζοντας από το πρώτο που εξήχθει από το *display name* κ.ο.κ. μέχρι το Genderize να βρει κάποιο από αυτά στη βάση δεδομένων του. Με την εύρεση ενός ονόματος η διαδικασία σταματάει και αποθηκεύεται στη βάση με τους χρήστες μας οι πληροφορίες που επέστρεψε το Genderize για το συγκεκριμένο όνομα, στα αντίστοιχα πεδία *genderize_gender*, *genderize_probability* και *genderize_count*. Αν προσπελασθούν όλα τα πιθανά ονόματα της λίστας χωρίς να αναγνωριστεί κάποιο από το Genderize, αποθηκεύουμε στα 3 παραπάνω πεδία την τιμή “unknown”.

Η όλη διαδικασία ολοκληρώθηκε σε μια ημέρα. Παρακάτω περιγράφονται τα αποτελέσματα που συγκεντρώσαμε με βάση αυτά τα 3 πεδία για όλους τους χρήστες της βάσης μας.

Σημειώνεται ότι μια άλλη τακτική που θα μπορούσαμε να έχουμε ακολουθήσει για την ανίχνευση του φύλου ενός χρήστη από το *display name* είναι να στέλνουμε όλα

τα πιθανά ονόματα που εξάγουμε στο Genderize και να κρατάμε αυτό που αντιστοιχεί σε μεγαλύτερα *genderize_count* ή/και *genderize_ probability.* Επιλέξαμε να μην ακολουθήσουμε αυτή την τακτική επειδή συνήθως τα ονόματα των χρηστών είναι στην πρώτη θέση του *display name*, οπότε θα μπορούσαμε να έχουμε θεωρήσει λανθασμένα πολλά επίθετα ως ονόματα.

4.3.3 Πειραματικά αποτελέσματα βάσει ονόματος χρήστη

Με την ολοκλήρωση της διαδικασίας επεξεργασίας των απαντήσεων του Genderize για όλους τους χρήστες της βάσης δεδομένων μας, αναλύσαμε τα πεδία *genderize_gender*, *genderize_probability* και *genderize_count* για την εξαγωγή αποτελεσμάτων.

Λαμβάνοντας υπόψην μόνο το πεδίο *genderize_gender* για κάθε χρήστη, πήραμε τα ακόλουθα στοιχεία:

- Για τους χρήστες που βρέθηκε κάποιο από τα πιθανά ονόματα που εξήχθησαν από το *display name,* στις βάσεις δεδομένων του Genderize (δεν υπήρχε *unknown* στο αντίστοιχο πεδίο *genderize_gender* της βάσης μας), είχαμε 81.55% ποσοστό σωστών εκτιμήσεων για το φύλο. Αναλυτικότερα, στο σύνολο των αντρών είχαμε 84.87% σωστές εκτιμήσεις ενώ αντίστοιχα στις γυναίκες 79.41%.

- Οι χρήστες για τους οποίους αναγνωρίστηκε κάποιο πιθανό όνομα από το Genderize (υπήρχε τουλάχιστον 1 σχετική εγγραφή στις βάσεις του Genderize) αποτελούσαν το 73.96% του συνόλου της βάσης μας. Πιο συγκεκριμένα, το ποσοστό των αντρών για τους οποίους έγινε ανίχνευση από το Genderize ως προς τον συνολικό αριθμό τους ήταν 74.87% ενώ το αντίστοιχο ποσοστό των γυναικών ήταν 73.38%.

- Το ποσοστό των σωστών προβλέψεων φύλου με βάση το *genderize_gender* ως προς το σύνολο όλων των χρηστών της βάσης μας ήταν 60.31% (θεωρώντας όλους τους χρήστες με *genderize_gender = unknown* ως λάθος προβλέψεις)

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι η ανίχνευση φύλου των χρηστών του Twitter που έχουν στο *display name* του προφίλ τους κάποιο όνομα

αναγνωρισμένο από το Genderize, παρουσιάζει μια σχετικά καλή απόδοση ακόμα και χωρίς να λάβουμε υπόψην τα πεδία *genderize_probability* και *genderize_count*. Επίσης περίπου 3 στους 4 χρήστες φαίνεται να επιλέγουν για το *display name* τους μια συμβολοσειρά που περιέχει gender-specific στοιχεία. Αυτό έρχεται σε αντίθεση με την παρατήρηση των Liu & Ruths [3] ότι μόνο 1 στους 3 χρήστες της βάσης τους είχαν όνομα με γνωστό συσχετισμό με κάποιο φύλο σύμφωνα με τα δεδομένα του US Census που χρησιμοποίησαν. Δεδομένου οτι χρησιμοποιήσαμε για την παρούσα μελέτη ένα υποσύνολο του dataset τους, θεωρούμε ότι η διαφορά αυτή οφείλεται στη χρήση του Genderize αντί του US Census. Το Genderize, έχοντας στις βάσεις δεδομένων του ονόματα από πολλές διαφορετικές χώρες και γλώσσες, παρέχει μεγαλύτερη κάλυψη ονομάτων. Επιπλέον επειδή όπως αναφέρεται από το Genderize, η συλλογή των ονομάτων έγινε από τα προφίλ χρηστών μεγάλων κοινωνικών δικτύων, μάλλον το Genderize προσφέρει ιδιαίτερη χρησιμότητα σε έρευνες που επικεντρώνονται στο Twitter. Αντίθετα το US Census είναι μεροληπτικό ως προς Δυτικά, Αμερικάνικα ονόματα όπως αναφέρουν και οι Liu & Ruths [3]. Ακόμα, παρατηρούμε ότι οι άντρες τείνουν να τονίζουν περισσότερο το φύλο τους μέσω του ονόματος σε αντίθεση με την περίπτωση της φωτογραφίας προφίλ όπως αναφέραμε στην Ενότητα 4.2.4. Σημειώνεται ότι στη μελέτη [3] τα πιο πολλά αναγνωρισμένα ονόματα ήταν γυναικεία. Η διαφορά αυτή μπορεί να οφείλεται στη χρησιμοποίηση του Genderize αντί του US Census όπως εξηγήσαμε προηγουμένως.

Βλέπουμε ότι το ποσοστό των σωστών προβλέψεων με βάση το *genderize_ gender* ως προς τον συνολικό αριθμό των χρηστών είναι χειρότερο συγκριτικά με τον *baseline gender classifier* ο οποίος προβλέπει ότι κάθε χρήστης στη βάση μας είναι γυναίκα. Αυτό μας δείχνει ότι η ανίχνευση φύλου των χρηστών του Twitter με χρήση του *genderize_gender*, δεν αρκεί και πρέπει να χρησιμοποιηθεί σε συνδυασμό με άλλες μεθόδους ανίχνευσης.

Στη συνέχεια της έρευνας μας, λάβαμε υπόψην και τα πεδία *genderize_ probability* και *genderize_count* που είχαμε αποθηκευμένα στη βάση δεδομένων μας, για να μελετήσουμε την επίδραση του κάθε πεδίου ξεχωριστά στις προβλέψεις φύλου με βάση το όνομα. Για τα πειράματα μας κρατήσαμε μόνο το σύνολο *X* των ονομάτων της βάσης μας που αναγνωρίστηκαν από το Genderize (κάθε όνομα αντιστοιχεί σε έναν μοναδικό χρήστη). Θεωρήσαμε τις μεταβλητές c_x και p_x που δηλώνουν τον αριθμό εγγραφών ενός ονόματος *x* στις βάσεις του Genderize (*genderize_count*) και την πιθανότητα σωστής ανίχνευσης φύλου για αυτό το όνομα

(*genderize_probability*) αντίστοιχα. Παρακάτω αναλύουμε την επίδραση των 2 αυτών πεδίων σε συνδυασμό με το *genderize_gender* στην ακρίβεια των προβλέψεων με βάση το Genderize καθώς και το ποσοστό των ονομάτων του συνόλου X το οποίο αντιστοιχεί στις σχετικές τιμές των 2 αυτών πεδίων (ποσοστό κάλυψης-coverage):

- ***genderize_count***: Από το πεδίο *genderize_count* της βάσης μας παρατηρήσαμε ότι για όλα τα ονόματα x είχαμε $c_x \in [1,12593]$. Θεωρήσαμε ένα count threshold t_c με σκοπό να μελετήσουμε την απόδοση του Genderize στους χρήστες που είχαν στο προφίλ τους όνομα με $c_x \geq t_c$ αλλά και το σχετικό ποσοστό των χρηστών αυτών ως προς τον συνολικό αριθμό των χρηστών με αναγνωρισμένα ονόματα. Για το πείραμα μας μεταβάλλαμε το t_c σε όλο το διάστημα $[1,12593]$ με βήμα 10 δηλαδή $t_c = \{1 + i * 10 : i = 0,\ldots,1260\}$. Οι επιδράσεις που παρατηρήσαμε στα accuracy και coverage με τη μεταβολή του t_c ήταν οι εξής:

  - **accuracy**: Όπως ήταν λογικό, οι προβλέψεις φύλου με βάση το *genderize_gender* σε ονόματα του συνόλου X με περισσότερες εγγραφές είχαν μεγαλύτερο ποσοστό επιτυχίας. Η πιο απότομη μεταβολή στην ακρίβεια των προβλέψεων παρατηρήθηκε με τη μεταβολή του t_c στο διάστημα [1,101]. Ειδικότερα, για ολόκληρο το σύνολο των αναγνωρισμένων ονομάτων (t_c=1) η ακρίβεια προβλέψεων (όπως έχουμε αναφέρει προηγουμένως) ήταν 81.55%. Για τα ονόματα με $c_x \geq 11$ (t_c=11) η ακρίβεια ήταν 85.31% ενώ για τα ονόματα με $c_x \geq 161$ (t_c=161) το αντίστοιχο ποσοστό επιτυχίας ήταν 91.55%. Για t_c=8641 όλες οι προβλέψεις με το Genderize ήταν σωστές (ποσοστό επιτυχίας 100%). Στο Σχήμα 4.6 φαίνεται η γραφική παράσταση της ακρίβειας προβλέψεων με το Genderize σε σχέση με το t_c. Όπως βλέπουμε, το πεδίο *genderize_count* παίζει μεγάλο ρόλο στην ποιότητα της ταξινόμησης των ονομάτων κατά φύλο χωρίς να λάβουμε υπόψην το *genderize_probability,* καθώς παρατηρούμε αύξηση 10% στο ποσοστό των σωστών προβλέψεων με την αύξηση του t_c από 1 σε 101. Επίσης το 100% των σωστών προβλέψεων για ονόματα με $c_x \geq 8641$ μας δείχνει ότι τα πιο δημοφιλή ονόματα δεν είναι διαμοιρασμένα και στα 2 φύλα αλλά απευθύνονται αποκλειστικά σε άντρες ή γυναίκες.

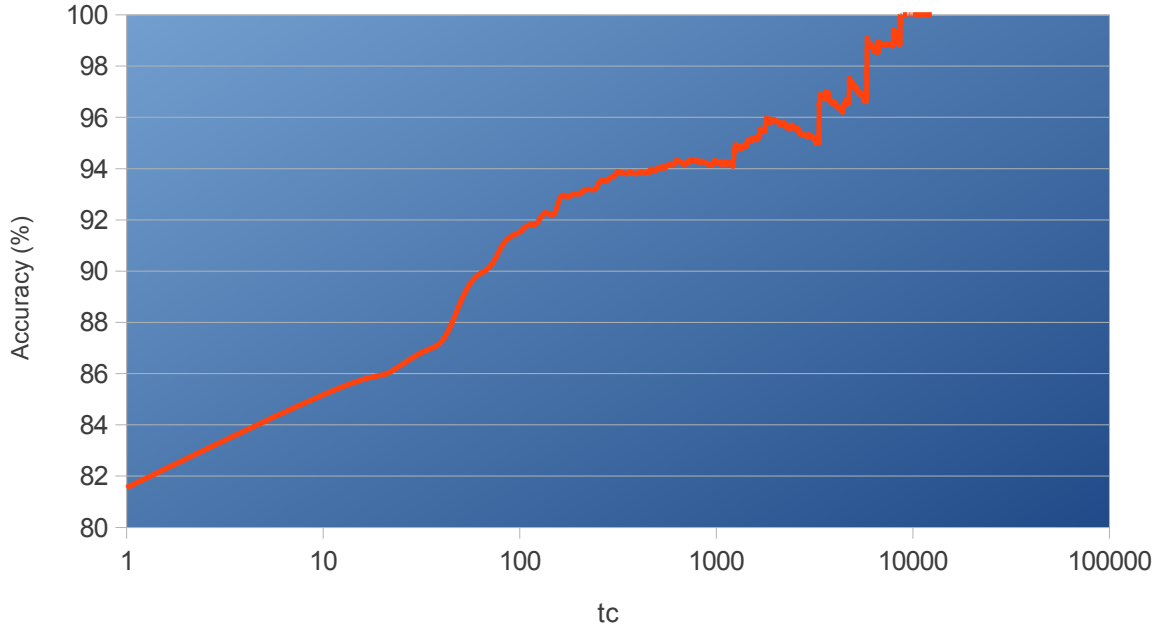
- **coverage**: Όσον αφορά το ποσοστό του συνόλου αναγνωρισμένων ονομάτων που ικανοποιούν τα διαφορετικά thresholds αριθμού εγγραφών, παρατηρούμε ότι με τη μεταβολή του t_c στο διάστημα [1,151] υπάρχει απότομη πτώση του ποσοστού. Πιο συγκεκριμένα, περίπου 1 στα 5 ονόματα (21.12%) είχε το πολύ 10 εγγραφές, ενώ το 48.98% των ονομάτων είχε το πολύ 150 εγγραφές. Εξετάζοντας για μεγαλύτερο t_c παρατηρούμε ότι το 10% των ονομάτων είχαν περισσότερες από 3750 εγγραφές και μόλις το 0.82% περισσότερες από 10000 εγγραφές. Στο Σχήμα 4.7 φαίνεται η γραφική παράσταση του ποσοστού των αναγνωρισμένων ονομάτων σε σχέση με το threshold αριθμού εγγραφών t_c. Από τα παραπάνω συμπεραίνουμε ότι σημαντικό ποσοστό από τα ονόματα που επιλέγουν οι χρήστες των κοινωνικών δικτύων δεν αντιπροσωπεύεται από αξιόπιστο αριθμό εγγραφών στις βάσεις του Genderize.

