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

Μηχανική μάθηση για την εκτίμηση τηλεθέασης βάσει δεδομένων κοινωνικών δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΡΙΟΣ Μ. ΠΑΡΑΣΚΕΥΟΠΟΥΛΟΣ

Επιβλέπων : Δρ. Ιωάννα Γ. Ρουσσάκη
Επ. Καθηγήτρια Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2018

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

Μηχανική μάθηση για την εκτίμηση τηλεθέασης βάσει δεδομένων κοινωνικών δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΡΙΟΣ Μ. ΠΑΡΑΣΚΕΥΟΠΟΥΛΟΣ

Επιβλέπων : Δρ. Ιωάννα Γ. Ρουσσάκη
Επ. Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Δεκεμβρίου 2018.

.....	.....	.....
Ιωάννα Ρουσσάκη	Μιλτιάδης Αναγνώστου	Συμεών Παπαβασιλείου
Επ. Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2018

.....

ΜΑΡΙΟΣ Μ. ΠΑΡΑΣΚΕΥΟΠΟΥΛΟΣ

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

Περίληψη

Τα κοινωνικά δίκτυα και ιδιαίτερα τα μέσα κοινωνικής δικτύωσης παρουσιάζουν μια διαρκώς αυξανόμενη δημοτικότητα τα τελευταία χρόνια, απόρροια της ευκολίας που παρέχουν στο χρήστη ως προς την επικοινωνία, τη διάδοση πληροφορίας και γνώσης και την ενημέρωση.

Τέτοια δίκτυα προσφέρουν αφθονία περιεχομένου δημιουργούμενου από τους χρήστες και αποτελούν μια ανοιχτή και πλούσια πηγή δεδομένων για τη μελέτη διαφόρων πτυχών της ανθρώπινης συμπεριφοράς από κοινωνικές και δικτυακές επιστήμες. Αυτός ο τεράστιος όγκος δεδομένων δύναται να αξιοποιηθεί και σε εφαρμογές, όπως το στοχευμένο μάρκετινγκ, η δημοσκόπηση και οι επιχειρηματικές αναλύσεις. Σημαντικό εργαλείο στα χέρια των αναλυτών αποτελεί η εξαγωγή πληροφοριών για τις προτιμήσεις και τα ενδιαφέροντα των χρηστών των κοινωνικών δικτύων.

Το αντικείμενο της παρούσης διπλωματικής εργασίας είναι η μελέτη και εξαγωγή συμπερασμάτων για το ενδιαφέρον των χρηστών για διάφορες τηλεοπτικές εκπομπές βασίζοντας την ανάλυση στη δραστηριότητα τους στα κοινωνικά δίκτυα. Συγκεκριμένα, αντλούνται στοιχεία από τις πλατφόρμες Twitter και Google Trends. Η πρώτη αποτελεί ένα από τα δημοφιλέστερα μέσα κοινωνικής δικτύωσης με μερικές εκατοντάδες εκατομμύρια χρήστες ανά την υφήλιο. Χαρακτηρίζεται από την ελευθερία πρόσβασης στα δεδομένα που παράγουν οι χρήστες, ωστόσο δε διαθέτει υποχρεωτικά πεδία συμπλήρωσης δημογραφικών στοιχείων και στοιχείων προτίμησης. Το γεγονός αυτό, σε συνδυασμό με το αυξανόμενο ερευνητικό και εμπορικό ενδιαφέρον για τέτοια δεδομένα, έχει οδηγήσει στη διεξαγωγή πληθώρας ερευνών και τη δημιουργία διαφόρων τρόπων έμμεσης ανάκτησης τέτοιων στοιχείων από τους λογαριασμούς των χρηστών.

Η διαφοροποίηση της προσπάθειας αυτής έγκειται στην καινοτόμο αξιοποίηση της δεύτερης πλατφόρμας εκ των αναφερθέντων. Τα Google Trends, προϊόν της περιώνυμης εταιρείας, είναι μια ιστοσελίδα σύγκρισης της δημοφιλίας των όρων αναζήτησης στην ομώνυμη μηχανή αναζήτησης με έμφαση στη γεωγραφική και γλωσσική ευελιξία της ανάλυσης.

Η εργασία αυτή καταδεικνύει τη δύναμη της πληροφορίας που εξάγεται από τη δραστηριότητα των χρηστών του Twitter και τη δυνατότητα αξιοποίησης δεδομένων από τα Google Trends στην εκτίμηση της ακροαματικότητας τηλεοπτικών εκπομπών.

Λέξεις κλειδιά

Επιστήμη δεδομένων, Μηχανική μάθηση, Κοινωνικά δίκτυα, Εξαγωγή γνώσης, Twitter, Google Trends, Παλινδρόμηση, Ενδιαφέροντα Χρηστών, Τηλεόραση, Τηλεθέαση, Τηλεοπτικές εκπομπές

Abstract

Social networks and media have increased their popularity over the last few years, as they facilitate the users' right in spreading and acquiring information and knowledge and chiefly in communication.

These networks hold a variety of content created and uploaded by the users and constitute a huge, free data source that can be exploited in both scientific and humanitarian research. Polls, targeted marketing and business analytics are only a few applications that these data are a valuable input in. Extracting information from the users' activity in social networks surely adds some sharp arrows inside the analyst's quiver.

Subject of the thesis is to study the activity in social networks and reach a conclusion about the interest of the users for live TV shows. Specifically, we create our dataset from Twitter and Google Trends platforms. The first stands among the most popular social media, counting hundreds of millions of users around the globe. Everyone may have free access to this treasure, though there are no obligatory fields for the users to fill in order to sign up. This fact, along with the increasing interest of similar data for scientific and commercial purposes, has led to a vast amount of studies that focus on indirect ways to detect the interests of Twitter users.

Including data from the latter is what makes this approach differ from the aforementioned. Google Trends is a website by Google that analyzes the popularity of top search queries in Google Search across various regions and languages. The website uses graphs to compare the search volume of different queries over time.

We conclude that Twitter is a powerful data source, a miniature on society at times and users' activity can be connected to their preferences and interests. Furthermore, Google trends can provide us with useful data in audience analytics and estimation.

Keywords

Data Science, Machine Learning, Data Analysis, Social Networks, Knowledge Extraction, Twitter, Google Trends, Regression, Users Interests, Television, Audience, TV Shows

Ευχαριστίες

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια Ιωάννα Ρουσσάκη για τη δυνατότητα που μου έδωσε να εργαστώ πάνω σε ένα σύγχρονο και άκρως ενδιαφέρον θέμα και για την εμπιστοσύνη που μου έδειξε όλους αυτούς τους μήνες. Οι συμβουλές και η καθοδήγηση που έλαβα ήταν πολύτιμες στην εκπόνηση της διπλωματικής αυτής εργασίας.

Στη συνέχεια, ευχαριστώ τον υποψήφιο διδάκτορα Νίκο Καλατζή για την πολύ σημαντική βοήθεια σε πρακτικά -και όχι μόνο- κομμάτια της εργασίας.

Ακόμα, ευχαριστώ ολόψυχα τους ΗΜΜΥτρελους για τα όσα όμορφα ζήσαμε όλα αυτά τα χρόνια και τα μέλη της EESTEC από όλη την Ευρώπη που κατάφεραν να χωρέσουν το κόκκινο χρώμα σε μια καταπράσινη καρδιά.

Η εργασία αυτή φυσικά αφιερώνεται στην οικογένειά μου για την αγάπη, τη στήριξη και την υπομονή που μου έδειξαν σε όλη μου τη ζωή.

Μάριος Παρασκευόπουλος
Αθήνα, 19 Δεκεμβρίου 2018

Περιεχόμενα

Ευρετήριο πινάκων

Ευρετήριο εικόνων

1. Εισαγωγή

1.1 Κοινωνικά δίκτυα

Παρότι συνδεδεμένη με τη σύγχρονη ψηφιακή εποχή, η ορολογία αυτή ακούστηκε πρώτη φορά στα τέλη του 19ου αιώνα όταν τόσο ο Émile Durkheim όσο και ο Ferdinand Tönnies προωθούσαν την ιδέα των κοινωνικών δικτύων στις θεωρίες τους και την έρευνα των κοινωνικών ομάδων. Ο Tönnies ισχυρίστηκε ότι οι κοινωνικές ομάδες μπορούν να υπάρχουν ως προσωπικοί και άμεσοι κοινωνικοί δεσμοί που είτε συνδέουν άτομα που μοιράζονται ίδιες, αξίες και πεποιθήσεις ("Gemeinshaft" ή "κοινότητα") είτε απρόσωπες, επίσημες και οργανικές κοινωνικές σχέσεις ("Gesellschaft" ή "κοινωνία").Ο Durkheim έδωσε μια μη-εξατομικευμένη εξήγηση των κοινωνικών γεγονότων υποστηρίζοντας ότι τα κοινωνικά φαινόμενα προκύπτουν όταν οι αλληλεπιδράσεις μεταξύ ατόμων αποτελούν μια πραγματικότητα που δεν μπορεί πλέον να ληφθεί υπόψη από την άποψη των ιδιοτήτων των μεμονωμένων παραγόντων. Ο Georg Simmel, στη στροφή του εικοστού αιώνα, επεσήμανε τη φύση των δικτύων και την επίδραση του μεγέθους του δικτύου στην αλληλεπίδραση και εξέτασε την πιθανότητα αλληλεπίδρασης σε χαλαρά πλεγμένα δίκτυα και όχι σε ομάδες.

Στη διάρκεια των επόμενων δεκαετιών και στη σκιά των πολέμων που ταλάνισαν την Ευρώπη και τον υπόλοιπο κόσμο, ο τομέας γνώρισε ιδιαίτερη άνθιση και πλήθος επιστημόνων ασχολήθηκαν με τη μελέτη τέτοιων δικτύων. Όντας κατεξοχήν διεπιστημονικό αντικείμενο, απασχόλησε μεταξύ άλλων ερευνητές από τους τομείς της ψυχολογίας, κοινωνιολογίας και ανθρωπολογίας πολύ πριν να έλθουμε στην εποχή της ψηφιοποίησης. Φυσικά σχετικές αναλύσεις γίνονται κι από επιστήμονες της βιολογίας, της στατιστικής, των μαθηματικών και της θεωρίας γράφων. Τα τελευταία χρόνια με την εξάπλωση των ηλεκτρονικών κοινωνικών δικτύων βρίσκει πρόσφορο έδαφος η μελέτη και ανάλυση για πολιτικούς, εμπορικούς και επιχειρηματικούς σκοπούς.

1.2 Μέσα κοινωνικής δικτύωσης

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. Μέσω αυτών παρέχεται η δυνατότητα σε άτομα να διευρύνουν τις κοινωνικές τους σχέσεις και να δημιουργήσουν φιλίες τόσο στα πλαίσια της τοπικής τους κοινωνίας όσο και σε διαφορετικές κοινωνίες σε όλο τον κόσμο. Επιπλέον μέσω των ηλεκτρονικών κοινωνικών δικτύων γίνεται η ενημέρωση των χρηστών για πρόσφατα γεγονότα και για σημαντικές ή λιγότερο σημαντικές πληροφορίες, τις οποίες άλλοι κοινοποιούν.