- ***genderize_probability***: Κατ' αναλογία με τη μέθοδο που ακολουθήσαμε για το *genderize_count*, μελετήσαμε την επίδραση της πιθανότητας σωστής ανίχνευσης φύλου που δίνει το Genderize για κάθε όνομα. Ορίσαμε ένα threshold πιθανότητας t_p για να δούμε πώς επηρεάζεται η ποιότητα των προβλέψεων με βάση το *genderize_gender*, κρατώντας μόνο τα ονόματα που ικανοποιούν αυτό το threshold. Επιπλέον μετρήσαμε και το σχετικό πλήθος τους σε μορφή ποσοστού του συνόλου X. Επειδή για κάθε όνομα x είναι $p_x \in [0.5,1]$, δώσαμε στο t_p τιμές $t_p = \{0.5 + i * 0.05 : i = 0,\ldots,10\}$ και κάθε φορά λαμβάνονταν υπόψην μόνο οι χρήστες που τα ονόματα τους είχαν $p_x \geq t_p$. Οι επιδράσεις που παρατηρήσαμε στα accuracy και coverage με τη μεταβολή του t_p ήταν οι εξής:

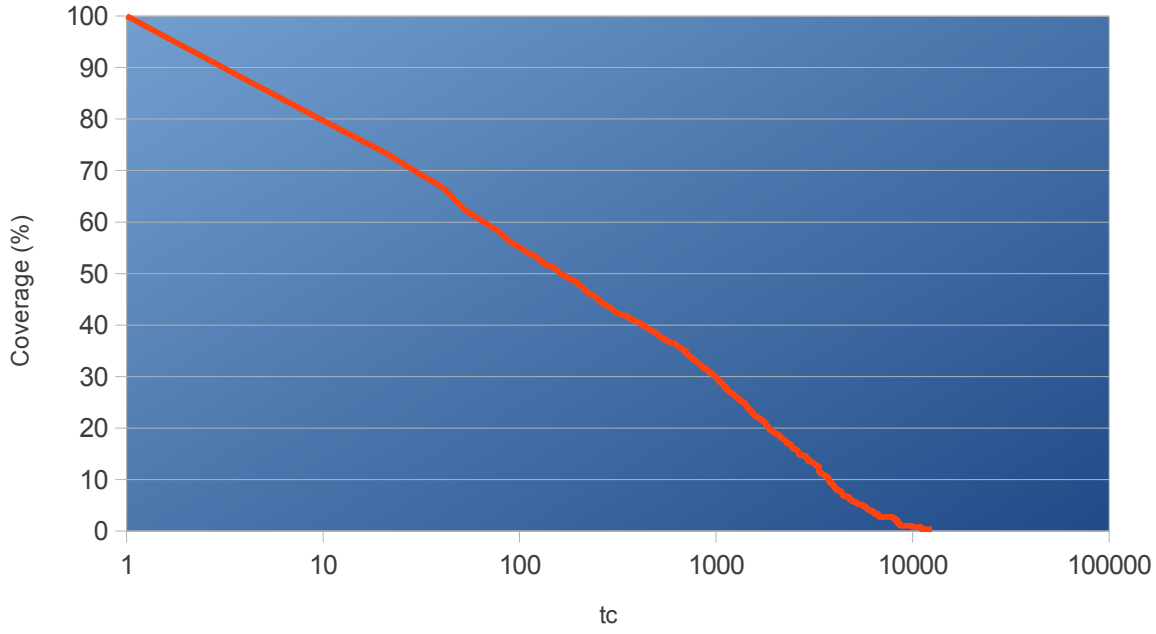
  - **accuracy**: Σχετικά με την επίδραση της πιθανότητας σωστής ανίχνευσης σύμφωνα με το Genderize, στο ποσοστό επιτυχίας των προβλέψεων παρατηρήσαμε 2 πράγματα. Πρώτον, το μέγιστο της ακρίβειας των προβλέψεων για όλα τα thresholds t_p ήταν 89.87% γεγονός το οποίο δείχνει ότι το *genderize_probability* από μόνο του δεν μπορεί να εγγυηθεί τόσο αξιόπιστες προβλέψεις όσο το *genderize_count*. Δεύτερον, για t_p=1 έχουμε μικρότερη ακρίβεια προβλέψεων από ότι για t_p=0.95. Όπως θα εξηγηθεί αναλυτικότερα στη συνέχεια, αυτή η πτώση στην απόδοση για t_p=1 οφείλεται στα ονόματα που έχουν μόνο μια

εγγραφή στο Genderize. Στο Σχήμα 4.8 φαίνεται η γραφική παράσταση της ακρίβειας προβλέψεων του Genderize σε σχέση με το threshold πιθανότητας σωστής ανίχνευσης t_p.

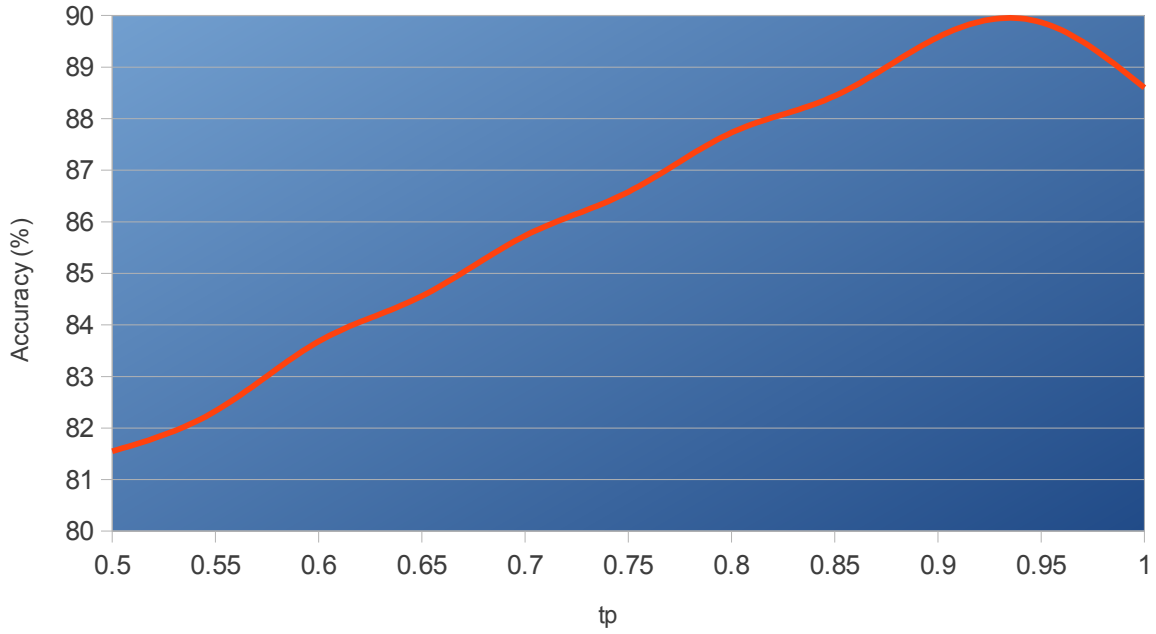
- **coverage**: Για το ποσοστό των αναγνωρισμένων ονομάτων σε σχέση με το *genderize_probability*, παρατηρούμε ότι περίπου 3 στα 4 ονόματα (74.09%) έχουν $p_x \geq 0.9$ και για τα μισά περίπου ονόματα (48.08%) το Genderize είναι σίγουρο για την εκτίμηση φύλου ($p_x = 1$). Στο Σχήμα 4.9 φαίνεται η γραφική παράσταση του ποσοστού των αναγνωρισμένων ονομάτων σε σχέση με το threshold πιθανότητας σωστής ανίχνευσης t_p. Από τα παραπάνω συμπεραίνουμε ότι το Genderize περιέχει στις βάσεις δεδομένων του, κατά κύριο λόγο ονόματα τα οποία κατά τη διάρκεια συλλογής τους από τα κοινωνικά δίκτυα απαντήθηκαν στη συντριπτική πλειοψηφία τους σε χρήστες συγκεκριμένου φύλου.



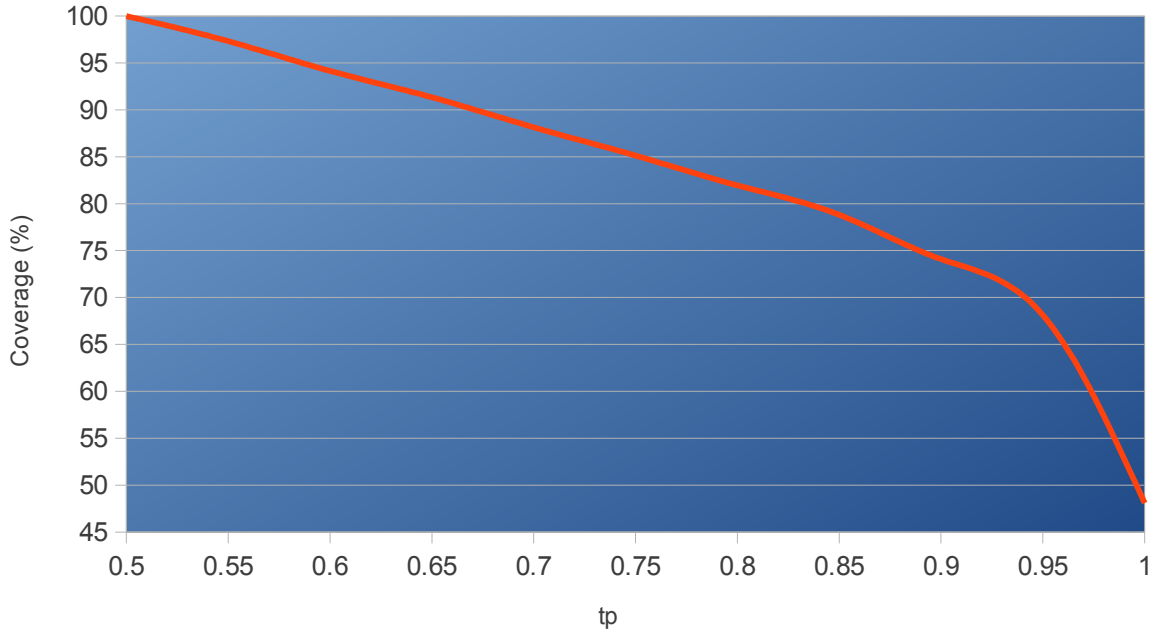
Σχήμα 4.6: Γραφική παράσταση του ποσοστού accuracy με το Genderize σε σχέση με το count threshold



Σχήμα 4.7: Γραφική παράσταση του ποσοστού coverage με το Genderize σε σχέση με το count threshold



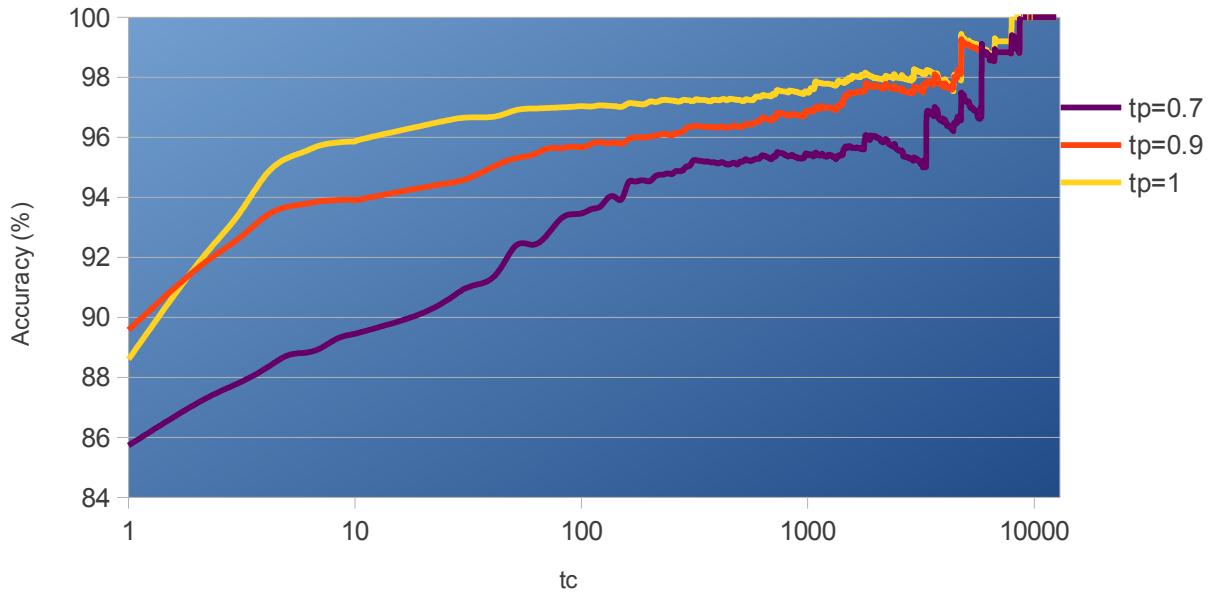
Σχήμα 4.8: Γραφική παράσταση του ποσοστού accuracy με το Genderize σε σχέση με το probability threshold



Σχήμα 4.9: Γραφική παράσταση του ποσοστού coverage με το Genderize σε σχέση με το probability threshold

Θέλοντας να μελετήσουμε την επίδραση αμφότερων των πεδίων *genderize_count* και *genderize_probability* στην ακρίβεια προβλέψεων φύλου με βάση το *genderize_gender*, μετρήσαμε το ποσοστό των σωστών προβλέψεων για μεταβαλλόμενο count threshold t_c στο διάστημα [1,12593] με επιλεγμένες σταθερές τιμές του probability threshold t_p. Στο Σχήμα 4.10, φαίνονται οι γραφικές παραστάσεις του accuracy για $t_p = \{0.7, 0.9, 1\}$ σε σχέση με το t_c (οι τιμές του t_p επιλέχτηκαν ενδεικτικά). Όπως βλέπουμε, το probability threshold t_p παίζει σημαντικό ρόλο στην ακρίβεια των προβλέψεων όταν συμπεριλάβουμε στα αποτελέσματα μας ονόματα με σχετικά μικρό αριθμό εγγραφών. Ωστόσο όσο το count threshold που επιβάλλουμε στα ονόματα αυξάνεται, τόσο η σημασία του t_p στην ποιότητα των προβλέψεων αμβλύνεται και καταλήγει να έχει μηδαμινή επίδραση για μεγάλες τιμές του t_c. Επίσης παρατηρούμε ότι για όλα τα ονόματα του συνόλου X ($t_c = 1$), όπως διατυπώσαμε και προηγουμένως, για t_p=0.9 έχουμε καλύτερη απόδοση από ότι με t_p=1. Αυτό οφείλεται στο ότι στα ονόματα που έχουν p_x=1 περιλαμβάνονται και αυτά που έχουν c_x=1 τα οποία δεν παρέχουν πολύ αξιόπιστη πληροφορία (προφανώς το Genderize τους δίνει πιθανότητα σωστής ανίχνευσης 100%). Βρήκαμε ότι τα ονόματα αυτά αποτελούν το 6.07% του συνόλου αναγνωρισμένων ονομάτων X και έχουν ποσοστό σωστών εκτιμήσεων

φύλου 66.84%. Όταν θέτουμε t_p=1 το ποσοστό των ονομάτων που παίρνουμε υπόψην μας για τα αποτελέσματα μικραίνει (48.08% του X) και άρα τα ονόματα με c_x=1 παίζουν μεγαλύτερο ρόλο στην ολική ακρίβεια προβλέψεων, δυσχεραίνοντας την περισσότερο σε σχέση με την περίπτωση του t_p=0.9 το οποίο περιλαμβάνει μεγαλύτερη γκάμα ονομάτων (74.09% του X). Αντίθετα, για $t_c \geq 2$ το threshold πιθανότητας t_p=1 παρουσιάζει την καλύτερη απόδοση όπως θα περιμέναμε.



Σχήμα 4.10: Γραφική παράσταση του ποσοστού accuracy με το Genderize σε σχέση με το count threshold για διαφορετικές τιμές του probability threshold

Η επιπλέον πληροφορία που δίνουν τα πεδία *genderize_count* και *genderize_probability* θα αξιοποιηθεί στο Κεφάλαιο 5 με χρήση αλγορίθμων μηχανικής μάθησης.

4.4 Ανίχνευση φύλου μέσω του Χρώματος Θέματος

4.4.1 Περιγραφή χρώματος θέματος

Όπως έχουμε αναφέρει, στην παρούσα διπλωματική εργασία κάναμε χρήση του χρώματος θέματος των προφίλ χρηστών του Twitter για την ανίχνευση του φύλου τους. Την τιμή του χρώματος αυτού την αποκτήσαμε μέσω του API του Twitter, παίρνοντας την τιμή που μας επιστράφηκε για το προγενέστερο πεδίο *profile link*

color, αφού παρατηρήσαμε ότι οι 2 αυτές τιμές ταυτίζονταν.

Στο πεδίο *theme color* της βάσης δεδομένων μας, είναι αποθηκευμένος για κάθε χρήστη ένας 6-ψήφιος δεκαεξαδικός αριθμός τον οποίο αποκτήσαμε από τα responses του Tweepy και αντιστοιχεί στο χρώμα που έχει επιλέξει ο χρήστης για τη διαμρφωση του προφίλ του. Ο αριθμός αυτός εκφράζει κάποιο χρώμα σε RGB μορφή[24] με τα πρώτα 2 ψηφία να αποτελούν το Red value, τα επόμενα 2 το Green value και τα 2 τελευταία το Blue value. Μερικά παραδείγματα από ζεύγη δεκαεξαδικών αριθμών και RGB αναπαραστάσεων για δημοφιλή χρώματα φαίνονται στο Σχήμα 4.11.

#9400D3	RGB 148, 0, 211
#4B0082	RGB 75, 0, 130
#0000FF	RGB 0, 0, 255
#00FF00	RGB 0, 255, 0
#FFFF00	RGB 255, 255, 0
#FF7F00	RGB 255, 127, 0
#FF0000	RGB 255, 0 , 0

Σχήμα 4.11: Δεκαεξαδικές και RGB αναπαραστάσεις για δημοφιλή χρώματα

4.4.2 Διεξαγωγή πειραμάτων με το χρώμα θέματος

Για τον προσδιορισμό του φύλου των χρηστών του Twitter με βάση το *χρώμα θέματος* του προφίλ τους, πραγματοποιήσαμε πειράματα με αλγορίθμους μηχανικής μάθησης. Πιο συγκεκριμένα, δοκιμάσαμε ως ταξινομητές φύλου τα μοντέλα επιβλεπόμενης μάθησης: Gaussian Naive Bayes, Multinomial Naive Bayes, Support Vector Machines (με RBF, γραμμικό, πολυωνυμικό και σιγμοειδή πυρήνα)

[24] https://en.wikipedia.org/wiki/RGB_color_model

και Probabilistic Neural Networks. Στο υπόλοιπο της Ενότητας ένας ταξινομητής φύλου με βάση το χρώμα θέματος θα αναφέρεται και ως *Color Algorithm*.

4.4.2.1 Είσοδος χρώματος στους αλγορίθμους

Για να τροφοδοτήσουμε το theme color του κάθε χρήστη σαν διάνυσμα χαρακτηριστικών στους αλγορίθμους μηχανικής μάθησης, χρησιμοποιήσαμε τις εξής 2 διαφορετικές μεθόδους:

- **Χωρίς προεπεξεργασία (Normal)**: Σε αυτή τη μέθοδο δίνουμε σαν διάνυσμα χαρακτηριστικών στους ταξινομητές μηχανικής μάθησης 3 ακέραιους αριθμούς οι οποίοι αντιστοιχούν στα Red, Green & Blue values του κάθε χρώματος. Όπως εξηγήσαμε και προηγουμένως, κάθε value αντιστοιχεί σε δυο δεκαεξαδικά ψηφία που συνεπάγονται εύρος ακεραίων 0-255. Για παράδειγμα για έναν χρήστη που έχει theme color: #4286f4 θα δοθούν οι τιμές (66, 134, 244) σαν διάνυσμα χαρακτηριστικών στον ταξινομητή.

- **Color Quantization & Sorting**: Με βάση την δουλειά των Alowibdi, Buy & Yu [2] δοκιμάσαμε την προεπεξεργασία των theme colors που πήραμε από την βάση δεδομένων μας για να αποφανθούμε αν αυτή η προεπεξεργασία θα οδηγήσει σε καλύτερη απόδοση ανίχνευσης φύλου όπως αναφέρουν οι προηγούμενοι συγγραφείς. Αρχικά χρησιμοποιήσαμε την τεχνική color quantization. Το color quantization είναι μια μορφή συμπίεσης που μειώνει τον τεράστιο αριθμό των διαφορετικών χρωμάτων. Κάθε μια Red, Green και Blue τιμή μειώνεται από τα 8 bits στα 3 bits και έτσι ο συνολικός αριθμός χρωμάτων αλλάζει από $256^3 \approx 16 * 10^6$ σε $8^3 = 512$ χρώματα. Κάθε αρχικό χρώμα μετατράπηκε στο συμπιεσμένο χρώμα από το οποίο είχε την μικρότερη Ευκλείδια απόσταση. Στη συνέχεια τα συμπιεσμένα χρώματα μετατράπηκαν από την RGB μορφή τους σε μορφή HSV (Hue,Saturation,Value)[25] και ταξινομήθηκαν πρωτίστως κατά Hue και δευτερευόντως κατά Value. Τέλος στα ταξινομημένα χρώματα ανατέθηκαν διαδοχικές ακέραιες αριθμητικές ταμπέλες οι οποίες αποτέλεσαν τα διανύσματα χαρακτηριστικών που τροφοδοτήθηκαν στους ταξινομητές . Για παράδειγμα, το αρχικό χρώμα (R:190,G:45,B:120) μετατράπηκε στο

[25] https://en.wikipedia.org/wiki/HSL_and_HSV

συμπιεσμένο χρώμα (R:182,G:36,B:109) το οποίο αντιστοιχεί σε (H:330^{o},S:80.2%,V:71.4%). Η θέση αυτής της HSV αναπαράστασης στην ταξινομημένη λίστα είναι η “ταμπέλα” που αποτέλεσε το διάνυσμα χαρακτηριστικών για τους χρήστες με το σχετικό αρχικό χρώμα. Η ιδέα πίσω από αυτή διαδικασία είναι ότι ίσως ένας Color Algorithm θα μπορεί να διαχωρίσει καλύτερα τα δυο φύλα παίρνοντας την επιπλέον πληροφορία των “κοντινών” χρωμάτων που έχουν διαλέξει οι χρήστες μεταξύ τους.

Στη συνέχεια του κειμένου οι 2 μέθοδοι θα αναφέρονται ως *Normal* και *Q&S.*

Σημειώνεται ότι για καμία από τις δυο μεθόδους δεν πραγματοποιήσαμε κανονικοποίηση των διανυσμάτων χαρακτηριστικών αφού για τη μέθοδο *Normal* οι R,G,B τιμές έχουν όλες εύρος [0-255] και για τη μέθοδο Q&S έχουμε μόνο έναν ακέραιο αριθμό ως χαρακτηριστικό των δειγμάτων μας.