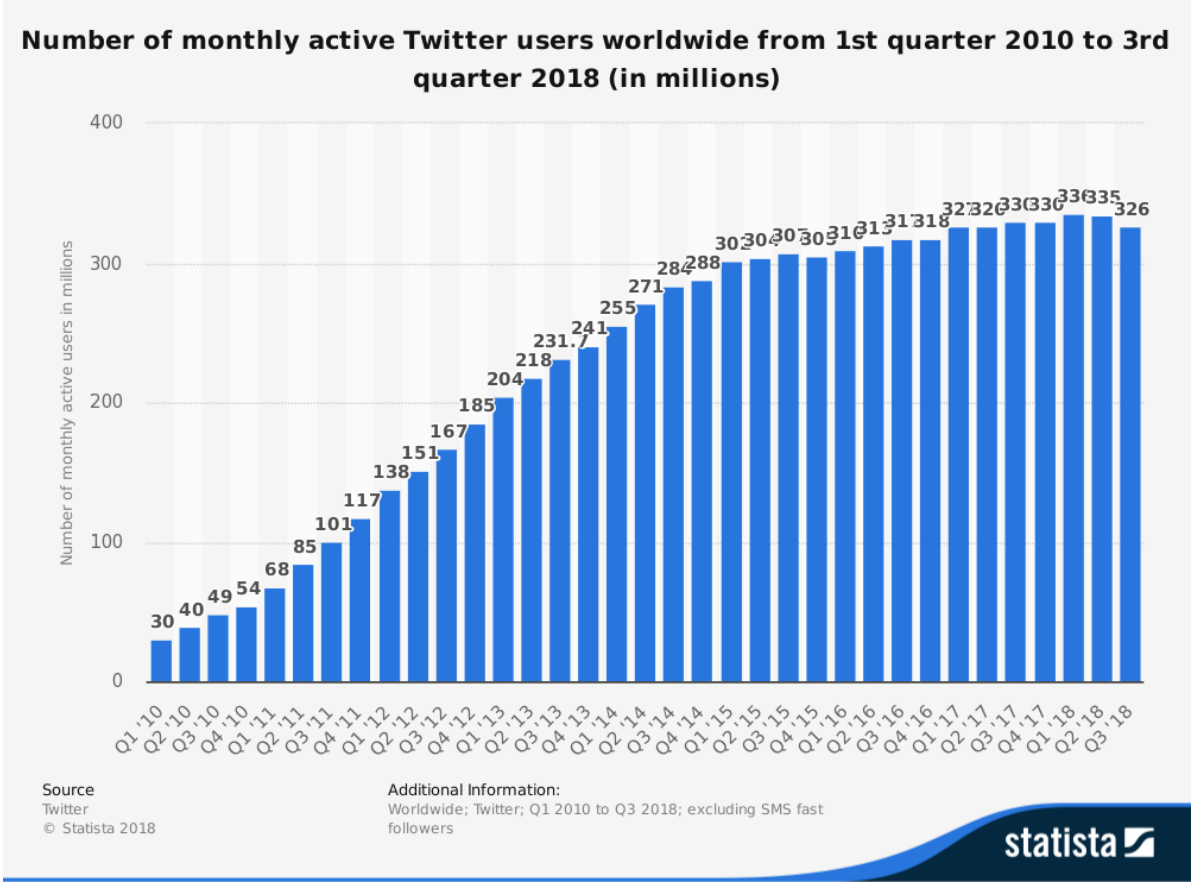
Μερικά παραδείγματα πολύ δημοφιλών ηλεκτρονικών κοινωνικών δικτύων είναι το Facebook, το Instagram, το Pinterest και φυσικά το Twitter.

1.2.1 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 280 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).

Ειδικότερα, το Twitter αυξάνει συνεχώς τους χρήστες του με αποκορύφωμα τους 336 εκατομμύρια ενεργούς χρήστες που κατέγραψε στο τέλος του πρώτου τριμήνου του 2018 σύμφωνα με την ιστοσελίδα Statista (βλ. Εικόνα 1.1). Κατά μέσο όρο οι χρήστες αυτοί δημοσιεύουν 500 εκατομμύρια tweets ημερησίως. 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 από ακαδημαϊκές 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:

- Επιστήμονες κοινωνιολογίας που μελετάνε τη συμπεριφορά των ανθρώπων στις online κοινωνίες και πιθανή διαφοροποίηση ανάλογα με τα δημογραφικά τους στοιχεία.
- 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.
- Ψυχολογικές έρευνες σχετικά με τα άτομα μιας συγκεκριμένης κοινότητας.
- Εταιρείες που θα μπορούν να προσαρμόσουν καλύτερα τις στοχευμένες διαφημίσεις και το δημόσιο πρόσωπο που δείχνουν ανάλογα με τα στοιχεία του κοινού στο οποίο απευθύνονται.
- Δημόσιους φορείς (πχ αστυνομία) που μπορεί να χρησιμοποιήσει δημογραφικά χαρακτηριστικά ως μέρος ερευνών.



Εικόνα 1.1: Αριθμός των ενεργών χρηστών του Twitter παγκοσμίως από το 2010 έως το 2018

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

- Όπως ανεφέρθη προηγουμένως, διαθέτει ένα μεγάλο αριθμό ενεργών χρηστών που παράγουν έναν ανεκτίμητο όγκο πληροφορίας, ο οποίος αποτελεί δικλείδα για την αξιοπιστία στα εξαγόμενα συμπεράσματα.
- Τα δεδομένα αυτά ανανεώνονται συνεχώς και έτσι είναι δυνατόν να επανεκτιμώνται παράλληλα τα χαρακτηριστικά που μελετούνται.
- Πρόκειται για μια απολύτως ελεύθερη πηγή πληροφορίας και δεδομένων και μάλιστα προσφέρονται easy-to-use APIs και λοιπές διεπαφές για την ανάκτηση του περιεχομένου και σχετικών μεταδεδομένων.
- Το Twitter έχει την επιλογή του geotagging, δηλαδή τη δημοσίευση ενός tweet μαζί με την πληροφορία της τοποθεσίας. Το geotagging χρησιμοποιείται όλο και περισσότερο πλέον και έτσι οι διάφορες έρευνες θα μπορούν στο μέλλον να στοχεύουν σε συγκεκριμένες γεωγραφικές τοποθεσίες.
- Έχει ήδη διεξαχθεί αρκετή έρευνα βασισμένη στο εν λόγω δίκτυο και υπάρχει υλικό από εφαρμοσμένες μεθοδολογίες και τα παραγόμενα αποτελέσματα και συμπεράσματα, τα οποία αποτελούν οδηγό για επέκταση και καινοτομία.

1.2.2 Google Trends

Το Google Trends είναι μια ιστοσελίδα που δημιουργήθηκε από την Google Inc το Μάιο του 2006 ώστε να παρέχει δεδομένα σχετικά με τη μηχανή αναζήτησης της εταιρείας. Συγκεκριμένα, δίνει συγκριτικά δεδομένα για οποιονδήποτε όρο αναζήτησης σε συνάφεια

με άλλους της επιλογής του χρήστη και μάλιστα με δυνατότητα γεωγραφικής και γλωσσικής επιλογής. Το ίδιο ισχύει και για σελίδες, εικόνες και πλέον και για το YouTube Search, μιας και η δημοφιλής πλατφόρμα φιλοξενίας βίντεο ανήκει στον όμιλο. Η ιστοσελίδα ανανεώνεται σε ικανοποιητικό βαθμό τελευταία, ωστόσο είναι γεγονός ότι έλαβε τη σημερινή της μορφή το Σεπτέμβριο του 2012, όταν και ένωσε τις δυνατότητες της με το Google Insights for Search.

Φαινομενικά η παραπάνω δυνατότητα είναι αρκετά περιορισμένη, ωστόσο η αξία της πλατφόρμας αυτής βρίσκεται στον όγκο δεδομένων τον οποίο συνοψίζει και οπτικοποιεί. Έχει υπολογιστεί ότι γίνονται περισσότερα από 2 τρις αναζητήσεις σε ετήσια βάση, που αντιστοιχεί σε περισσότερα από 5.5 δισεκατομμύρια ημερησίως. Μιλάμε για ένα τεράστιο όγκο δεδομένων, τον οποίον συνοψίζει η ιστοσελίδα σε αποτελέσματα ευπρόσιτα και εύκολα στην επεξεργασία. Αξίζει να σημειωθεί ότι μεταξύ των περιορισμών της πλατφόρμας είναι και η αδυναμία της Google να πρωτοστατήσει σε κάποιες αγορές για διάφορους λόγους, όπως πολιτικοί (Ρωσία) ή γλωσσολογικοί (Άπω Ανατολή).

Τα κυριότερα στοιχεία της πλατφόρμας είναι οι χρονοσειρές και οι χρωματικοί χάρτες. Οι χρονοσειρές είναι ακολουθίες δεδομένων που αναφέρονται στη διαχρονική εξέλιξη ενός φαινομένου. Στην προκειμένη αποδίδεται ο (σχετικός) όγκος αναζητήσεων ενός όρου με το χρόνο ως οριζόντιο άξονα. Στα γραφήματα αυτά δίνεται και η δυνατότητα ταυτόχρονης εμφάνισης περισσοτέρων χρονοσειρών για σύγκριση των όγκων αναζήτησης. Οι χρωματικοί χάρτες προσφέρονται για σύγκριση όρων σε γεωγραφικό επίπεδο για μια δεδομένη χρονική περίοδο που καθορίζεται από το χρήστη. Συγκεκριμένα, κάθε περιοχή χρωματίζεται με το χρώμα που υποδεικνύει το υπόμνημα για το δημοφιλέστερο όρο μεταξύ των επιλεχθέντων για τη σύγκριση. Αυτές είναι μόνο ορισμένες από τις δυνατότητες που δίνει η πλατφόρμα στους χρήστες της για αξιοποίηση αυτού του τεράστιου όγκου δεδομένων που διαθέτει η Google.

1.3 Συναφής βιβλιογραφία

1.3.1 Πρόβλεψη βάσει δεδομένων Twitter

Το Twitter έχει αξιοποιηθεί σε μεγάλο βαθμό τα τελευταία χρόνια με σκοπό την ανάπτυξη μοντέλων πρόβλεψης σε διάφορους τομείς. Οι Sinha, Dyer, Gimpel και Smith (2013) επιχειρούν ένα μοντέλο πρόβλεψης αποτελεσμάτων αμερικάνικου ποδοσφαίρου χρησιμοποιώντας των όγκο των δημοσιεύσεων. Συγκεκριμένα, το μοντέλο ορίζει μια συνάρτηση που συσχετίζει τις μεταβολές στους αριθμούς αυτούς για ορισμένες κατηγορίες. Η απόδοση των προβλέψεων κυμαίνεται στο 68%.

Οι O'Connor, Balasubramanyan, Routledge, Smith (2010) χρησιμοποιούν τεχνικές ανάλυσης περιεχομένου για τα tweet ώστε να διεξάγουν δημοσκοπήσεις και εκλογικές προβλέψεις. Αυτό επιτυγχάνεται αποδίδοντας σε κάθε λέξη θετικά και αρνητικά βάρη ανάλογα με τη σημασία τους βάσει ειδικών λεξικών. Το μοντέλο απέφερε συσχέτιση της τάξης του 80% σε σχέση με μετρήσεις κοινής γνώσης στην εκλογή Obama. Παρόμοια, οι Tumasjan, Sprenger, Sandner, Welpe (2010) αναφορικά με τις εκλογές του 2009 στη Γερμανία μετράνε τις εμφανίσεις των κομμάτων σε δημοσιεύσεις χωρίς κάποιο μοντέλο πρόβλεψης.

Οι Paul, Dredze (2011) χρησιμοποιούν ανάλυση συχνότητας για ορολογία σχετική με συμπτώματα και προβλήματα με σκοπό την εποπτεία μεταδοτικών παθήσεων. Αντίστοιχα έχει μελετηθεί η εποχική γρίπη από αρκετές ομάδες, μεταξύ άλλων από τους Achrekar, Gandhe, Lazarus, Yu, Liu (2012).

Η πλατφόρμα έχει αξιοποιηθεί και στον τομέα της οικονομίας και του μάρκετινγκ. Οι Bollen, Mao, Zeng (2011) χρησιμοποιούν νευροασαφή δίκτυα και χρονοσειρές

επιτυγχάνοντας απόδοση 86% στην πρόβλεψη της πορείας χρηματιστηριακών δεικτών. Άλλες σχετικές εφαρμογές που αναδεικνύουν τη δύναμη της πληροφορίας από το Twitter είναι η ανίχνευση εγκληματικών ενεργειών μέσω της δυνατότητας να εντοπιστεί η απαρχή συγκεκριμένων κρίσιμων περιπτώσεων, όπως συζητήσεις επί μεταφορικών δυστυχημάτων και πυρκαγιών μετά από έρευνα των Wang, Gerber, Brown (2012) και η εκτίμηση πλήθους σε συγκεκριμένες τοποθεσίες, όπως τα αεροδρόμια με τη χρήση γραμμικού μοντέλου επί του όγκου δημοσιεύσεων στο Twitter και συνδυασμό με δεδομένα κινητής τηλεφωνίας. Σε πλήρη αντιστοιχία με την προηγούμενη έρευνα οι Chauhan, Kummamuru, Toshniwal (2016) ανιχνεύουν κατάλληλα μέρη για επίσκεψη και τουρισμό. Οι Grasso, Zaza, Zabini, Pantaleo, Nesi, Crisci και Gozzini (2016) εξέδωσαν δύο δημοσιεύσεις σχετικές με την αξιολόγηση των πληροφοριών από τα δελτία καιρού.

Μια ακόμα μεγάλη συνεισφορά των δεδομένων από το twitter είναι στην πρόβλεψη της επιτυχίας που θα σημειώσουν οι τελευταίες παραγωγές του κινηματογράφου. Οι Asur και Huberman (2010) προτείνουν ένα μοντέλο που δέχεται το μέσο ρυθμό δημοσιεύσεων, την πιθανή ύπαρξη ηλεκτρονικών διευθύνσεων (URLs) στα tweets και το πλήθος αναδημοσιεύσεων ως χαρακτηριστικά εισόδου. Χρησιμοποιούν χρονοσειρές μήκους μερικών ημερών και επιτυγχάνουν R^2 της τάξης του 0,94 μέσω ενός γραμμικού μοντέλου που αξιοποιεί επιπρόσθετα την ανάλυση περιεχομένου. Στο ίδιο πεδίο έχουν επιχειρηθεί αρκετές ακόμα προσεγγίσεις που συνδυάζουν μετρικές όγκου δραστηριότητας με ανάλυση περιεχομένου σε μακριές χρονοσειρές χωρίς ωστόσο να προτείνονται συγκεκριμένα μοντέλα. Μερικές τέτοιες απόπειρες βρίσκονται στις σημοσιεύσεις των Mishne και Glance (2006), Sitaram και Huberman (2010), Leskovec (2011) και Lu, Kruger, Thom, Wang, Koch, Ertl και Maciejewski (2014). Τέλος, οι Reddy, Kasat και Jain (2012) υλοποιούν ένα σύστημα ασαφών συμπερασμάτων αξιοποιώντας μετρικές όπως ο αριθμός των tweets, οι followers και μετρικές ανάλυσης περιεχομένου, καθώς επίσης και συμπληρωματικές πληροφορίες για τη βαθμολογία ηθοποιών. Σε ορισμένες περιπτώσεις παρατηρείται μεγάλο μέσο τετραγωνικό σφάλμα που κυμαίνεται μεταξύ 6% και 27%.

1.3.2 Μηχανική μάθηση για την εκτίμηση τηλεθέασης από κοινωνικά δίκτυα

Έχουν πραγματοποιηθεί αρκετές έρευνες τα τελευταία χρόνια βασισμένες σε δεδομένα από το Twitter για ανάλυση σχετική με το ενδιαφέρον για τηλεοπτικές εκπομπές. ο Giglietto (2013) αξιοποιεί δεδομένα από την πλατφόρμα σε συνδυασμό με τον αριθμό τηλεθεατών προηγούμενων επεισοδίων για την πρόβλεψη του ποσοστού σε μακροχρόνιες πολιτικές εκπομπές (από 14 έως 280 εκπομπές), χρησιμοποιώντας κυρίως αριθμητκά δεδομένα, όπως το πλήθος των δημοσιεύσεων κατά τη διάρκεια της ροής της εκπομπής, καθώς και το σχετικό ρυθμό. Το αποτέλεσμα είναι της τάξης του 0,95 για τη μετρική R^2.

Οι Sereday και Cui από την ομάδα Nielsen Media Research χρησιμοποιούν μηχανική μάθηση πάνω σε δεδομένα του Twitter για να προβλέψουν την τηλεθέαση διαφόρων τηλεοπτικών προγραμμάτων και ιδιαίτερα τη μεγάλη διασπορά της τάξης των 2/3 που παρατηρείται στην απόκλιση των μεγεθών που σημείωσαν στην πρεμιέρα τους. Η τηλεθέαση συνήθως υπολογίζεται από μηχανήματα εγκατεστημένα σε δέκτες επιλεγμένους δειγματοληπτικά και από μετρητές στους αποκωδικοποιητές συνδρομητικών υπηρεσιών. Η ομάδα βασίζεται σε δημογραφικά στοιχεία και τα προηγούμενα νούμερα που έχουν καταγραφεί και εφαρμόζουν τεχνικές όπως τα δέντρα αποφάσεων για τις προβλέψεις τους.

Οι Hsieh, Chou, Cheng, Wu (2013) δημιούργησαν ένα πολυεπίπεδο νευρωνικό δίκτυο (MLP) και πραγματοποίησαν ανάλυση τηλεθέασης βασιζόμενοι σε δεδομένα από το Facebook. Προβλέποντας για διαφορετικές εκπομπές επέτυχαν μέσο (απόλυτο) ποσοστιαίο σφάλμα μεταξύ 6% και 24%.

Οι Molteni και Ponce De Leon (2016) αναλύουν το περιεχόμενο των δημοσιεύσεων από το Twitter για να εκτιμήσουν το κοινό αμερικάνικων σειρών. Μια σημαντική καινοτομία που εισήγαγαν είναι η ομαδοποίηση των εκπομπών σε 3 clusters και η αναζήτηση ενός μοντέλου για κάθε ομάδα, μια και παρατηρήθηκαν σημαντικές διαφορές τόσο στη δημοτικότητα όσο και στα δεδομένα από το κοινωνικό δίκτυο για τις εκπομπές που ανήκουν σε κάθε cluster. Το μοντέλο που προέκυψε βασίζεται στη γραμμική παλινδρόμηση με τη χρήση δεδομένων όπως ο όγκος δημοσιεύσεων και η βαθμολογία από ανάλυση περιεχομένου. Η μετρική R^2 κυμαίνεται μεταξύ 0,73 και 0,94 για τα 3 groups. Χαρακτηριστικό είναι ότι δεν επιλέχθηκε η διασταύρωση (cross validation) των προβλέψεων, μιας και χρησιμοποιήθηκαν στιβαρές εκπομπές με μεγάλο αριθμό επεισοδίων. Ταυτόχρονα υπάρχουν σημαντικές διαφορές σε σχέση με τις εκπομπές ζωντανού χρόνου, όπου συχνά το κοινό καλείται να συμμετάσχει μέσω των κοινωνικών δικύων, εκτοξεύοντας τη σχετική δραστηριότητα.

Οι Sommerdijk, Sanders και van den Bosh διαπιστώνουν σημαντικές συσχετίσεις ανάμεσα στον αριθμό των δημοσιεύσεων σε εύρος μισής ώρας πριν και μετά την εκπομπή και σε διαδοχικά επεισόδια χωρίς ωστόσο να προτείνεται ένα μοντέλο πρόβλεψης. Στην εργασία των Wakamiya, Lee και Sumiya (2011) επιχειρείται μια λειτουργική σύγκριση μεταξύ των κλασικών λύσεων για την εκτίμηση των ποσοστών τηλεθέασης εκπομπών με χρήση σχετικών δεδομένων, μαζί με μια πρώιμη λύση που βασίζεται από κοινού σε χωρική, χρονική και συνάφεια σε επίπεδο κειμένου, επίσης χωρίς μοντέλο.

Σύμφωνα με τη σχετική ανάλυση πάνω στις σύγχρονες τεχνολογίες εκτίμησης τηλεθέασης εκπομπών, οι δυνατότητες πρόβλεψης που προσφέρει το Twitter ερμηνεύονται κυρίως στη χρήση μετρικών όγκου, όπως ο συνολικός αριθμός (ανα)δημοσιεύσεων που συνδέονται με συγκεκριμένους χρήστες ή ετικέτες. Ωστόσο σε ορισμένες περιπτώσεις απαιτείται μία βαθύτερη κατανόηση των tweets, ώστε να δημιουργηθούν ισχυρότερα προγνωστικά μοντέλα. Εξού και η ανάγκη για τη δημιουργία αλγορίθμων επεξεργασίας φυσικής γλώσσας (NLP) για ανάλυση περιεχομένου, ώστε η σημασία των δημοσιεύσεων να λαμβάνεται υπόψη στα μοντέλα, όπως υποδεικνύουν και οι O' Connor, Balasubramanyan, Routledge και Smith (2010). Σημαντικό ρόλο στην ακρίβεια των προβλέψεων μπορεί να έχουν και τεχνικές κατάτμησης, φιλτραρίσματος και ομαδοποίησης βάσει περιεχομένου (πχ NLP για την εξακρίβωση συνάφειας tweet με το θέμα) ή και στοιχείων των χρηστών (φύλο, ηλικία, τόπος διαμονής). Μια σύνοψη των προγνωστικών μεθόδων επιχειρείται από τους Sikdar, Adali, Amin, Abdelzaher, Chan, Cho, Kang, O'Donovan (2014).