Στη συνέχεια περιγράφουμε τα πειράματα που πραγματοποιήσαμε με βάση το theme color των χρηστών της βάσης δεδομένων μας αναφέροντας τα σχετικά αποτελέσματα.

4.4.2.2 Μετρήσεις απόδοσης διαφορετικών αλγορίθμων

Το κάθε μοντέλο επιβλεπόμενης μάθησης, κατά τη διαδικάσια της εκπαίδευσης δέχεται για κάθε χρήστη του συνόλου εκπαίδευσης, ένα διάνυσμα χαρακτηριστικών μαζί με την επιθυμητή έξοδο που αντιστοιχεί στο φύλο του. Τα διανύσματα χαρακτηριστικών είναι βασισμένα στο *theme color* με τη μέθοδο *Normal* ή *Q&S* και η επιθυμητή έξοδος έχει δυαδική τιμή (ο αριθμός εξόδου 1 αντιστοιχεί στο αντρικό και ο αριθμός 0 στο γυναικείο φύλο).

Μετά τη διαδικασία της εκπαίδευσης, το κάθε μοντέλο έχει την δυνατότητα να προβλέπει το φύλο ενός νέου χρήστη βάσει του αντίστοιχου διανύσματος χαρακτηριστικών του. Επιπλέον, υπάρχει η δυνατότητα εξαγωγής των πιθανοτήτων που αποδίδει το μοντέλο στο να ανήκει ο χρήστης σε κάθε φύλο. Για παράδειγμα για κάποιον χρήστη μπορεί να δοθεί η πιθανότητα 63% να είναι γυναίκα και 37% να είναι άντρας. Σε αυτό το στάδιο της μελέτης, δεν κάναμε χρήση των πιθανοτήτων που εξάγονται από τα μοντέλα επιβλεπόμενης μάθησης.

Η αξιολόγηση της απόδοσης των μοντέλων επιβλεπόμενης μάθησης έγινε με την

εξαγωγή της μετρικής *accuracy* μέσω της μεθόδου stratified 5-fold cross-validation.

Για την βελτιστοποίηση της απόδοσης των Μηχανών Διανυσμάτων Υποστήριξης και των Πιθανολογικών Νευρωνικών Δικτύων, αναζητήσαμε τις κατάλληλες παραμέτρους του κάθε μοντέλου προκειμένου να βρούμε τις τιμές τους που μεγιστοποιούν το accuracy. Από τις επιλογές που μας έδιναν οι βιβλιοθήκες scikit-learn και nuepy για τη ρύθμιση παραμέτρων, πειραματιστήκαμε με τις εξής:

- Για τα Support Vector Machines με πυρήνα RBF, σιγμοειδή ή πολυωνυμικό εξετάσαμε τις τιμές των υπερπαραμέτρων *gamma* και *C*. Για τον πολυωνυμικό πυρήνα, πειραματιστήκαμε επιπλέον με τον βαθμό του πολυωνύμου.

- Για το Support Vector Machine με γραμμικό πυρήνα, εξετάσαμε τις τιμές της παραμέτρου *C* σε συνδυασμό με δυο μορφές της συνάρτησης απώλειας, την βασιζόμενη σε συντελεστή εξάρτησης (hinge) και τη βασιζόμενη στο τετράγωνο του συντελεστή εξάρτησης (squared hinge).

- Για το Probabilistic Neural Network δοκιμάσαμε διαφορετικές τιμές για την τυπική απόκλιση *σ*.

Στο σημείο αυτό αναφέρουμε τη διαδικασία εύρεσης των υπερπαραμέτρων *C* και *gamma* για τον ταξινομητή SVM με RBF πυρήνα. Η μέθοδος που ακολουθήσαμε είναι η αναζήτηση πλέγματος (βλ. Κεφαλαιο 2.2.3.3). Για την υπερπαράμετρο *C* πραγματοποιήσαμε αναζήτηση στο εύρος τιμών $\left\{10^i : i \in \mathbb{Z}, -2 \le i \le 3\right\}$, ενώ για την *gamma* στο εύρος $\left\{10^i : i \in \mathbb{Z}, -9 \le i \le 3\right\}$. Τα accuracies που αντιστοιχούν στα αποτελέσματα που πήραμε με τη μέθοδο stratified 5-fold cross-validation για κάθε δυνατό συνδυασμό παρουσιάζονται στον Πίνακα 4.1. Όπως βλέπουμε το βέλτιστο accuracy επιτεύχθει για C=100 και gamma=10^{-4}.

gamma \ *C*	0.01	0.1	1	10	100	1000
10^{-9}	61.19%	61.19%	61.19%	61.19%	61.19%	61.19%
10^{-8}	61.19%	61.19%	61.19%	61.19%	61.19%	61.19%
10^{-7}	61.19%	61.19%	61.19%	61.19%	61.19%	61.56%
10^{-6}	61.19%	61.19%	61.19%	61.54%	63.87%	63.97%
10^{-5}	61.19%	60.96%	63.87%	64.19%	64.26%	64.16%
10^{-4}	60.95%	63.83%	64.69%	65.52%	**65.67%**	65.37%
10^{-3}	63.15%	65.28%	65.58%	64.52%	64.03%	63.36%
10^{-2}	63.02%	65.41%	65.46%	64.68%	64.58%	64.4%
10^{-1}	63.02%	65.42%	65.48%	65.3%	65.3%	65.3%
1	63.02%	65.42%	65.33%	65.31%	65.31%	65.31%
10	63.02%	65.42%	65.3%	65.3%	65.3%	65.3%
100	63.02%	65.42%	65.3%	65.3%	65.3%	65.3%
1000	63.02%	65.42%	65.3%	65.3%	65.3%	65.3%

Πίνακας 4.1: Αναζήτηση πλέγματος για SVM με RBF kernel

Στον Πίνακα 4.2 φαίνονται συγκεντρωμένα όλα τα αποτελέσματα των διαφορετικών αλγορίθμων μηχανικής μάθησης που χρησιμοποίησαμε ως Color Algorithm για την ανίχνευση του φύλου των χρηστών του Twitter με βάση το χρώμα θέματος του προφίλ τους, με τις μεθόδους *Normal* και *Q&S*. Για κάθε αλγόριθμο αναφέρουμε την βέλτιστη απόδοση που πετύχαμε με δοκιμές των ανάλογων παραμέτρων. Όλα τα accuracies εξήχθησαν με τη μέθοδο stratified 5-fold cross-validation.

Όπως βλέπουμε, την καλύτερη απόδοση παρουσίασε η ταξινόμηση των χρηστών με χρήση Μηχανής Διανυσμάτων Υποστήριξης με πυρήνα RBF χωρίς την προεπεξεργασία των χρωμάτων, με 65.58% accuracy. Το Πιθανολογικό Νευρωνικό Δίκτυο είχε αρκετά κοντινό accuracy με 65.16%, επίσης χωρίς color quantization & sorting.

	Normal	Q&S
GNB	61.86%	61.19%
MNB	61.6%	61.19%
SVM-RBF	**65.67%**	64.64%
SVM-Sigmoid	61.21%	61.19%
SVM-Polynomial	62.7%	-[26]
SVM-Linear	61.1%	61.19%
PNN	65.16%	62.31%

Πίνακας 4.2: Accuracies για όλα τους υποψήφιους Color Algorithms με τις μεθόδους Normal και Q&S

4.4.3 Συμπεράσματα

Από τα παραπάνω αποτελέσματα φαίνεται ότι το *theme color* δεν είναι ικανό χαρατηριστικό για να προσδιορίσει το φύλο ενός χρήστη του Twitter από μόνο του. Το υψηλότερο accuracy που πετύχαμε μέσω ενός Color Algorithm ήταν 65.67% που είναι ελάχιστα καλύτερο από το 61.19% του *baseline gender classifier*. Παρατηρούμε επίσης από την δοκιμή διαφορετικών υλοποιήσεων του Color Algorithm και τον πειραματισμό με διαφορετικές τιμές των παραμέτρων τους, ότι φαίνεται να υπάρχει ένα threshold για το accuracy το οποίο δεν φτάνει σε καμία περίπτωση το 66%.

Όσον αφορά την τεχνική που χρησιμοποιήσαμε για τη μετατροπή των χρωμάτων από 24-bit RGB αναπαράσταση σε 8-bit RGB και την μετέπειτα ταξινόμηση τους σύμφωνα με την HSV αναπαράσταση τους, παρατηρήσαμε γενικά χειρότερα αποτελέσματα στην ανίχνευση φύλου σε σχέση με την περίπτωση της μη προεπεξεργασίας. Αυτό έρχεται σε αντίθεση με τα αποτελέσματα των Alowibdi, Buy & Yu [2] που είχαν παρατηρήσει 13% αύξηση στο accuracy με χρήση παρόμοιας τεχνικής. Επιπλέον το accuracy που πετύχαμε χρησιμοποιώντας μόνο το *theme color* είναι πολύ χαμηλότερο από το 71.6% που σημείωσαν οι προαναφερθέντες με τη χρήση Πιθανολογικού Νευρωνικού Δικτύου. Όπως αναφέραμε και στην Ενότητα 3.2, οι Alowibdi, Buy & Yu είχανε χρησιμοποιήσει 5 χρώματα που αντιστοιχούσαν σε 5 διαφορετικά σημεία του προφίλ ενός χρήστη, ενώ πλέον οι χρήστες επιλέγουν μόνο ένα theme color. Φαίνεται ότι η επιλογή του

[26] Ο πολυωνυμικός πυρήνας δεν δοκιμάστηκε για τη μέθοδο Q&S καθώς έχουμε μόνο ένα χαρακτηριστικό εισόδου.

Twitter να αφήσει στους χρήστες του την ελευθερία να επιλέγουν μόνο ένα ενιαίο χρώμα για το προφίλ τους, είχε άμεσο αντίκτυπο στην πληροφορία που μπορεί να εξαχθεί για το φύλο των χρηστών από τον χρωματισμό του προφίλ τους. Περισσότερες λεπτομέρειες παρουσιάζονται στην επόμενη Ενότητα.

4.4.4 Στατιστικά επιλογής χρωμάτων

Θέλοντας να μελετήσουμε την αιτία της χαμηλής απόδοσης του *theme color* ως χαρακτηριστικό προσδιορισμού του φύλου των χρηστών αλλά και από ερευνητικό ενδιαφέρον, πραγματοποίησαμε στατιστική ανάλυση των χρωμάτων που επιλέγουν τα 2 φύλα. Αρχικά μετρήσαμε τους χρήστες των δυο φύλων που έχουν στο προφίλ τους το default theme color του Twitter (R:29, G:161, B:242). Τα ποσοστά για τους άντρες και τις γυναίκες της βάσης μας ήταν 19.98% και 10.85% αντίστοιχα. Αυτό μας δείχνει ότι οι γυναίκες χρήστες του Twitter είναι πιο πιθανό σε σχέση με τους άντρες να ασχοληθούν με μια χρωματική διαμόρφωση του προφίλ τους που εκφράζει το προσωπικό τους στυλ.

Στη συνέχεια, εντοπίσαμε τα 10 πιο δημοφιλή χρώματα για το κάθε φύλο ξεχωριστά και μετρήσαμε το ποσοστό επιλογής του κάθε χρώματος ως προς το σύνολο των χρηστών του αντίστοιχου φύλου. Τα δημοφιλή χρώματα προσδιορίσθηκαν με color quantization (3-bit RGB τιμές) για να ομαδοποιήσουμε τις κοντινές αποχρώσεις με σκοπό την καλύτερη εποπτεία. Τα αποτελέσματα φαίνονται στον Πίνακα 4.3. Παρατηρούμε αρχικά ότι το πιο δημοφιλές χρώμα και για τα 2 φύλα είναι το συμπιεσμένο χρώμα που είναι πιο “κοντά” στο default theme color με τιμή (R:0, G:145, B:218). Αυτό ήταν αναμενόμενο με βάση τα ποσοστά που αναφέραμε πιο πάνω. Επιπλέον, και το δεύτερο πιο δημοφιλές χρώμα είναι κοινό για τα 2 φύλα. Εξετάζοντας τα υπόλοιπα χρώματα των δυο λιστών, βλέπουμε ότι τα πιο πολλά χρώματα είναι κοινά στις δυο λίστες εκτός από μια απόχρωση του γκρι και μια απόχρωση του γαλάζιου που υπάρχουν αποκλειστικά στους αντρες και 2 αποχρώσεων του ροζ που υπάρχουν αποκλειστικά στις γυναίκες. Με βάση αυτά τα στοιχεία, βγάζουμε το συμπέρασμα ότι η στερεοτυπική φράση “Οι άντρες προτιμούν το μπλε και οι γυναίκες το ροζ” φαίνεται να μην ισχύει στο Twitter.