Οι Ιταλοί Crisci κ.α. αναλύουν γλωσσικά tweets που προέρχονται από τη δραστηριότητα χρηστών που παρακολουθούν και σχολιάζουν reality shows της γείτονος χώρας και σε συνδυασμό με μετρικές όπως ο αριθμός των tweets και retweets ανά επεισόδιο προβλέπουν την ακροαματικότητα με το μοντέλο της γραμμικής παλινδρόμησης. Οι Madlberger και Almansour (2014) επιπρόσθετα σχολιάζουν κάποιες από τις προγνωστικές δυνατότητες των προτεινόμενων μοντέλων. Πράγματι, κάποιες προσεγγίσεις προτείνουν κάποια γενικευμένα μοντέλα, τα οποία έχουν προέλθει από φιλτράρισμα και κατηγοριοποίηση βασισμένη (και) στον ανθρώπινο παράγοντα. Κατ' αυτόν τον τρόπο γίνεται πιο δύσκολο να αναπαραχθεί και εν τέλει να γενικευτεί η λύση.

1.3.3 Προβλέψεις βάσει δεδομένων Google Trends

Όσο για το Google Trends, δεν έχει υπάρξει μέχρι τώρα κάποια απόπειρα για τη συσχέτιση της τηλεθέασης με τα δεδομένα από το Google Search, ωστόσο έχουν υπάρξει ικανοποιητικά αποτελέσματα σε άλλους τομείς που σχετίζονται με την ανθρώπινη δραστηριότητα. Οι Ginsberg, Mohebbi, Patel, Brammer, Smolinski και Brilliant (2009), εργαζόμενοι της Google ήταν οι πρώτοι που έσπευσαν να αναδείξουν τις δυνατότητες της

πλατφόρμας και αξιοποίησαν δεδομένα για την ανίχνευση εξάρσεων του ιού της γρίπης. Συγκεκριμένα, παρατήρησαν υψηλά ποσοστά συνάφειας μεταξύ αναζητήσεων που σχετίζονται με την ασθένεια της γρίπης και του αριθμού των αντίστοιχων ασθενών που επισκέπτονται το σύστημα υγείας. Αναλύοντας τα παραπάνω δεδομένα από το Google Trends καταφέρνουν να αποδείξουν τη δυναμική της πλατφόρμας ως αντιπροσωπευτική πηγή ανθρώπινης δραστηριότητας και εργαλείο ανίχνευσης τέτοιων φαινομένων.

Η ομάδα αυτή επέτυχε το σκοπό της, αφού δεδομένα από την πλατφόρμα χρησιμοποιήθηκαν αρκετά για επιδημιολογικούς και γενικότερους ιατρικούς σκοπούς. Χαρακτηριστική είναι η έρευνα των Schootman, Toor, Cavazos-Rehg, Jeffe, McQueen, Eberth και Davidson (2014) όπου μελέτησαν τη συνάφεια δεδομένων από το Google Trends για τον προσυμπτωματικό έλεγχο για καρκίνο σε σχέση με τα πραγματικά δεδομένα που προκύπτουν από πολυδάπανες αναλύσεις. Τα αποτελέσματα κυμάνθηκαν από το 0,14 στο 0,55 δείχνοντας πως η πλατφόρμα αποτελεί ιδανικό συμπλήρωμα, αλλά δεν μπορεί να αντικαταστήσει τις παραδοσιακές τεχνικές ανάλυσης.

Σημαντικές έρευνες σημειώθηκαν και σε άλλους κλάδους, όπως για εμπορικούς σκοπούς και συγκεκριμένα για ανάλυση της συμπεριφοράς των καταναλωτών. Οι Goel, Hofman, Lahaie, Pennock και Watts (2010) αποδεικνύουν ότι αυτά που αναζητούν οι χρήστες στο διαδίκτυο μπορούν να χρησιμοποιηθούν για την πρόβλεψη της συλλογικής συμπεριφοράς μέρες ή και βδομάδες νωρίτερα. Μεταξύ άλλων καταπιάστηκαν με τον κλάδο των πωλήσεων, τόσο για οχήματα όσο και για κατοικίες. Οι Choi και Varian (2012) έκαναν σχετικές έρευνες.

Ιδιαίτερο ενδιαφέρον στις μέρες μας παρουσιάζει και η μελέτη κοινωνικοπολιτικών δεδομένων. Οι Askitas και Zimmerman (2009) μελετούν τη σύνδεση και συσχέτιση συγκεκριμένων όρων αναζήτησης και των δεικτών ανεργίας στην απαρχή της παγκόσμιας κρίσης. Οι Mavragani και Tsagarakis (2015) διερεύνησαν την αναλογία Ναι/Όχι στις σχετικές αναζητήσεις πριν το ελληνικό δημοψήφισμα. Το αποτέλεσμα της έρευνας έδειξε σαφή υπεροχή της επιλογής "Όχι", σε σύμπνοια με το τελικό εκλογικό αποτέλεσμα και κατατροπώνοντας τις πρότερες προγνώσεις που έδειχναν εκλογικό ντέρμπι. Έτσι συνέβαλαν στη συζήτηση για το αν η πλατφόρμα μπορεί να αποτελέσει εργαλείο πρόγνωσης εκλογικών αναμετρήσεων στο μέλλον. Οι Bulut και Dogan (2018) χρησιμοποιούν το Google Trends προσπαθώντας να προσομοιώσουν την ισοτιμία αμερικάνικου δολλαρίου και τούρκικης λίρας που ως γνωστόν διαγράφει τρομερές μεταβολές τα τελευταία χρόνια, λόγω της αστάθειας του τουρκικού νομίσματος.

1.4 Συμβολή παρούσας προσέγγισης

Η παρούσα διπλωματική εργασία επικεντρώνεται στο συνδυασμό των δεδομένων από το Twitter και το Google Trends για την εκτίμηση της ακροαματικότητας τηλεοπτικών εκπομπών βάσει δραστηριότητας των χρηστών κοινωνικών δικτύων, όπως τα αναφερθέντα. Για τους σκοπούς της μελέτης μας χρησιμοποιούνται δεδομένα από το Ιταλικό reality show "Amici di Maria de Filippi". Η προσέγγιση είναι ανεξάρτητη της γλώσσας και μπορεί να γενικευτεί για οποιαδήποτε περιοχή παρουσιάζει σχετικό ενδιαφέρον. Τα ανακτηθέντα δεδομένα επεξεργάζονται με μεθόδους μηχανικής μάθησης και κυρίως τη γραμμική παλινδρόμηση. Για την μοντελοποίηση και αυτοματοποίηση χρησιμοποιείται η γλώσσα προγραμματισμού Python. Στόχοι της εργασίας είναι οι ακόλουθοι:

- Η επαλήθευση για τη συνάφεια και συσχέτιση μεταξύ των δεδομένων από τις δύο πλατφόρμες και συνεπώς η δημιουργία πρόσφορου εδάφους για την αξιοποίηση των Google trends στα audience analytics.

- Η δημιουργία μοντέλου μηχανικής μάθησης για την εκτίμηση της τηλεθέασης για την υπό μελέτη εκπομπή με χρήση των παραπάνω δεδομένων ως είσοδο.
- Η ανάλυση των παραπάνω δεδομένων για την εύρεση μοτίβων δημογραφικής φύσεως.

1.5 Διάρθρωση εργασίας

Στο κεφάλαιο 2 παρουσιάζεται αρχικά η δομή και τα χαρακτηριστικά που έχει μια ανάρτηση στο Twitter, ένα tweet. Έπειτα, γίνεται μια εισαγωγή στη Μηχανική Μάθηση ως κλάδος της Πληροφορικής με έμφαση στην Παλινδρόμηση και δίνεται το θεωρητικό υπόβαθρο για τις τεχνικές που χρησιμοποιήθηκαν στην παρούσα εργασία. Στο κλείσιμο του κεφαλαίου παρουσιάζεται η γλώσσα προγραμματισμού Python και οι βιβλιοθήκες που χρησιμοποιήθηκαν για να παρασταθούν και πραγματοποιηθούν τα πειράματα.

Στο κεφάλαιο 3 επιχειρείται μια πιο πρακτική ανάλυση των αλγορίθμων που χρησιμοποιήθηκαν στα πειράματα και παρουσίαση του μαθηματικού και βιολογικού υπόβαθρου. Συγκεκριμένα γίνεται μια περιγραφή της μεθόδου ελαχίστων τετραγώνων και του αλγορίθμου καθόδου, καθώς και μια εκτενής ανάλυση της λειτουργίας των γενετικών αλγορίθμων με περιγραφή τόσο των βιολογικών όρων που προσομοιώνονται προγραμματιστικά όσο και των μεθόδων και τεχνικών που χρησιμοποιούνται κατά την εφαρμογή ενός γενετικού αλγόριθμου.

Στο κεφάλαιο 4 ξεκινάει το πιο πρακτικό κομμάτι της εργασίας, περιγράφοντας τη διαδικασία ανάκτησης των δεδομένων από τις δύο πλατφόρμες που μελετάμε, τα εργαλεία και προγράμματα που χρησιμοποιήθηκαν καθώς και οι μέθοδοι επεξεργασίας των αρχικών δεδομένων για τη δημιουργία των datasets, πάνω στα οποία βασίζονται τα πειράματα της παρούσας εργασίας.

Στο κεφάλαιο 5 παρουσιάζονται τα κυριότερα αποτελέσματα των πειραμάτων και αναλύονται σημαντικές λεπτομέρειες στις υλοποιήσεις των επιμέρους τμημάτων κώδικα και scripts στη γλώσσα Python. Αρχικά γίνεται από τη σύγκριση μεταξύ των δεδομένων από τις δημοσιεύσεις στο Twitter και των δεδομένων αναζήτησης από τα Google Trends και την επαλήθευση της συνάφειάς τους. Έπειτα περιγράφεται η διαδικασία επιλογής της κατάλληλης μορφής των δεδομένων εισόδου, όπως το χρονικό παράθυρο επιλογής δημοσιεύσεων. Στη συνέχεια εκτελούνται τα πειράματα για διαφορετικούς τύπους παλινδρόμησης, προκειμένου να επιλεγεί το καταλληλότερο μοντέλο και τα αντίστοιχα δεδομένα εισόδου για το πρόβλημα της εκτίμησης της ακροαματικότητας τηλεοπτικών εκπομπών. Χρησιμοποιείται γραμμική, πολυωνυμική και γενική μη γραμμική παλινδρόμηση καθώς επίσης και μια υβριδική υλοποίηση με ανάμεικτες εισόδους από το γραμμικό και μη γραμμικό μοντέλο. Τα αποτελέσματα των παραπάνω μοντέλων επαληθεύονται με μια υλοποίηση γραμμικής παλινδρόμησης ως γενετικό αλγόριθμο. Στο τέλος γίνεται μια σύνοψη των πειραματικών εξαγομένων.