Συγκρίνοντας τον Πίνακα 4.3 με την αντίστοιχη κατανομή χρωμάτων που κατέγραψαν οι Alowibdi, Buy & Yu, παρατηρούμε ότι με την ελευθερία επιλογής 5 διακριτών χρωμάτων για το κάθε προφίλ, υπήρχε εμφανής διαφορά στις χρωματικές επιλογές των 2 φύλων ενώ πλέον με την μοναδική επιλογή του

χρώματος θέματος οι διαφορές αυτές έχουν αμβλυνθεί θεαματικά. Επιβεβαιώνουμε λοιπόν ότι η διαφορά του accuracy που πετύχαμε στην ανίχνευση φύλου των χρηστών του Twitter σε σχέση με το accuracy που επιτεύχθει στη μελέτη [2], οφείλεται σε μεγάλο βαθμό στην αλλαγή που επέβαλε το Twitter σχετικά με την επιλογή χρωμάτων από τους χρήστες του.

Άντρες	Γυναίκες
21.90%	11.61%
17.20%	10.29%
10.44%	6.18%
4.84%	4.90%
4.02%	4.28%
3.10%	3.63%
1.97%	3.61%
1.97%	3.61%
1.70%	2.55%
1.64%	2.26%

Πίνακας 4.3: Τα 10 πιο δημοφιλή χρώματα για τα 2 φύλα

Τέλος, μετρήσαμε τον αριθμό των διαφορετικών χρωμάτων (χωρίς quantization) που επέλεξαν οι χρήστες της βάσης δεδομένων μας. Οι άντρες επέλεξαν 761 και οι γυναίκες 1724 διαφορετικά χρώματα. Παρατηρείται λοιπόν μεγάλη διασπορά των χρωμάτων που επιλέγονται από τα 2 φύλα. Αυτό το γεγονός είναι ένας επιπλέον παράγοντας της μικρής απόδοσης που πετύχαμε, καθώς η μη μαζική επιλογή συγκεκριμένων χρωμάτων από κάθε φύλο, δυσχεραίνει την ανίχνευση του.

4.5 Τελικά Συμπεράσματα

Έχοντας ολοκληρώσει τα πειράματα μας για την ανίχνευση φύλου των χρηστών του Twitter με χρήση της *profile picture*, του *display name* και του *theme color* ξεχωριστά, βγάλαμε αρκετά χρήσιμα συμπεράσματα όσον αφορά τον βαθμό με τον οποίο το κάθε ένα από τα τρία πεδία βοηθάει στον χαρακτηρισμό του φύλου ενός χρήστη. Επιπλέον κάναμε μερικές ενδιαφέρουσες παρατηρήσεις σχετικά με τις διαφορές που έχει η χρησιμοποιήση του κάθε πεδίου ανάμεσα στα 2 φύλα.

Σύμφωνα με τα αποτελέσματα μας, καμία από τις τρεις προσεγγίσεις που ακολουθήσαμε δεν ήταν ικανή να οδηγήσει σε ταξινόμηση των χρηστών της βάσης μας κατά φύλο με αξιόλογο ποσοστό επιτυχίας. Για το λόγο αυτό επιχειρήσαμε να εισάγουμε αλγορίθμους μηχανικής μάθησης κατά την χρησιμοποίηση της *profile picture* και του *display name* και κατόπιν να συνδυάσουμε τις τρεις αμιγείς προσεγγίσεις σε έναν υβριδικό αλγόριθμο μηχανικής μάθησης, έχοντας ως στόχο την επίτευξη ενός ικανοποιητικού ποσοστού ακρίβειας στις προβλέψεις φύλου. Στην υβριδική προσέγγιση, δεν πραγματοποιείται color quantization & sorting, καθώς όπως αναφέραμε αυτή η μέθοδος μας έδωσε χειρότερα αποτελέσματα.

Κεφάλαιο 5

Προτεινόμενη Υβριδική Προσέγγιση και Αξιολόγηση

5.1 Χρησιμότητα Υβριδικού Αλγορίθμου

Ένας Υβριδικός Αλγόριθμος συνδυάζει δύο ή περισσότερους αλγορίθμους οι οποίοι επιλύουν το ίδιο πρόβλημα, είτε διαλέγοντας τον καλύτερο από αυτούς ανάλογα με την περίπτωση ή συνδυάζοντας τα ξεχωριστά αποτελέσματα τους για την εξαγωγή ενός ενιαίου αποτελέσματος. Μια συνήθης περίπτωση είναι οι επι μέρους αλγόριθμοι να διαφέρουν ως προς την απόδοση τους ανάλογα με τις ξεχωριστές περιπτώσεις του προβλήματος που καλούνται να επιλύσουν. Η όλη διαδικασία σκοπεύει στον συνδυασμό των επιθυμητών χαρακτηριστικών των επι μέρους αλγορίθμων έτσι ώστε ο Υβριδικός Αλγόριθμος να έχει καλύτερη απόδοση από τις ξεχωριστές συνιστώσες του.

Ειδικότερα για τους σκοπούς της παρούσας εργασίας, επιχειρήσαμε να συνδυάσουμε τις προβλέψεις τριών αλγορίθμων μηχανικής μάθησης οι οποίοι ταξινομούν κατά φύλο τους χρήστες του Twitter βασισμένοι ξεχωριστά ο καθένας στην φωτογραφία προφίλ, στο όνομα και στο χρώμα. Η λογική πίσω από αυτό το έγχειρημα είναι ότι συνδυάζοντας τα αποτελέσματα των τριών μεθόδων ανίχνευσης του φύλου θα μπορέσουμε να εκμεταλλευτούμε τις περιπτώσεις χρηστών για τις οποίες μια ή δυο από τις μέθοδους μπορεί να μην παράγουν αξιόπιστες προβλέψεις, λαμβάνοντας υπόψην τις προβλέψεις των μεθόδων που είναι πιο “σίγουρες” για τα αποτελέσματα τους. Με την υβριδική προσέγγιση, οι δυνάμεις μιας μεθόδου συμπληρώνουν τις αδυναμίες της άλλης και με αυτό τον τρόπο η υβριδική μέθοδος γενικεύει πολύ καλύτερα στις διάφορες περιπτώσεις χρηστών με υψηλότερη συνολική ποιότητα προβλέψεων φύλου για όλους τους χρήστες.

Στη συνέχεια, περιγράφουμε την διαδικασία που ακολουθήσαμε για την υλοποίηση του υβριδικού αλγορίθμου μηχανικής μάθησης και παρουσιάζουμε τα αποτελέσματα μας. Στο τέλος διατυπώνουμε τα συμπεράσματα μας.

5.2 Υλοποίηση Υβριδικής Προσέγγισης

Για να συνδυάσουμε τις προβλέψεις φύλου που γίνονται για τον κάθε χρήστη με βάση την φωτογραφία προφίλ, το όνομα και το χρώμα με σκοπό να παράξουμε μια τελική υβριδική ανίχνευση φύλου, επιχειρήσαμε να βρούμε έναν ομοιογενή τρόπο για να παίρνουμε κάποιον παράγοντα βεβαιότητας ότι ο χρήστης ανήκει σε κάποιο φύλο με βάση το κάθε πεδίο ξεχωριστά.

Όπως περιγράψαμε στο Κεφάλαιο 4, οι διαδικασίες ανίχνευσης του φύλου των χρηστών που ακολουθήσαμε με βάση την φωτογραφία προφίλ, το όνομα και το χρώμα θέματος είναι οι εξης:

- Για την χρησιμοποίηση της φωτογραφίας προφίλ, στείλαμε στο Face Detection API του Face++ το URL της φωτογραφίας και με βάση την επεξεργασία της απάντησης του Face++ αποθηκεύσαμε στη βάση δεδομένων μας το πεδίο *faceplusplus_gender* για κάθε χρήστη το οποίο έχει πιθανές τιμές "Άντρας", "Γυναίκα" ή "Άγνωστο". Με βάση αυτές τις τιμές κάναμε "πρόβλεψη" του φύλου κάθε χρήστη.

- Για την χρησιμοποίηση του ονόματος, στείλαμε στο API του Genderize κάθε λέξη που περιέχεται στη συμβολοσειρά του display name μέχρι να βρούμε κάποια που αντιστοιχεί σε αναγνωρισμένο όνομα από το Genderize ή μέχρι να τελειώσουν οι λέξεις. Με βάση τις απαντήσεις του Genderize για αυτή τη διαδικασία αποθηκεύσαμε στη βάση δεδομένων μας το πεδίο genderize_*gender* για κάθε χρήστη το οποίο έχει πιθανές τιμές "Άντρας", "Γυναίκα" ή "Άγνωστο. Στις περιπτώσεις που το *genderize_gender* δεν είναι "Άγνωστο" έχουμε τα επιπλέον στοιχεία: *genderize_count* (αριθμός εγγραφών του ονόματος) και *genderize_probability* (πιθανότητα σωστής ανίχνευσης). Με βάση τις τιμές του *genderize_gender* κάναμε "πρόβλεψη" του φύλου κάθε χρήστη χωρίς να δοκιμάσουμε τη χρήση των άλλων δύο πεδίων με συστηματικό τρόπο.

- Για την χρησιμοποίηση του χρώματος, εκπαιδεύσαμε αλγορίθμους επιβλεπόμενης μάθησης με χαρακτηριστικά που αντιστοιχούν στο *theme color* των χρηστών και στη συνέχεια με βάση τους εκπαιδευμένους αλγορίθμους πήραμε για έναν χρήστη την πρόβλεψη "Άντρας" ή "Γυναίκα" μαζί με τις αντίστοιχες πιθανότητες του να ανήκει ο χρήστης σε κάθε φύλο.

Με βάση τα παραπάνω, βλέπουμε ότι οι 3 ξεχωριστές προσεγγίσεις διαφέρουν αρκετά ώστε να συνδυαστούν όπως έχουν και να παράξουν ομοιόμορφα έναν παράγοντα βεβαιότητας για την ανίχνευση φύλου ενός χρήστη. Ο πιο λογικός τρόπος για να το πετύχουμε αυτό, είναι η επέκταση των διαδικασιών ανίχνευσης του φύλου με βάση την φωτογραφία προφίλ και το όνομα, χρησιμοποιώντας αλγορίθμους επιβλεπόμενης μάθησης σε συνάφεια με τη διαδικασία που ακολουθήσαμε για το χρώμα. Με αυτό τον τρόπο αποκτούμε με ομοιογενή τρόπο την βεβαιότητα που δίνει η κάθε προσέγγιση στην ταύτιση ενός χρήστη με ένα φύλο. Στη συνέχεια οι αλγόριθμοι επιβλεπόμενης μάθησης που χρησιμοποιήθηκαν για τις ξεχωριστές προσεγγίσεις ανίχνευσης φύλου θα αναφέρονται ως *Color Algorithm*, *Name Algorithm* και *Photo Algorithm.*

Οι διαδικασίες που ακολουθήσαμε για την ενσωμάτωση αλγορίθμων μηχανικής μάθησης στις μεθόδους ανίχνευσης φύλου μέσω της φωτογραφίας προφίλ και του ονόματος είναι οι εξής:

- Για τη μέθοδο με χρήση της φωτογραφίας προφίλ, κωδικοποιήσαμε τις τιμές “Άντρας”, “Γυναίκα”, “Άγνωστο” του *faceplusplus_gender* για κάθε χρήστη με τους αριθμούς 1, -1, 0 αντίστοιχα. Ο αριθμός που αντιστοιχούσε σε κάθε χρήστη αποτέλεσε το διάνυσμα χαρακτηριστικών που εισήχθη στον Photo Algorithm.

- Για την προσέγγιση με χρήση του ονόματος, έχουμε για κάθε χρήστη με βάση τις απαντήσεις του Genderize τα στοιχεία *genderize_gender*, *genderize_count* και *genderize_probability*. Διακρίνουμε τις εξής περιπτώσεις:

    - Αν το *genderize_gender* για κάποιον χρήστη έχει τιμή “Άντρας” τότε του δίνουμε το διάνυσμα χαρακτηριστικών [1, genderize_count, genderize_ probability].

    - Αν το *genderize_gender* για κάποιον χρήστη έχει τιμή “Γυναίκα” τότε του δίνουμε το διάνυσμα χαρακτηριστικών [-1, -genderize_count, -genderize_ probability].

    - Αν το *genderize_gender* για κάποιον χρήστη έχει τιμή “'Αγνωστο” τότε του δίνουμε το διάνυσμα χαρακτηριστικών [0,0,0].

Για να αποκτήσουμε παράγοντες βεβαιότητας σωστής ανίχνευσης φύλου που υπολογίζονται με τον ίδιο τρόπο από κάθε μέθοδο, χρησιμοποίησαμε ίδιους αλγορίθμους επιβλεπόμενης μάθησης για τους Color Algorithm, Photo Algorithm και Name Algorithm. Όπως αναφέραμε και στην Ενότητα 2.3.2.1, μπορούμε με τη χρήση κατάλληλης μεθόδου να πάρουμε για κάθε αλγόριθμο επιβλεπόμενης μάθησης τις πιθανότητες που δίνει στο να ανήκει ένα ς χρήστης σε κάθε φύλο. Πιο συγκεκριμένα, ο κάθε αλγόριθμος επιβλεπόμενης μάθησης δίνει μια πιθανότητα P_m ο χρήστης να είναι άντρας και μια πιθανότητα P_f ο χρήστης να είναι γυναίκα όπου προφανώς ισχύει $P_m+P_f=1$. Θέλοντας να έχουμε μια μοναδική έξοδο για κάθε μέθοδο ανίχνευσης φύλου που να συμβολίζει την βεβαιότητα της για κάθε πρόβλεψη, υπολογίσαμε έναν δεκαδικό *αριθμό φύλου* ως έξοδο για κάθε έναν από τους Color Algorithm, Photo Algorithm και Name Algorithm. Οι *αριθμοί φύλου* υπολογίστηκαν ως εξής:

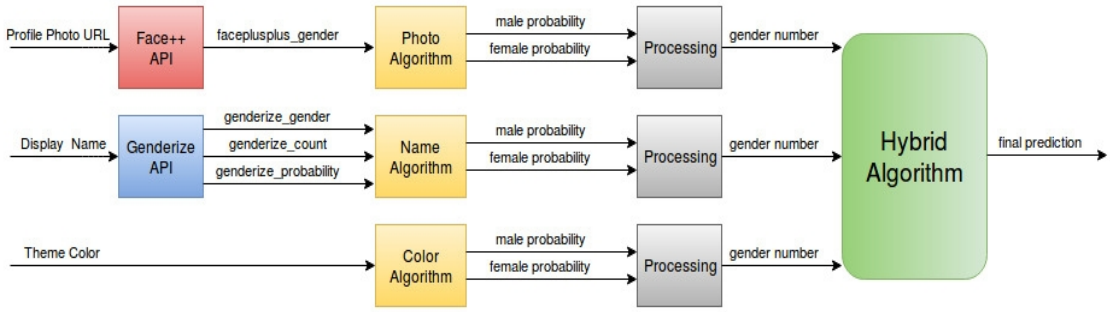
- Αν $P_m \geq P_f$ τότε ο *αριθμός φύλου* είναι ίσος με $1-P_m$.
- Αν $P_m < P_f$ τότε ο *αριθμός φύλου* είναι ίσος με P_f .

Για παράδειγμα για κάποιον χρήστη που ένας αλγόριθμος του δίνει 0.3 πιθανότητα να είναι άντρας και 0.7 να είναι γυναίκα, παίρνουμε σαν έξοδο από τον αλγόριθμο τον αριθμό 0.3, ενώ για έναν χρήστη που του δίνεται πιθανότητα 0.7 να είναι άντρας και 0.3 να είναι γυναίκα, παίρνουμε σαν έξοδο τον αριθμό 0.3.

Η κωδικοποίηση αυτή των αριθμών φύλου, έγινε με τη λογική ότι μια τελείως αβέβαιη πρόβλεψη φύλου θα πάρει τιμή κοντά στο 0.5. Μια πρόβλεψη με μεγάλη αυτοπεποίθηση ότι ο χρήστης είναι γυναίκα θα πάρει τιμή κοντά στο 0 και μια πρόβλεψη με μεγάλη αυτοπεποίθηση ότι ο χρήστης είναι άντρας θα πάρει τιμή κοντά στο 1. Με αυτό τον τρόπο περιέχεται σε έναν μόνο αριθμό με εύρος τιμών [0,1] όλη η απαραίτητη πληροφορία που χρειάζεται ο υβριδικός αλγόριθμος από κάθε έναν από τους αλγορίθμους των ξεχωριστών προσεγγίσεων για να εκτιμήσει τη βεβαιότητα τους για την πρόβλεψη σε κάθε χρήστη.

Ο υβριδικός αλγόριθμος ο οποίος είναι υλοποιημένος κι αυτός με ένα μοντέλο επιβλεπόμενης μάθησης, εκπαιδεύεται με διάνυσματα χαρακτηριστικών που περιέχουν τους τρεις αριθμούς φύλου που προέκυψαν από τις εξόδους των Color Algorithm, Photo Algorithm και Name Algorithm. Στο Σχήμα 5.1 συνοψίζεται η αρχιτεκτονική της υβριδικής προσέγγισης την οποία περιγράψαμε. Οι τρεις

αλγόριθμοι των διακριτών μεθόδων χρωματίζονται με ίδιο χρώμα για να τονιστεί ότι στηρίζονται σε ίδιο μοντέλο μηχανικής μάθησης με σκοπό την συνάφεια των εξόδων τους.



Σχήμα 5.1: Αρχιτεκτονική Υβριδικής Προσέγγισης

5.3 Πειράματα και Αποτελέσματα

5.3.1 Προετοιμασία

Για τα πειράματα μας δοκιμάσαμε ως Color Algorithm, Photo Algorithm και Name Algorithm, SVM με RBF πυρήνα και PNN τα οποία είχαν την καλύτερη απόδοση κατά τα πειράματα με βάση το χρώμα (βλ. Πίνακα 4.2).

Για την απόκτηση των πιθανοτήτων φύλου που δίνουν οι Color Algorithm, Photo Algorithm και Name Algorithm για κάθε χρήστη της βάσης δεδομένων μας πραγματοποιήσαμε ξεχωριστά για κάθε έναν 5-fold cross-validation. Πιο συγκεκριμένα, και για τις τρεις περιπτώσεις σε κάθε fold εκπαιδεύσαμε τον ανάλογο αλγόριθμο με βάση τα διανύσματα χαρακτηριστικών του training set που σχετίζονται με την κάθε προσέγγιση και αποθηκεύσαμε τις πιθανότητες που εξήχθησαν για το test set. Στο τέλος της όλης διαδικασίας είχαμε στη διάθεση μας τις πιθανότητες που εξήχθησαν από τους τρεις αλγορίθμους για όλο το σύνολο των χρηστών της βάσης μας. Με βάση αυτές τις πιθανότητες δόθηκε ένας “αριθμός φύλου” σε κάθε χρήστη για κάθε διακριτή προσέγγιση ανίχνευσης φύλου

όπως αναφέρθηκε στην Ενότητα 5.2

Στον Πίνακα 5.1 παραθέτουμε για λόγους πληρότητας, τα accuracies που πετύχαμε με την διαδικασία του stratified 5-fold cross-validation για κάθε προσέγγιση με χρήση SVM με RBF πυρήνα και PNN. Παρατηρούμε ότι το “φιλτράρισμα” των εξόδων του Face++ και του Genderize με αλγορίθμους μηχανικής μάθησης αύξησε κατά πολύ το accuracy των δυο αυτών προσεγγίσεων. Στις Ενότητες 4.2.4 και 4.3.3, είχαμε αναφέρει τις σχετικές αποδόσεις ανίχνευσης φύλου με βάση την εικόνα προφίλ και το display name μετρώντας τον αριθμό των σωστών εκτιμήσεων φύλου με βάση το *faceplusplus_gende*r και το *genderize_gender* ως προς το σύνολο των χρηστών της βάσης μας. Με το *faceplusplus_gende*r είχαμε 65.75% ακρίβεια για την ανίχνευση φύλου όλων των χρηστών, ενώ με το *genderize_gender* 60.31%. Με αυτή τη “naive” μέθοδο δεν είχαμε εκμεταλλευτεί κάποιες επιπλέον χρήσιμες πληροφορίες για την ανίχνευση του φύλου. Ειδικότερα, για την προσέγγιση με το Genderize δεν λάβαμε υπόψην τα επιπλέον στοιχεία που δίνουν τα πεδία *genderize_count* και *genderize_probability*. Επίσης δεν εκμεταλλευτήκαμε το γεγονός ότι η τιμή “Άγνωστο” μπορεί να εμφανίζεται συχνότερα για κάποιο φύλο βάσει της κάθε προσέγγισης. Οι ταξινομητές επιβλεπόμενης μάθησης, μέσω της διαδικασίας εκπαίδευσης μπορούν να χρησιμοποιήσουν αυτές τις πρόσθετες πληροφορίες, μαθαίνοντας τα μοτίβα που διέπουν τις εκτιμήσεις των Genderize και Face++ σε συνδυασμό με τα πραγματικά φύλα των χρηστών της βάσης μας.

	SVM-RBF	PNN
Color Algorithm	65.67%	65.16%
Name Algorithm	78.12%	77.43%
Photo Algorithm	79.73%	77.10%

Πίνακας 5.1: Accuracies για τους Color Algorithm, Photo Algorithm και Name Algorithm με SVM-RBF και PNN

Για την βελτιστοποίηση της απόδοσης, δοκιμάσαμε για κάθε έναν από τους Color Algorithm, Photo Algorithm και Name Algorithm ξεχωριστά τις κατάλληλες παραμέτρους των μοντέλων SVM και PNN που χρησιμοποιήσαμε. Επίσης, για τα διανύσματα χαρακτηριστικών που αφορούν τα σχετικά πεδία του Genderize, πραγματοποιήσαμε κανονικοποίηση των χαρακτηριστικών με χρήση της μεθόδου *scale* του scikit-learn πριν την τροφοδότηση τους στον Name Algorithm. Ο λόγος

που πραγματοποιήσαμε αυτή τη διαδικασία είναι ότι το πεδίο *genderize_probability* έχει εύρος ακέραιων τιμών στο διάστημα [0,12593] ενώ τα άλλα 2 πεδία έχουν τιμές μεταξύ του 0 και του 1. Συνεπώς, μπορεί να κυριαρχήσει στις προβλέψεις του Name Algorithm μη αφήνοντας τον να μάθει από τα άλλα χαρατηριστικά όπως αναμένεται. Η μέθοδος *scale* κανονικοποιεί ξεχωριστά το κάθε χαρακτηριστικό έτσι ώστε να έχει μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση.

5.3.2 Απόδοση Υβριδικού Αλγορίθμου

Στο επόμενο στάδιο της διαδικασίας, πραγματοποιήσαμε stratified 5-fold cross-validation με τον υβριδικό αλγόριθμο παίρνοντας από κάθε training set ως διάνυσματα χαρακτηριστικών τους 3 *αριθμούς φύλου* που αντιστοιχούσαν στους χρήστες του συνόλου και με βάση την εκπαίδευση πάνω σε αυτά κάναμε προβλέψεις για το φύλο των χρηστών του κάθε test set. Στο τέλος της διαδικασίας μετρήσαμε τον μέσο όρο των accuracies που είχαμε σε κάθε γύρο και διαλέξαμε το καλύτερο μοντέλο επιβλεπόμενης μάθησης για τον υβριδικό αλγόριθμο. Οι δοκιμές έγιναν με Gaussian Naive Bayes, SVM και PNN με εύρεση των καλύτερων παραμέτρων όπου χρειάστηκε.

Στον Πίνακα 5.2 φαίνονται τα accuracies που πετύχαμε με τον υβριδικό αλγόριθμο. Στις γραμμές του πίνακα είναι τα μοντέλα επιβλεπόμενης μάθησης που δοκιμάσαμε για την υλοποίηση του υβριδικού αλγορίθμου και στις 2 στήλες του πίνακα διαχωρίζουμε τις περιπτώσεις όπου χρησιμοποιήσαμε SVM με RBF πυρήνα ή PNN για την υλοποίηση των Color Algorithm, Photo Algorithm και Name Algorithm.