Στο κεφάλαιο 6 παρουσιάζονται τα τελικά συμπεράσματα που προκύπτουν από την εργασία και προτείνονται επεκτάσεις που είτε θα αποτελέσουν συνέχεια της δουλειάς μας είτε μπορούν να αξιοποιηθούν ως ιδέες από άλλους ερευνητές.

2. Υπόβαθρο διπλωματικής

2.1 Δομή ενός tweet

Το βασικότερο αντικείμενο που χρησιμοποιείται στην παρούσα μελέτη είναι το tweet, δηλαδή η ανάρτηση που κάνει ένας χρήστης στο Twitter. Τα στοιχεία που το απαρτίζουν, όπως φαίνονται -τα περισσότερα- και στην Εικόνα 2.1 είναι τα εξής:

- Το αναγνωριστικό/id (φαίνεται στο permalink του tweet κι όχι στην εικόνα)
- Το permalink όπου βρίσκεται και είναι διαθέσιμο το tweet
- Το μοναδικό όνομα χρήστη/username, με το οποίο αναγνωρίζεται ο χρήστης
- Το πλήρες όνομα του χρήστη
- Η μικρογραφία (thumbnail) της εικόνας χρήστη
- Το περιεχόμενο του tweet ως κείμενο, σύνδεσμος ή/και εικόνα με μέγιστο αριθμό 280 χαρακτήρων
- Η ημερομηνία και ώρα ανάρτησης
- Το πλήθος πιθανών αναδημοσιεύσεων (retweets) από άλλους χρήστες
- Η επισήμανση "μου αρέσει" (favorites) που μπορεί να γίνει από ακόλουθους
- Οι αναφορές (mentions) σε άλλους χρήστες με χρήση του '*@username*' στο κείμενο
- Οι "ετικέτες" (hashtags) με χρήση του '*#hashtag*' για κατηγοριοποίηση του tweet
- Προαιρετικά, η τοποθεσία του χρήστη τη στιγμή του tweet (geotagging)

Στην παρούσα εργασία ιδιαίτερη σημασία παίζουν τα hashtags, η ημερομηνία και το username, ωστόσο πολύ σημαντικά στοιχεία μπορεί να είναι και τα retweets, τα favorites και φυσικά το κείμενο.

Εικόνα 2.1: Παράδειγμα ενός tweet πλήρους από χαρακτηριστικά

2.2 Μηχανική μάθηση

Η Μηχανική Μάθηση (Machine Learning, ML) είναι ένας κλάδος της Επιστήμης Υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή Νοημοσύνη. Μπορεί να οριστεί ως το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν (τυπικά "να βελτιώνουν προοδευτικά την επίδοσή τους σε μια συγκεκριμένη εργασία") από δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά, χωρίς να προγραμματίζονται ρητά για αυτό το σκοπό. Το πεδίο της ML έχει εφαρμογή μεταξύ άλλων στην Υπολογιστική Όραση, στην Οπτική Αναγνώριση Χαρακτήρων (OCR) και στο φιλτράρισμα των email.

Οι αλγόριθμοι ML χωρίζονται σε 3 κατηγορίες: Επιτηρούμενη Μάθηση (Supervised Learning), Μη Επιτηρούμενη Μάθηση (Unsupervised Learning), Ενισχυτική Μάθηση (Reinforcement Learning). Ως επιτήρηση (ή επίβλεψη) νοείται η ύπαρξη των επιθυμητών αποτελεσμάτων ως είσοδος στο πρόγραμμα και στόχος είναι η εύρεση ενός γενικού κανόνα αντιστοίχισης των δεδομένων στα αποτελέσματα. Στη δεύτερη περίπτωση, δεν παρέχεται κάποια εμπειρία στον αλγόριθμο και στόχος είναι η αναγνώριση μοτίβων στα δεδομένα, είτε ως αυτοσκοπός (ομαδοποίηση) ή για την ανακάλυψη κάποιων χαρακτηριστικών. Η Ενισχυτική Μάθηση αφορά ένα πρόγραμμα που αλληλεπιδρά με ένα δυναμικό περιβάλλον ώστε να επιτύχει κάποιο στόχο, χωρίς να γνωρίζει ρητά αν οι κινήσεις που επιλέγει είναι προς τη σωστή κατεύθυνση. Το μόνο που γνωρίζει είναι κάποια συνάρτηση "ανταμοιβής"

που θέλει να μεγιστοποιήσει, είτε πρόκειται για το σκορ σε κάποιο παίγνιο ή το αναμενόμενο σήμα από το περιβάλλον.

Τα δεδομένα που αποτελούν είσοδο και επεξεργάζονται από τους αλγόριθμους ML διαθέτουν κάποια χαρακτηριστικά (features). Πρόκειται ουσιαστικά για μετρήσιμες ιδιότητες των δεδομένων και χωρίζονται ανάλογα τις τιμές τους σε συνεχή και διακριτά (ή κατηγορικά). Ένα χαρακτηριστικό συνεχών τιμών είναι το ύψος και αντίστοιχα για διακριτές τιμές το χρώμα ή μια οποιαδήποτε δυαδική μεταβλητή. Τα χαρακτηριστικά των δεδομένων δίνουν την απαραίτητη γνώση στους αλγόριθμους ML ώστε να εκτελέσουν με επιτυχία την επιθυμητή εργασία (ομαδοποίηση, εκτίμηση/πρόβλεψη εξόδου).

Στη συνέχεια παρουσιάζονται οι τεχνικές και αλγόριθμοι που χρησιμοποιήθηκαν στην εργασία αυτή.

2.2.1 Η τεχνική της παλινδρόμησης

Η παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων σε μία πραγματική συνεχή μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν αυτή είναι διακριτή. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη εξ αυτών, που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο προϊόν ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

Το Γενικευμένο Νευρωνικό Δίκτυο Παλινδρόμησης (Generalized Regression Neural Network - GRNN) μπορεί να εκτελέσει διεργασίες παλινδρόμησης για να κατασκευάσει ένα μοντέλο παλινδρόμησης. Τα μοντέλα παλινδρόμησης περιλαμβάνουν τις ακόλουθες μεταβλητές: Οι άγνωστες παράμετροι συσχέτισης που δηλώνονται ως ένα διάνυσμα β (διάνυσμα βαρών), οι ανεξάρτητες μεταβλητές (διάνυσμα X) και φυσικά η εξαρτώμενη μεταβλητή Y.

Ένα μοντέλο παλινδρόμησης συσχετίζει το Y με μια συνάρτηση των X και β, δηλαδή $Y \cong F(X, \beta)$. Ο συνήθης φορμαλισμός είναι $E(Y|X) = F(X, \beta)$.

2.2.1.1 Γραμμική παλινδρόμηση

Η μοντελοποίηση μπορεί να γίνει χωρίς να είναι γνωστή από πριν η γνώση για τον τρόπο με τον οποίο συνδέεται η εξαρτημένη μεταβλητή από τις ανεξάρτητες και τότε ονομάζεται εμπειρική μοντελοποίηση. Στην γραμμική παλινδρόμηση συγκεκριμένα, η απαίτηση του μοντέλου που θα παραχθεί είναι μία: η εξαρτημένη μεταβλητή y να είναι ένας γραμμικός συνδυασμός των ανεξαρτήτων μεταβλητών x_i.

Στο απλό γραμμικό μοντέλο υπάρχει μόνο μια ανεξάρτητη μεταβλητή και το μοντέλο έχει τη μορφή $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ όπου ε το σφάλμα της εκτίμησης. Για περισσότερες μεταβλητές περνάμε στο γενικό γραμμικό μοντέλο και αυξάνεται ανάλογα το μέγεθος των διανυσμάτων β και X, πάντα όμως παραμένει η γραμμική εξάρτηση του y από τις ανεξάρτητες μεταβλητές X.

2.2.1.2 Πολυωνυμική παλινδρόμηση

Το παρόν μοντέλο έχει μεγάλη συνάφεια με το γραμμικό. Συγκεκριμένα αν το δούμε από στατιστικής απόψεως, πρόκειται επίσης για μοντέλο γραμμικό ως προς τα βάρη β, τα οποία και υπολογίζει ο αλγόριθμος. Η διαφορά έγκειται στο ότι πλέον οι μεταβλητές y συσχετίζονται με κάποιο πολυώνυμο των ανεξάρτητων μεταβλητών x κι όχι με τις ίδιες τις τιμές απαραίτητα. Χαρακτηριστικά, για μία μεταβλητή έχουμε το πολυωνυμικό μοντέλο $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \varepsilon$ σε πλήρη αντιστοιχία με τα προηγούμενα. Για περισσότερες μεταβλητές έχουμε όλα τα δυνατά πεπλεγμένα πολυώνυμα μέχρι n βαθμού.

2.2.1.3 Μη Γραμμική παλινδρόμηση

Σε αυτή την κατηγορία εντάσσονται όλα τα μοντέλα που δεν υπάρχει γραμμικότητα στο διάνυσμα β. Ένας τρόπος να χειριστούμε δεδομένα που φαίνεται να αντιστοιχούν σε τέτοια μοντέλα είναι η γραμμικοποίηση. Αυτό μπορεί να συμβεί με δύο τρόπους, είτε με μετατροπή των δεδομένων ή με κατάτμηση σε γραμμικές συστάδες. Παρακάτω θα εξηγηθεί περισσότερο η διαδικασία της μετατροπής.

2.2.2 Γενετικοί Αλγόριθμοι

Οι Γενετικοί αλγόριθμοι ανήκουν στο κλάδο της επιστήμης υπολογιστών και αποτελούν μια μέθοδο αναζήτησης βέλτιστων λύσεων σε συστήματα που μπορούν να περιγραφούν ως μαθηματικό πρόβλημα. Είναι χρήσιμοι σε προβλήματα που περιέχουν πολλές παραμέτρους και δεν υπάρχει (προφανής) αναλυτική μέθοδος που να μπορεί να βρει το βέλτιστο συνδυασμό τιμών για τις μεταβλητές ώστε το υπό εξέταση σύστημα να αντιδρά με όσο το δυνατόν με το επιθυμητό τρόπο.