Hybrid \ *Photo*, *Name*, Color	SVM-RBF	PNN
Gaussian Naive Bayes	83.21%	86.23%
SVM-RBF	84.40%	**87.2%**
SVM-Polynomial	84.24%	86.23%
SVM-Linear	83.82%	87.06%
PNN	84.38%	86.28%

Πίνακας 5.2: Accuracies υβριδικού αλγορίθμου

Παρατηρούμε ότι η απόδοση του υβριδικού ταξινομητή είναι πολύ καλύτερη όταν λαμβάνει “αριθμούς φύλων” που εξάγονται από τις πιθανότητες που δίνουν Πιθανωτικά Νευρωνικά Δίκτυα αντί για Μηχανές Διανυσμάτων Υποστήριξης. Αυτό έρχεται σε αντίθεση με τα αποτελέσματα του Πίνακα 5.1 που είδαμε ότι τα SVMs έχουν καλύτερη απόδοση σε σχέση με τα PNNs σε κάθε προσέγγιση. Αυτό οφείλεται σε 2 λόγους. Πρώτον, τα PNNs έχουν την ικανότητα να παράγουν πολύ αξιόπιστες πιθανότητες για την κλάση που ανήκει κάποιο δείγμα δεδομένων όπως αναφέραμε και στην Ενότητα 2.2.2.3. Δεύτερον, ανατρέχοντας στο documentation του scikit-learn[27], βλέπουμε ότι για την εξαγωγή πιθανοτήτων από την κλάση *SVC* με τη μέθοδο *predict_proba* χρησιμοποιείται ένα επιπλέον 5-fold cross-validation στο training set και οι πιθανότητες αυτές μπορεί να μην συμβαδίζουν με τις προβλέψεις που κάνει το SVM μέσω της μεθόδου *predict* (μπορεί η *predict_proba* να δίνει μικρότερη πιθανότητα στο φύλο που προέκυψε από την *predict*). Βάσει των παραπάνω συνάγεται ότι τα PNN είναι καταλληλότερα για την εξαγωγή των *αριθμών φύλου*.

Η καλύτερη απόδοση επιτεύχθει με την υλοποίηση των Color Algorithm, Photo Algorithm και Name Algorithm με Πιθανολογικά Νευρωνικά Δίκτυα και την υλοποίηση του Hybrid Algorithm με Μηχανή Διανυσμάτων Υποστήριξης με πυρήνα RBF. Το accuracy αυτού του συνδυασμού ήταν 87.2%.

5.3.3 Επιρροή κάθε ξεχωριστού πεδίου στην συνολική απόδοση

Για να εξετάσουμε την επιρροή που είχαν το χρώμα, το όνομα και η φωτογραφία στην υβριδική προσέγγιση, καταγράψαμε την απόδοση του υβριδικού ταξινομητή για τις περιπτώσεις αφαίρεσης του κάθε πεδίου από τη διαδικασία (κατά την αφαίρεση του χρώματος και του ονόματος δεν χρησιμοποιήθηκαν ούτε οι Photo και Name Algorithms ούτε τα Genderize και Face++). Στον Πίνακα 5.3 φαίνονται τα accuracies που καταγράψαμε για κάθε συνδυασμό πεδίων που χρησιμοποιήσαμε.

[27] http://scikit-learn.org/stable/modules/svm.html

Πεδία	Accuracy
χρώμα+εικόνα	82.52%
χρώμα+όνομα	79.31%
εικόνα+όνομα	86.45%
χρώμα+εικόνα+όνομα	87.2%

Πίνακας 5.3: Accuracies για διαφορετικούς συνδυασμούς των συνιστωσών του υβριδικού αλγορίθμου

Παρατηρούμε ότι η χρησιμοποίηση και των τριών πεδίων, επιφέρει την καλύτερη απόδοση για την υβριδική προσέγγιση όπως αναμέναμε. Παρατηρούμε όμως, ότι η αφαίρεση του χρώματος από την υβριδική προσέγγιση οδηγεί σε ελάττωση μόλις 0.75% του accuracy. Συμπεραίνουμε λοιπόν ότι το χρώμα δεν προσφέρει πολύ μεγάλη βοήθεια για την ανίχνευση του φύλου των χρηστών του Twitter, πράγμα που αναμέναμε και με βάση τα αποτελέσματα των Πίνακων 5.1 και 4.2. Έστω και αυτή η μικρή αύξηση της απόδοσης όμως που προσφέρει η συμπερίληψη του χρώματος, είναι σημαντική λόγω του εν γένει δύσκολου προβλήματος της ανίχνευσης του φύλου.

5.3.4 Ποσοστά εκτιμήσεων με χαμηλή πιθανότητα σφάλματος

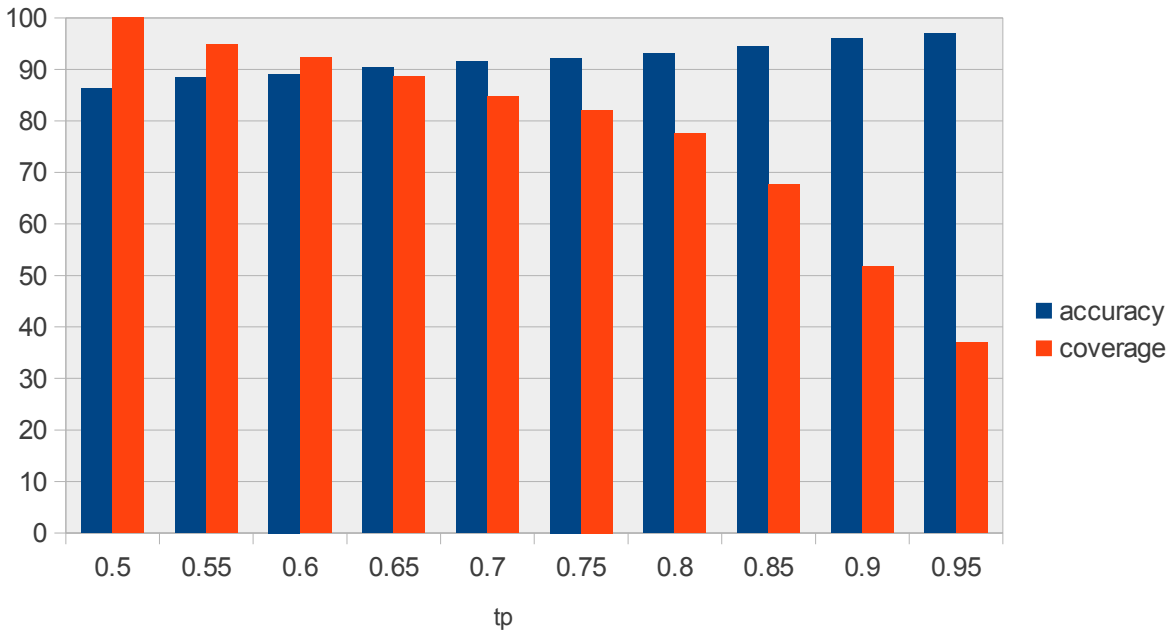
Όπως έχουμε αναφέρει, η ανίχνευση του φύλου από χρήστες του Twitter μπορεί να γίνει για ποικίλους σκοπούς. Ο λόγος για τον οποίο επιχειρείται η ανίχνευση, διαφοροποιεί την ανοχή σχετικά με την πιθανότητα σφάλματος. Παραδείγματος χάριν, μια εταιρεία που έχει προφίλ στο Twitter και θέλει να έχει μια σφαιρική εικόνα για τα ποσοστά των δυο φύλων που απαρτίζουν τους followers της, προτιμά να λάβει εκτιμήσεις φύλου για όλους τους χρήστες που εξετάζει παρά να περιορίσει τις εκτιμήσεις σε κάποιο πιο “σίγουρο” υποσύνολο. Αντίθετα, μια εταιρεία που θέλει να στείλει απευθείας μηνύματα (direct messages) στους followers της για να διαφημίσει προιόντα που απευθύνονται αποκλειστικά σε άντρες ή γυναίκες, θα προτιμήσει να λάβει εκτιμήσεις φύλου για τις πιο σίγουρες περιπτώσεις χρηστών.

Με κίνητρο τα παραπάνω, ακολουθήσαμε την εξής διαδικασία:

- Χρησιμοποιήσαμε την υβριδική προσέγγιση χρησιμοποιώντας PNNs για τους

Color, Name και Photo Algorithms αλλά και για τον Υβριδικό Αλγόριθμο. Ο λόγος που δεν χρησιμοποίησαμε το SVM για τον Υβριδικό Αλγόριθμο είναι όπως αναφέραμε προηγουμένως ότι δεν παράγει τόσο αξιόπιστες πιθανότητες ορθών πρόβλεψεων όσο το PNN, παρόλο που έχει καλύτερο accuracy. Από τον Υβριδικό Αλγόριθμο, με χρήση της μεθόδου *predict_proba*, πήραμε για κάθε χρήστη τις πιθανότητες ταύτισης του με κάθε φύλο και κρατήσαμε την μεγαλύτερη η οποία αποτελεί την "αυτοπεποίθηση" που έχει ο αλγόριθμος ότι η πρόβλεψη φύλου που κάνει για αυτόν τον χρήστη είναι σωστή.

- Θεωρήσαμε ένα threshold αυτοπεποίθησης t_p το οποίο παίρνει τιμές από 0.5 έως 0.95 (δεν υπήρχαν προβλέψεις με πιθανότητα 1). Ορίζοντας ως coverage (ποσοστό κάλυψης) το ποσοστό των χρηστών των οποίων η αυτοπεποίθηση ορθής πρόβλεψης είναι μεγαλύτερη ή ίση του t_p και μετρώντας το accuracy των προβλέψεων για κάθε τέτοιο υποσύνολο χρηστών, πήραμε για τις διαφορετικές τιμές του t_p τα αποτελέσματα που φαίνονται στο Σχήμα 5.2



Σχήμα 5.2: Συνδυασμοί accuracy και coverage για διαφορετικές πιθανότητες ορθής πρόβλεψης του υβριδικού αλγορίθμου

Ένας πολύ ενδιαφέρον συνδυασμός που πετύχαμε είναι accuracy 96.08% για ένα υποσύνολο χρηστών που είναι παραπάνω από το μισό του αρχικού συνόλου (51.73%). Συμπεραίνουμε ότι περιορίζοντας την ανίχνευση φύλου στους μισούς χρήστες του Twitter, μπορούμε χρησιμοποιώντας την προσέγγιση μας η οποία κάνει χρήση ελάχιστων χαρακτηριστικών για τον κάθε χρήστη, να πετύχουμε ένα πάρα πολύ υψηλό accuracy. Αυτό είναι πολύ ενθαρρυντικό για οντότητες που θέλουν να έχουν πολύ υψηλής ακρίβειας προβλέψεις για ένα ικανοποιητικό ποσοστό χρηστών του Twitter χωρίς τεράστιο χρονικό και υπολογιστικό κόστος.

5.4 Συμπεράσματα και Παρατηρήσεις

Με βάση τα αποτελέσματα που εξάγαμε με τη χρήση του υβριδικού αλγορίθμου μηχανικής μάθησης, παρατηρούμε ότι το accuracy 87.2% που επιτεύχθει για τις προβλέψεις φύλου του συνόλου των χρηστών Twitter της βάσης δεδομένων μας, είναι αρκετά ανταγωνιστικό συγκριτικά με παρόμοιες μελέτες που χρησιμοποιούν τεράστιο αριθμό από χαρακτηριστικά και χώρους υψηλών διαστάσεων, κάνοντας χρήση *n*-grams και άλλων χαρακτηριστικών εξαγόμενων από tweets. Επιπλέον το 96.08% accuracy που πετύχαμε για παραπάνω από τους μισούς χρήστες, μας δείχνει ότι μπορεί να πραγματοποιηθεί ιδιαίτερα ακριβής ανίχνευση του φύλου για ένα σημαντικό ποσοστό των χρηστών Twitter με πολύ αποδοτικό τρόπο.

Με βάση τα παραπάνω ωθούμαστε στο συμπέρασμα ότι η χρησιμοποίηση μεθόδων ανίχνευσης φύλου που βασίζονται σε απλά στοιχεία από τα προφίλ των χρηστών του Twitter μπορεί να προσφέρει αξιοσημείωτα ανταγωνιστική απόδοση συγκριτικά με πολύ πιο περίπλοκες προσεγγίσεις. Υπάρχει λοιπόν αρκετό ερευνητικό ενδιαφέρον στην τροποποίηση και βελτιστοποίηση παρόμοιων μεθόδων.

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη και Συμπεράσματα

Στα πλαίσια αυτής της διπλωματικής εργασίας προτάθηκε ένας υβριδικός αλγόριθμος μηχανικής μάθησης για την ανίχνευση του φύλου των χρηστών του Twitter. Ο αλγόριθμος αυτός εκμεταλλεύεται τα στοιχεία που προκύπτουν από την φωτογραφία προφίλ, το όνομα και το χρώμα θέματος ενός προφίλ στο Twitter, για να εκτιμήσει το φύλο του αντίστοιχου χρήστη. Τα στοιχεία αυτά είναι πολύ εύκολο και γρήγορο να συλλεχθούν και να υποστούν κατάλληλη επεξεργασία για την χρησιμοποίηση τους. Επιπλέον, λόγω της απλότητας τους κάνουν δυνατή την εκπαίδευση των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται, σε χώρους χαμηλών διαστάσεων. Συνεπώς, η διαδικασία της ανίχνευσης φύλου πραγματοποιείται με φιλικό τρόπο ως προς την απαιτούμενη χρονική διάρκεια και κατανάλωση πόρων. Αυτό καθιστά την προτεινόμενη προσέγγιση επεκτάσιμη σε πολύ μεγάλους πληθυσμούς του Twitter.

Τα βασικότερα συμπεράσματα που προέκυψαν από την εργασία αυτή είναι τα εξής:

1. Υπάρχει η δυνατότητα επίτευξης μεγάλης ακρίβειας στις προβλέψεις φύλου σε ένα δείγμα χρηστών αντιπροσωπευτικό του συνολικού πληθυσμού του Twitter, χωρίς να εφαρμοστούν μέθοδοι που εξαρτώνται από την γλώσσα και χρησιμοποιούν στοιχεία από ανάλυση κειμένων, καταλήγοντας με πολυάριθμα εξεταζόμενα χαρακτηριστικά. Προτείνεται στους ερευνητές του πεδίου, η κατεύθυνση της πιο επιμελούς εκμετάλλευσης των χρήσιμων στοιχείων που περιέχονται στα προφίλ του Twitter, αντί της ανάλυσης των tweets.

2. Οι αλλαγές που πραγματοποίησε το Twitter σχετικά με την επιλογή των χρωμάτων διαμόρφωσης του προφίλ, επηρέασε αρνητικά την δυνατότητα ανίχνευσης του φύλου των χρηστών. Προτείνεται ιδιαίτερη μέριμνα από

τους ερευνητές που στηρίζονται σε ευρήματα παλιότερων μελετών για τον σχεδιασμό της προσέγγισης τους.

6.2 Μελλοντικές Επεκτάσεις

Η παρούσα μελέτη δίνει χώρο για πολλές πιθανές επεκτάσεις. Κάποιες από αυτές τις επεκτάσεις χρησιμεύουν στην περαιτέρω αξιολόγηση της παρούσας προσέγγισης ανίχνευσης φύλου και κάποιες στοχεύουν στη βελτίωση της.

Ένας προφανής τρόπος αξιολόγησης της παρούσας προσέγγισης είναι η χρησιμοποίηση μιας πολύ μεγαλύτερης βάσης δεδομένων από χρήστες του Twitter, για να γίνει βέβαιο ότι η απόδοση που καταγράφηκε στην παρούσα μελέτη γενικεύει σε μεγαλύτερους πληθυσμούς χρηστών, με πιθανή βελτίωση λόγω του μεγαλύτερου μεγέθους πληροφορίας που θα έχουμε στη διάθεση μας.

Επιπλέον, εκτός από τη χρήση συγκεκριμένης βάσης δεδομένων, η προσέγγιση μας θα μπορούσε να εφαρμοστεί σε δεδομένα πραγματικού χρόνου. Για παράδειγμα, μπορούμε να συλλέξουμε τα tweets που αναφέρονται μέσω hashtags σε κάποιο τρέχον κοινωνικό ή ψυχαγωγικό γεγονός και να ταξινομήσουμε τους αντίστοιχους χρήστες κατά φύλο μέσω της μεθόδου που προτείναμε. Με αυτό τον τρόπο, θα εξακριβώσουμε αν τα ποσοστά των φύλων που λάβαμε είναι τα αναμενόμενα (π.χ. συλλέγοντας tweets με το hashtag #NBA, περιμένουμε η συντριπτική πλειοψηφία των χρηστών να είναι άντρες).

Μια καλή πρόταση για την βελτίωση της προσέγγισης μας είναι η χρήση περισσότερων βάσεων δεδομένων με ονόματα, ώστε να διαλέγουμε για τον προσδιορισμό του φύλου κάθε ονόματος αυτή που φαίνεται πιο αξιόπιστη (π.χ. αυτή που περιέχει περισσότερες εγγραφές του ονόματος).

Όσον αφορά τη χρησιμοποίηση επιπλέον πληροφορίας για κάθε χρήστη με σκοπό την αύξηση της ακρίβειας προβλέψεων φύλου, η μόνη πηγή που είναι πιθανό να περιέχει στοιχεία για το φύλο του χρήστη και επιπλέον δεν επηρεάζει την κλιμακωσιμότητα της προσέγγισης μας είναι το *bio*. Ένας αποδοτικός τρόπος χρησιμοποίησης του είναι μέσω αναζήτησης για λέξεις-κλειδιά που μπορεί να προδώσουν το φύλο του χρήστη (π.χ. “father” για τους άντρες και “mother” για τις γυναίκες). Εντούτοις, η χρησιμοποίηση του *bio* έχει το πολύ βασικό μειονέκτημα ότι εξαρτάται από την γλώσσα, οπότε θα πρέπει να περιοριστούμε στην ανίχνευση

φύλου ορισμένης μερίδας του συνολικού πληθυσμού του Twitter. Επίσης χρειάζεται ιδιαίτερη μέριμνα για την αποφυγή περιπτώσεων όπου μια χαρακτηριστική λέξη μπορεί να χρησιμοποιείται διαφορετικά από τα 2 φύλα (π.χ. “proud husband” από τους άντρες, “my husband” από τις γυναίκες).

Σχετικά με την επέκταση της προσέγγισης μας χωρίς να εκμεταλλευτούμε παραπάνω πληροφορίες για τους χρήστες, μια ενδιαφέρουσα κατεύθυνση είναι η δοκιμή διαφορετικών μοντέλων επιβλεπόμενης μάθησης για την ανίχνευση του φύλου, όπως είναι τα Δέντρα Αποφάσεων. Επίσης μπορούμε να πειραματιστούμε με μοντέλα μη επιβλεπόμενης μάθησης όπως η ομαδοποίηση Κ-μέσων (k-means clustering). Τα μοντέλα μη επιβλεπόμενης μάθησης έχουν το μεγάλο πλεονέκτημα ότι μπορούν να εκπαιδευτούν χωρίς να χρειάζονται σύνολο δεδομένων από χρήστες επισημασμένους με φύλο. Θα μπορούσαμε λοιπόν να χρησιμοποιήσουμε έναν αλγόριθμο μη επιβλεπόμενης μάθησης σε ένα μεγάλο σύνολο τυχαίων χρηστών του Twitter που δεν γνωρίζουμε το φύλο τους για την κατασκευή clusters, και στη συνέχεια να αξιολογήσουμε την απόδοση του με το σύνολο χρηστών της βάσης δεδομένων μας που έχουν γνωστό φύλο.

Βιβλιογραφία

[1] Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, SMUC ’10 (pp. 37–44). New York, NY, USA: ACM.

[2] J. S. Alowibdi, U. A. Buy, and P. Yu. Language independent gender classification on twitter. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 739–743. ACM, 2013.

[3] Liu, W. & Ruths, D. (2013). What’s in a name? using first names as features for gender
inference in twitter. In AAAI Spring Symposium: Analyzing Microtext.

[4] Vicente, M., Batista, F., Carvalho, J.P.: Twitter gender classification using user unstructured information. In: Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istambul, Turkey, August 2015.

[5] Alowibdi JS, Buy UA, Yu PS, Stenneth L (2014) Detecting deception in online social networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research , 12:2825–2830, 2011

[7] Jisun An and Ingmar Weber. 2016. # greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter. In AAAI ICWSM.

[8] Nguyen DP, Trieschnigg R, Doğruöz A, Gravel R, Theune M, Meder T, et al. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In: Proceedings of the 25th International Conference on Computational Linguistics. COLING; 2014. p. 1950–1961.

[9] Cesare, Nina & Grant, Christan & O. Nsoesie, Elaine. (2017). Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices.

[10] Eckert, P. & McConnell-Ginet, S. (2013). Language and gender. Cambridge University Press.

[11] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’11 (pp. 1301–1309). Stroudsburg, PA, USA: Association for Computational Linguistics.

[12] Al Zamal, F., Liu, W., & Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. ICWSM, 270.

[13] T. development team,“Api overview.” https://dev.twitter.com/overview/api, 2015.

[14] A. Culotta, “Detecting influenza outbreaks by analyzing twitter messages,” arXiv preprint arXiv:1007.4748, 2010.

[15] P. Earle, M. Guy, R. Buckmaster, C. Ostrum, S. Horvath, and A. Vaughan, “Omg earthquake! can twitter improve earthquake response?,” Seismological Research Letters, vol. 81, no. 2, pp. 246–251, 2010.

[16] J. P. Carvalho, V. Pedro, and F. Batista, “Towards intelligent mining of public social networks’ influence in society,” in IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), (Edmonton, Canada), pp. 478 – 483, June 2013

[17] Kohavi, Ron. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. 14.

[18] Tang C., Ross K., Saxena N., Chen R. (2011) What’s in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook. In: Xu J., Yu G., Zhou S., Unland R. (eds) Database Systems for Adanced Applications. DASFAA 2011. Lecture Notes in Computer Science, vol 6637. Springer, Berlin, Heidelberg

[19] Quanzeng, Y., Bhatia, S., Tong, S., Jiebo, L., “The Eyes of the Beholder: Gender Prediction Using Images Posted in Online Social Networks,”in IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1026-1030, 14 Dec 2014.

[20] Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412

[21] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119–123.

[22] Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers age and gender. In Third International AAAI Conference on Weblogs and Social Media.

[23] Cheng, N., Chandramouli, R., & Subbalakshmi, K. (2011). Author gender identification from text. Digital Investigation, 8(1), 78–88.

[24] Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 37–44).: ACM.

[25] Aravantinou, C., Simaki, V., Mporas, I., & Megalooikonomou, V. (2015). Gender classification of web authors using feature selection and language models. In A. Ronzhin, R. Potapova, & N. Fakotakis (Eds.), Speech and Computer, volume 9319 of Lecture Notes in Computer Science (pp. 226–233). Springer International Publishing.

[26] Baumann, A., Krasnova, H., Veltri, N. F., & Ye, Y. (2015). Men, women, microblogging: Where do we stand?

[27] N. Garera and D. Yarowsky. Modeling latent biographic attributes in conversational genres. In Proceedings of the Joint Conference of Association of Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 710–718, 2009.

[28] Bill Heil and Mikolaj Jan Piskorski. 2009. New Twitter research: Men follow men and nobody tweets. Harvard Business Review, June 1.

[29] Robin Wauters. 2010. Only 50% of Twitter messages are in English, study says. TechCrunch, February 1. http://techcrunch.com/2010/02/24/twitter-languages/.

[30] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, (pp. 415–444).

[31] Bamman, D., Eisenstein, J., & Schnoebelen, T. (2012). Gender in twitter: Styles, stances, and social networks. CoRR abs/1210.4567.

[32] Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W. (2012). Gender identification on twitter using the modified balanced winnow.

[33] Miller, Z., Dickinson, B., & Hu, W. (2012). Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. International Journal of Intelligence Science,2(24).

[34] Mislove, A.; Lehmann, S.; Ahn, Y.; Onnela, J.; and Rosenquist, J. 2011. Understanding the Demographics of Twitter Users. In Proceedings of the International
Conference on Weblogs and Social Media.

[35] https://en.wikipedia.org/wiki/Curse_of_dimensionality

[36] H. Bechar-Israeli, “From< bonehead> to< clonehead>: Nicknames, play, and identity on internet relay chat1,” Journal of Computer-Mediated Communication, vol. 1, no. 2, pp. 0–0, 1995.

[37] S. L. Calvert, B. A. Mahler, S. M. Zehnder, A. Jenkins, and M. S. Lee, “Gender differences in preadolescent children’s online interactions: Symbolic modes of self-presentation and self-expression,” Journal of Applied Developmental Psychology, vol. 24, no. 6, pp. 627–644, 2003.

[38] Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2008) ‘A Practical Guide to Support Vector Classification’, BJU international, 101(1), pp. 1396–400.

[39] http://neupy.com/pages/home.html

[40] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN: 978-0137903955.

[41] Haykin S (2009) Neural Networks & Learning Machines (3rd ed.). Prentice-Hall. ISBN: 978-0-13-147139-9

[42] D. F. Specht, "Probabilistic neural network for classification, mapping, or associative memory," in Proc. IEEE Int. Conf. Neural Network, vol.1, San Diego, CA, pp. 525-532, July

1988

[43] http://scikit-learn.org/stable/

[44] Holmes, J. & Meyerhoff, M. (2008). The handbook of language and gender, volume 25. John Wiley & Sons.

[45] Bucholtz, M. & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. Discourse studies, 7(4-5), 585–614.

[46] Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. Word, 14(1), 47–56.

[47] Labov, W. (2006). The social stratification of English in New York city. Cambridge University Press.