Ο τρόπος λειτουργίας των Γενετικών Αλγορίθμων είναι εμπνευσμένος από τη βιολογία. Όταν το 1859 ο Δαρβίνος ανέπτυσσε στα νησιά Γκαλαπάγκος τη θεωρία της εξέλιξης, σίγουρα δεν περίμενε ότι η ιδέα αυτή θα αποτελέσει τη σπουδαιότερη -κατά πολλούς- ανακάλυψη στον τομέα της βιολογίας, πολλώ δε μάλλον ότι θα εμπνεύσει επιστήμονες διαφόρων άλλων τομέων, όπως της πληροφορικής. Περισσότερα από 150 χρόνια μετά έχουμε αλγορίθμους που προσομοιώνουν τις λειτουργίες της γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης.

Στο επόμενο κεφάλαιο περιγράφεται αναλυτικά ο τρόπος υλοποίησης των Γενετικών Αλγορίθμων και οι συσχετίσεις με τις αντίστοιχες βιολογικές λειτουργίες.

2.2.3 Άλλοι αλγόριθμοι παλινδρόμησης

Φυσικά υπάρχουν πολλά διαφορετικά μοντέλα και αλγόριθμοι για το πρόβλημα αυτό, μεταξύ άλλων και παραλλαγές ταξινομητών για να μπορούν να εφαρμόζουν παλινδρόμηση. Οι αλγόριθμοι Support Vector Machines (SVM) και k-nearest neighbors (k-nn) είναι οι χαρακτηριστικότεροι αλγόριθμοι ταξινόμησης που μπορούν να τροποποιηθούν για να λύσουν προβλήματα παλινδρόμησης. Ωστόσο αποδίδουν καλύτερα σε προβλήματα λογιστικής παλινδρόμησης και δεν αποτελούν καλές λύσεις στο πρόβλημα της εκτίμησης τηλεθέασης.

Ένας αλγόριθμος που θα μπορούσε να χρησιμοποιηθεί είναι τα δέντρα αποφάσεων (decision trees). Τα δέντρα αυτά υπολογίζουν την τιμή εξόδου παίρνοντας διαδοχικές αποφάσεις έπειτα από παρατηρήσεις επί των δεδομένων εισόδου. Παρότι έχουν αρκετά καλή απόδοση σε διάφορα προβλήματα παλινδρόμησης -με χαρακτηριστικότερο την

εκτίμηση αξίας ακινήτων- δεν αποτελούν καλή επιλογή στο συγκεκριμένο πρόβλημα. Ο σημαντικότερος ανασταλτικός παράγοντας είναι το γεγονός ότι το πρόβλημα της εκτίμησης τηλεθέασης έχει πολύ στενό εύρος τιμών και ειδικά στο δικό μας dataset είναι ακόμα πιο έντονο το φαινόμενο. Επομένως η εκτίμηση βάσει διχοτόμησης και διαδοχικών αποφάσεων ενέχει τον κίνδυνο μεγάλης αστοχίας, ακόμα και σε δεδομένα που δεν αποτελούν outliers (έκτοπα στοιχεία).

2.3 Η γλώσσα προγραμματισμού Python

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού η οποία δημιουργήθηκε από τον Ολλανδό Guido van Rossum το 1990 και αναπτύσσεται ως ελεύθερο λογισμικό υπό τη διαχείριση του μη κερδοσκοπικού οργανισμού Python Software Foundation. Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της και το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από όσες πιθανόν να χρειάζονταν σε γλώσσες όπως η C++ ή η Java. Το όνομά της προέρχεται από την περίφημη σατυρική ομάδα Βρετανών κωμικών Monty Python.

Αρχικά, η Python ήταν γλώσσα σεναρίων που χρησιμοποιούνταν στο λειτουργικό σύστημα Amoeba, ικανή και για κλήσεις συστήματος. Η Python 2.0 κυκλοφόρησε στις 16 Οκτωβρίου του 2000. Στις 3 Δεκεμβρίου 2008 κυκλοφόρησε η έκδοση 3.0 (γνωστή και ως py3k ή python 3000). Πολλά από τα καινούργια χαρακτηριστικά αυτής της έκδοσης έχουν μεταφερθεί στις εκδόσεις 2.6 και 2.7 που είναι προς τα πίσω συμβατές. Μέχρι σήμερα χρησιμοποιείται πολύ ενεργά και η Python 2 και η Python 3 (μάλιστα στην παρούσα εργασία χρησιμοποιείται Python 2) και υποστηρίζονται και οι δύο από την κοινότητα.

Συν τοις άλλοις, διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες και για την ταχύτητα εκμάθησης της. Ένα ιδιαίτερο χαρακτηριστικό της γλώσσας είναι η χρήση της στοίχισης για το διαχωρισμό των συντακτικών δομών που προγράμματος, σε αντίθεση με τη συνήθη πρακτική άλλων γλωσσών όπου για τον ίδιο σκοπό χρησιμοποιούνται ειδικά σύμβολα (πχ αγκύλες). Αυτό, σε συνδυασμό με το ότι χρησιμοποιεί πλήρεις αγγλικές λέξεις στη θέση συμβόλων, καθιστούν τον κώδικά της ευανάγνωστο από όσους έχουν βασική γνώση της αγγλικής γλώσσας.

Πάμε να δούμε τις βιβλιοθήκες και τα περιβάλλοντα που ήταν απαραίτητα κατά την ανάπτυξη του κώδικα των πειραμάτων.

2.3.1 SciPy

Η SciPy είναι μια βιβλιοθήκη ελεύθερου λογισμικού που χρησιμοποιείται στην υπολογιστική επιστήμη. Περιλαμβάνει μεθόδους για πολλές σημαντικές εργασίες της επιστήμης και της μηχανικής, όπως βελτιστοποίηση, γραμμική άλγεβρα, επεξεργασίας σημάτων και εικόνας, επίλυση διαφορικών εξισώσεων και πολλά άλλα.

Ουσιαστικά ενθυλακώνει τις βιβλιοθήκες NumPy, Pandas, scikit-learn, matplotlib (για τις οποίες θα πούμε στη συνέχεια) και άλλες σε ένα σύνολο επιστημονικού σκοπού βιβλιοθηκών που καλούνται SciPy stack (ή NumPy stack).

2.3.2 Pandas

Τα Pandas είναι μια βιβλιοθήκη ελεύθερου λογισμικού που πρωτογράφτηκε σε C και Python από τον Wes McKinney το 2008 και χρησιμοποιείται ευρέως για διαχείριση και

ανάλυση δεδομένων. Το όνομα προέρχεται από τη σύντμηση του όρου panel data, τα οποία αποτελούν πολυδιάστατες δομές δεδομένων στην Οικονομετρία.

Βασικό αντικείμενο της βιβλιοθήκης είναι το Data Frame, το οποίο θυμίζει πίνακες άλλων γνωστών λογισμικών (πχ Microsoft Excel). Η βιβλιοθήκη διευκολύνει την ανάκτηση δεδομένων από άλλες πηγές, όπως αρχεία Excel ή βάσεις δεδομένων και τα αποθηκεύει σε data frames. Τα data frames διευκολύνουν πάρα πολύ δραστηριότητες σχετικές με τα δεδομένα αυτά, όπως το "πέρασμα" από όλες τις γραμμές και είναι ιδιαίτερα αποδοτικά σε μεγάλη κλίμακα. Αυτό καθιστά τη βιβλιοθήκη ιδιαιτέρως σημαντική και απαραίτητη σε εφαρμογές Big Data.

2.3.3 Scikit-learn

Η Scikit-learn είναι μια βιβλιοθήκη ελεύθερου λογισμικού που ξεκίνησε ως Google Summer of Code project το 2007 γραμμένη σε Python και δημοσιεύτηκε πρώτη φορά το 2010. Πρόκειται για μια βιβλιοθήκη εργαλείο Μηχανικής Μάθησης. Περιέχει τους κορυφαίους αλγόριθμους για όλα τα είδη μάθησης (Κατηγοριοποίηση, Παλινδρόμηση, Συσταδοποίηση) και επικοινωνεί με τις SciPy και NumPy.

2.3.4 Numpy

Η NumPy (Numerical Python) είναι μια βιβλιοθήκη ελεύθερου λογισμικού που δημιουργήθηκε το 2005 από τον Travis Oliphant, ο οποίος ένωσε τις λειτουργικότητες δύο ανταγωνιστικών μέχρι τότε μοντέλων. Παρέχει υποστήριξη για μεγάλους, πολυδιάστατους πίνακες και περιλαμβάνει μια σειρά από μαθηματικές συναρτήσεις υψηλού επιπέδου, σχεδιασμένες να αποδίδουν πάνω στους πίνακες.

2.3.5 Matplotlib

Η Matplotlib είναι μια βιβλιοθήκη ελεύθερου λογισμικού που δημιουργήθηκε από τον John Hunter το 2003. Πρόκειται για μια βιβλιοθήκη που παράγει υψηλής ποιότητας γραφήματα, όπως ιστογράμματα ή διαγράμματα "πίτες". Σκοπός είναι να παράγονται διαγράμματα με λίγες γραμμές κώδικα και μεταξύ άλλων προσφέρει στο χρήστη περιβάλλον που θυμίζει το λογισμικό MATLAB.

2.3.6 Math

Η βιβλιοθήκη math ανήκει στις βασικές βιβλιοθήκες της Python και προσφέρει παντός είδους μαθηματικές συναρτήσεις, όπως αυτές ορίζονται από τα standards της γλώσσας C. Οι συναρτήσεις αυτές λειτουργούν μόνο για πραγματικούς αριθμούς.

2.3.7 CSV & RE

Αυτές οι δύο βιβλιοθήκες είναι βιβλιοθήκες ειδικού σκοπού, τον οποίο μαρτυρούν και στο όνομά τους. Η csv παρέχει κλάσεις για την ανάγνωση και την αποθήκευση αντίστοιχα από ή σε αρχεία με την κατάληξη csv. Η re επιτρέπει λειτουργίες ταιριάσματος κανονικών εκφράσεων (Regular Expressions).

3. Χρησιμοποιούμενοι αλγόριθμοι εκμάθησης

Στη συνέχεια θα προσπαθήσουμε να αναλύσουμε τους αλγορίθμους που χρησιμοποιήθηκαν κατά την επεξεργασία των δεδομένων, κάνοντας τις απαραίτητες αναφορές στην επιστήμη των μαθηματικών και της βιολογίας! Οι αλγόριθμοι αυτοί είναι κατά σειρά η μέθοδος των ελαχίστων τετραγώνων, ο αλγόριθμος απότομης καθόδου και οι γενετικές τεχνικές της μετάλλαξης και της διασταύρωσης.

3.1 Μέθοδος των ελαχίστων τετραγώνων

Η πρώτη αναφορά με ολοκληρωμένη ανάπτυξη της μεθόδου εμφανίζεται το 1805 σε μια εργασία του Γάλλου μαθηματικού Legendre (1752-1833) και αμέσως μετά από το Γερμανό μαθηματικό Gauss (1777-1855) στην αστρονομική του πραγματεία "Theoria Motus" για τον προσδιορισμό της τροχιάς του μικρού πλανήτη Δήμητρα.

Πρόκειται για τη μέθοδο που χρησιμοποιείται για την εύρεση του Απλού Γραμμικού Μοντέλου. Όπως είναι ευρέως γνωστό, η εξίσωση της ευθείας δίνεται από τον τύπο $y = ax + b$ με α,b τις παραμέτρους που θέλουμε να εκτιμήσουμε ώστε να τραβήξουμε την ευθεία που περιγράφει καλύτερα τη συσχέτιση μεταξύ των μεταβλητών x και y. Η μέθοδος αυτή συνίσταται στον υπολογισμό των παραμέτρων α,b ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των αποστάσεων των σημείων (x,y) από την ευθεία $y = a + bx$, δηλαδή να γίνεται ελάχιστη η ποσότητα:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - bx_i - a)^2$$

Οι τιμές των παραμέτρων που ελαχιστοποιούν το παραπάνω άθροισμα καλούνται εκτιμήτριες ελάχιστων τετραγώνων και συμβολίζονται με $\hat{a}$, $\hat{b}$. Βρίσκοντας την κλίση του αθροίσματος και μηδενίζοντας για να πάρουμε το ελάχιστο αποδεικνύεται ότι οι παραπάνω παράμετροι υπολογίζονται από τους εξής τύπους:

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

όπου $\bar{x}$, $\bar{y}$ οι μέσοι όροι των μεταβλητών x και y αντίστοιχα.

Η ευθεία $\hat{y} = \hat{a} + \hat{b}x$ αποκαλείται ευθεία ελαχίστων τετραγώνων ή ευθεία παλινδρόμησης και αποτελεί την τελική εκτίμηση για την ευθεία που προσαρμόζεται καλύτερα στα δεδομένα.

3.2 Αλγόριθμος απότομης καθόδου

Η μέθοδος της προηγούμενης παραγράφου μπορεί να γενικευτεί εύκολα για περισσότερες παραμέτρους, ωστόσο υπάρχει ένας καλύτερος αλγόριθμος ο οποίος εφαρμόζεται στα προβλήματα μηχανικής μάθησης.

Έχουμε τη συνάρτηση υπόθεσης (που είναι συνάρτηση γραμμικής παλινδρόμησης) $h_\theta(\boldsymbol{X}) = \theta_0 + \boldsymbol{\theta X}$ με $\boldsymbol{\theta}$, $\boldsymbol{X}$ τα διανύσματα βαρών και ανεξάρτητων μεταβλητών αντίστοιχα

και θέλουμε να ελαχιστοποιήσουμε τη συνάρτηση κόστους ελάχιστων τετραγώνων $J(\boldsymbol{\theta}) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - \mathrm{y_i})^2$ με m τον αριθμό των δειγμάτων. Ο αλγόριθμος ξεκινάει με μια τυχαία αρχικοποίηση του διανύσματος $\boldsymbol{\theta}$ και αλλάζει τιμές προσπαθώντας να συγκλίνει στις τιμές που ελαχιστοποιούν τη συνάρτηση κόστους. Σε κάθε βήμα τα βάρη αλλάζουν με ταυτόχρονη ενημέρωση ως εξής:

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\boldsymbol{\theta})$$

3.3 Γενετικοί αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι είναι στοχαστικοί αλγόριθμοι και χρησιμοποιούνται σε προβλήματα βελτιστοποίησης όπου επαρκεί μια προσεγγιστική λύση και είναι αρκετά απλοί στην υλοποίησή τους. Οι τιμές για τις παραμέτρους του συστήματος πρέπει να κωδικοποιούνται με τρόπο ώστε να αναπαρασταθούν από μια μεταβλητή που περιέχει σειρά χαρακτήρων ή δυαδικών ψηφίων (0/1). Αυτή η μεταβλητή μιμείται το γενετικό κώδικα (γονιδίωμα) που υπάρχει στους ζωντανούς οργανισμούς.
Αρχικά, ο Γενετικός Αλγόριθμος παράγει πολλαπλά αντίγραφα της μεταβλητής/γενετικού κώδικα, συνήθως με τυχαίες τιμές, δημιουργώντας ένα πληθυσμό λύσεων. Κάθε λύση (τιμές για τις παραμέτρους του συστήματος) δοκιμάζεται για το πόσο κοντά φέρνει την αντίδραση του συστήματος στην επιθυμητή, μέσω μιας συνάρτησης που δίνει το μέτρο ικανότητας της λύσης και η οποία ονομάζεται συνάρτηση ικανότητας (Σ.Ι).
Οι λύσεις που βρίσκονται πιο κοντά στην επιθυμητή, σε σχέση με τις άλλες, σύμφωνα με το μέτρο που μας δίνει η Σ.Ι, αναπαράγονται στην επόμενη γενιά λύσεων και λαμβάνουν μια τυχαία μετάλλαξη. Επαναλαμβάνοντας αυτή τη διαδικασία για αρκετές γενιές, οι τυχαίες μεταλλάξεις σε συνδυασμό με την επιβίωση και αναπαραγωγή των γονιδιωμάτων/λύσεων που πλησιάζουν καλύτερα το επιθυμητό αποτέλεσμα θα παράγουν ένα γονίδιο/λύση που θα περιέχει τις τιμές για τις παραμέτρους που ικανοποιούν όσο καλύτερα γίνεται την Σ.Ι.

3.3.1 Ορολογία

Ας προσπαθήσουμε να εξηγήσουμε κάποιους όρους που θα χρησιμοποιήσουμε στη συνέχεια και να τους ερμηνεύσουμε με μαθηματικά ανάλογα. Ακολουθείται αλφαβητική σειρά κι όχι σειρά εμφάνισης.

- Άτομο: Ένα χρωμόσωμα που μπορεί να αποτελεί μία λύση του προβλήματος. Μαθηματικά, είναι το αποτέλεσμα μιας συνάρτησης.
- Γενιά: Ένα επίπεδο εξέλιξης προς την επιθυμητή "υγεία".
- Γονίδιο: Ένας φορέας γενετικής πληροφορίας σχετικά με ένα συγκεκριμένο χαρακτηριστικό, με άλλα λόγια μία μονάδα κληρονομικότητας. Μαθηματικά, είναι μια παράμετρος συνάρτησης που επηρεάζει μια συγκεκριμένη μεταβλητή.
- Γονιδίωμα: Ο συνδυασμός γενετικής πληροφορίας για ένα άτομο. Ανάλογα, θα ήταν μια μαθηματική συνάρτηση ως όλον.
- Διασταύρωση: Η δημιουργία απογόνου με το συνδυασμό δύο γονέων. Αποτελεί έναν από τους τελεστές γενετικής εξέλιξης του πληθυσμού. Μαθηματικά, θα μπορούσε να αντιστοιχιστεί με μια μαθηματική συνάρτηση που προκύπτει από το συνδυασμό παραμέτρων άλλων συναρτήσεων.
- Μετάλλαξη: Αλλαγή στην ακολουθία ενός γονιδίου. Αποτελεί έναν από τους τελεστές γενετικής εξέλιξης του πληθυσμού. Μαθηματικά, θα μπορούσε να αντιστοιχιστεί με μια αλλαγή παραμέτρου που οδηγεί σε αλλαγή αποτελέσματος.

- Πληθυσμός: Ένα σύνολο ατόμων σε μια έρευνα. Μαθηματικά, είναι ένα σύνολο λύσεων από το οποίο διαλέγουμε την καλύτερη.
- Τόπος: Αγγλιστί locus. Συγκεκριμένη τοποθεσία σε ένα χρωμόσωμα, για παράδειγμα η τοποθεσία ενός γονιδίου.
- Υγεία (ή Καταλληλότητα): Αγγλιστί το fitness. Πρόκειται για ένα μέτρο της προσαρμογής ενός ατόμου. Μαθηματικά, θα ήταν η καταλληλότητα μιας λύσης σε ένα πρόβλημα βελτιστοποίησης.
- Χρωμόσωμα: Είναι μια συλλογή γονιδίων που μεταφέρει γενετική πληροφορία. Θα μπορούσε να είναι ένα τμήμα μαθηματικής συνάρτησης με διάφορες παραμέτρους.

3.3.2 Διασταύρωση

Η διασταύρωση, αποκαλούμενη και ανασυνδυασμός, είναι ένας γενετικός τελεστής που χρησιμοποιείται ώστε να συνδυάσει τη γενετική πληροφορία δύο γονέων για να παραχθούν απόγονοι. Υπάρχουν διάφοροι τρόποι να προκύψει ο νέος απόγονος από τους γονείς, τους οποίους και αναφέρουμε στη συνέχεια.

3.3.2.1 Διασταύρωση ενός σημείου

Σε αυτό τον αλγόριθμο, επιλέγεται ένας (κοινός) τόπος πάνω στα χρωμοσώματα των γονέων και προκύπτουν δύο απόγονοι: ένας με τη γενετική πληροφορία αριστερά από τον πρώτο γονέα και δεξιά από τον δεύτερο και το αντίστροφο (βλ. Εικόνα 3.1).

3.3.2.2 Διασταύρωση k σημείων

Πρόκειται για τη γενίκευση του προηγούμενου αλγορίθμου, όπου επιλέγονται οσαδήποτε σημεία και χρησιμοποιείται εναλλάξ η γενετική πληροφορία από τους γονείς. Στην Εικόνα 3.1 βλέπετε παράδειγμα για 2 σημεία.

3.3.2.3 Ομοιόμορφη διασταύρωση

Τυπικά αποτελεί διασταύρωση k σημείων, ωστόσο η διαφορά έγκειται στο ότι δεν προκαθορίζεται ο αριθμός των τόπων αλλαγής. Απλά σε κάθε τόπο επιλέγεται τυχαία, δηλαδή με ίση πιθανότητα, ο γονιός του οποίου η γενετική πληροφορία θα χρησιμοποιηθεί. Μια παραλλαγή του συγκεκριμένου αλγορίθμου χρησιμοποιείται στην υλοποίηση της παρούσας εργασίας.

3.3.2.4 Αριθμητική διασταύρωση

Σε αυτό τον αλγόριθμο εκτελείται κάποια αριθμητική (λογική) πράξη μεταξύ των bits των γονέων για να παραχθεί ο απόγονος. Συνηθέστερες πράξεις είναι το λογικό XOR και XNOR, λόγω της ομοιομορφίας των εξόδων τους.

Εικόνα 3.1: Παράδειγμα διασταύρωσης ενός σημείου (αριστερά) και δύο σημείων (δεξιά)

3.3.3 Μετάλλαξη

Η μετάλλαξη είναι ένας ακόμα γενετικός τελεστής. Χρησιμοποιείται για τη διατήρηση και ενίσχυση της γενετικής ποικιλομορφίας ενός πληθυσμού από τη μία γενιά στην άλλη. Η διαδικασία αυτή είναι χρήσιμη και συχνά απαραίτητη ώστε να αποφευχθούν μεγάλες ομοιότητες μέσα στον πληθυσμό και κατά συνέπεια το "σκάλωμα" σε τοπικά ελάχιστα της συνάρτησης σφάλματος.

3.3.3.1 Μετάλλαξη ενός σημείου

Στο συγκεκριμένο αλγόριθμο παράγεται μία τυχαία μεταβλητή για κάθε bit γενετικής πληροφορίας και αυτή η μεταβλητή καθορίζει αν θα αντιστραφεί το bit στο οποίο αναφέρεται.

3.3.3.2 Αντιστροφή

Σε αυτό τον αλγόριθμο αντιστρέφονται όλα τα bit του επιλεγμένου γονιδιώματος.

3.3.3.3 Ομοιόμορφη μετάλλαξη

Πρόκειται για μετάλλαξη που πραγματοποιείται σε γονιδιώματα πραγματικών αριθμών. Το επιλεγμένο γονίδιο αντικαθίσταται με μια τυχαία (ομοιόμορφα επιλεγμένη) τιμή εντός των ορίων που έχουν τεθεί.

3.3.3.4 Μη ομοιόμορφη μετάλλαξη

Η διαφορά από το προηγούμενο είναι ότι όσο περνούν οι γενιές μειώνεται η πιθανότητα μετάλλαξης. Έτσι διευκολύνεται το σύστημα στην αρχή, αλλά διατηρείται ο πληθυσμός και οι πιθανές λύσεις προς το τέλος.

3.3.3.5 Γκαουσιανή μετάλλαξη

Εδώ δεν έχουμε ολοκληρωτική μετάλλαξη του γονιδίου, αλλά πρόσθεση σε αυτόν γκαουσιανού θορύβου, μέχρι τα όρια που έχουν τεθεί φυσικά, αλλιώς αποκόπτεται.

3.4 Αξιολόγηση αλγορίθμων

Στα προβλήματα ταξινόμησης υπάρχει μια πολύ προφανής μετρική για την αξιολόγηση των αλγορίθμων μηχανικής μάθησης. Απλούστατα μετράς επιτυχίες-αποτυχίες και μπορείς να αξιολογήσεις το μοντέλο με βάση την αποδοτικότητα του αλγορίθμου σε αυτό το απλό κριτήριο. Φυσικά δεν είναι μόνο αυτή η μετρική αξιολόγησης και υπάρχουν πολλές και πιο περίπλοκες, ωστόσο φαίνεται να υπάρχουν επιλογές που προκρίνονται διαισθητικά έναντι των άλλων για να παρουσιάσεις με απλότητα και σαφήνεια την απόδοση ενός μοντέλου για τέτοιο πρόβλημα.

Στην παλινδρόμηση δυστυχώς δεν υπάρχει κάποια τέτοια προφανής ή διαισθητική λύση. Υπάρχουν όμως και εδώ αρκετές ικανοποιητικές φόρμουλες. Στις υποενότητες θα αναφερθούμε στις δημοφιλέστερες, ανάμεσα στις οποίες θα επιλέξουμε για να απεικονίσουμε τα αποτελέσματα των πειραμάτων.

3.4.1 Μέσο απόλυτο σφάλμα

Το μέσο απόλυτο σφάλμα (mean absolute error or MAE) είναι το άθροισμα των απόλυτων διαφορών ανάμεσα στις εκτιμήσεις και τις πραγματικές τιμές (καλούνται υπόλοιπα ή residuals) διαιρούμενο με το πλήθος των παρατηρήσεων. Πρόκειται για την απλούστερη και διαισθητικά πιο "εύπεπτη" μετρική για συνεχείς τιμές.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}$$

Στα πειράματα χρησιμοποιούμε μια παραλλαγή της, όπου το σφάλμα είναι σχετικό, δηλαδή διαιρείται με την προσδοκώμενη τιμή και μετατρέπεται σε ποσοστιαίο. Η μετρική αυτή ονομάζεται μέσο ποσοστιαίο σφάλμα (mean percentage error).

3.4.2 Μέσο τετραγωνικό σφάλμα

Γνωστό στα αγγλικά ως mean squared error (MSE) ή mean squared deviation (MSD). Η μοναδική υπολογιστική διαφορά με το αποπάνω είναι ότι παίρνουμε το άθροισμα των τετραγώνων των υπολοίπων. Για πληρότητα έχουμε:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^{n} {e_i}^2}{n}$$

Η μετρική αυτή αποτελεί ουσιαστικά (και τυπικά) τη διασπορά της εκτιμήτριας, εφόσον είναι αμερόληπτη. Το ελάττωμά της είναι ότι δεν έχει τις ίδιες μονάδες με τα αποτελέσματα. Για αυτό πολλοί ερευνητές παίρνουν την τετραγωνική ρίζα της συγκεκριμένης ποσότητας, γνωστή ως RMSE εκ του αγγλικού root. Εντελώς ανάλογα για αμερόληπτες εκτιμήτριες η ποσότητα αυτή αποτελεί την τυπική απόκλιση.

3.4.3 Συντελεστής προσδιορισμού R^2

Ο συντελεστής προσδιορισμού αποτελεί μια μετρική της καταλληλότητας ή του ταιριάσματος του συνόλου των εκτιμήσεων σε αυτό των πραγματικών τιμών. Θεωρητικά παίρνει τιμές από 0 έως 1, όπου 1 το απόλυτα ταίριασμα, ωστόσο πρακτικά στην υλοποίηση της Python είναι δυνατό να πάρει αρνητικές τιμές για πολύ κακά μοντέλα. Ο γενικός τύπος υπολογισμού είναι ο εξής:

$$R^2 = 1 - \frac{SSE}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \mu_y)^2}{\sum_{i=1}^{n} {e_i}^2}$$

4. Συλλογή πειραματικών δεδομένων

Σε αυτό το σημείο παρουσιάζουμε βήμα βήμα την προετοιμασία των δεδομένων που χρησιμοποιήθηκε στην εργασία, από την ανάκτηση των δεδομένων από το διαδίκτυο μέχρι τη δημιουργία του dataset που χρησιμοποιούμε σχεδόν στο σύνολο των πειραμάτων.

4.1 Twitter API

Το twitter, όπως και τα περισσότερα social media πλέον, προσφέρει μια διεπαφή μέσα από την οποία κάθε ενδιαφερόμενος μπορεί να προγραμματίσει μια εφαρμογή και να αξιοποιήσει μέρος των υπηρεσιών τις πλατφόρμας και δεδομένα που είναι δημόσια και διαθέσιμα για ανάκτηση και χρήση. Οι κυριότερες υπηρεσίες στις οποίες έχει πρόσβαση ένας developer μέσω του API είναι οι παρακάτω:

- Διαχείριση λογαριασμών: Οι προγραμματιστές έχουν τη δυνατότητα να αλλάξουν στοιχεία και ρυθμίσεις σε λογαριασμούς, να μπλοκάρουν άλλους χρήστες, να αιτηθούν για πληροφορίες δραστηριότητας λογαριασμών και διάφορα άλλα. Η συγκεκριμένη δυνατότητα είναι ιδιαίτερα χρήσιμη σε υπηρεσίες, όπως τμήματα διαχείρισης κινδύνων και επειγόντων περιστατικών.
- Διαχείριση tweets και replies: Οι χρήστες δύνανται να προγραμματίσουν δημοσιεύσεις για τους λογαριασμούς τους και επίσης όλα τα δημόσια tweets και απαντήσεις σε αυτά είναι διαθέσιμα στο API. Το δεύτερο είναι ιδιαίτερα σημαντικό για πάρα πολλούς λόγους. Οι προγραμματιστές μπορούν να αναζητήσουν λέξεις κλειδιά ή να πάρουν δείγματα δημοσιεύσεων με συγκεκριμένες παραμέτρους πχ χωροχρονικές. Αυτή η δυνατότητα είναι πολύ σημαντική για χρήση στον τομέα της υγείας, καθώς διάφοροι οργανισμοί μπορούν να ανιχνεύσουν ξεσπάσματα ασθενειών σε συγκεκριμένες περιοχές ή ακόμα και να διακρίνουν και να αποτρέψουν την παραπληροφόρηση. Φυσικά μπορεί να αξιοποιηθεί και από οποιαδήποτε ομάδα επιστημόνων για ερευνητικούς σκοπούς, όπως και στην περίπτωση μας.
- Στο παραπάνω πλαίσιο το Twitter παρέχει και άμεσα μηνύματα (distant messages) μόνο εφόσον έχει επιτραπεί ρητά από το χρήστη που αφορούν. Αυτή η παροχή διευκολύνει τους προγραμματιστές chatbots.
- Η πλατφόρμα παρέχει επίσης μια σουίτα διεπαφών, ώστε να διευκολύνει τις επιχειρήσεις να ανιχνεύσουν ενδιαφέροντα και καυτά θέματα και με βάση την ανάλυση που πραγματοποιούν να δημιουργήσουν μια διαφημιστική καμπάνια που θα μπορέσει να καλύψει όσο δυνατόν περισσότερο το ποικιλόμορφο κοινό του Twitter.
- Μια τελευταία δυνατότητα που δίνει το Twitter API είναι αυτή της ενσωμάτωσης. Έτσι διευκολύνεται σημαντικά η τοποθέτηση σε ιστοσελίδες υλικού από το Twitter, όπως δημοσιεύσεις και συζητήσεις, πλήκτρα, χρονοδιαγράμματα και γενικότερα περιεχόμενο της πλατφόρμας.

Η πλατφόρμα τονίζει ότι δίνει ιδιαίτερη βαρύτητα στην προστασία του προϊόντος της και την ασφάλεια των χρηστών που απολαμβάνουν τις υπηρεσίες της. Σε αυτά τα πλαίσια εφαρμόζει αυστηρούς περιορισμούς στη χρήση των δεδομένων που παρέχονται από το API και αποκλείει την πρόσβαση σε τρίτους που φαίνεται να κάνουν χρήση των υπηρεσιών αυτών.

Φυσικά υπάρχει περιορισμός και στον αριθμό των αιτήσεων που μπορείς να καταθέσεις στη διεπαφή. Χοντρικά, το συγκεκριμένο όριο είναι 15 κλήσεις ανά 15 λεπτά, το οποίο είναι το χρονικό παράθυρο μηδενισμού των περιορισμών. Αυτό συνεπάγεται 180 αιτήσεις ανά χρήστη και 450 αιτήσεις ανά εφαρμογή και αναγνωριστικό πρόσβασης (access token) το οποίο συνοδεύει κάθε εφαρμογή.

4.2 GetOldTweets script

Στον τίτλο αναφέρεται ένα script γραμμένο σε python το οποίο χρησιμοποιούμε για να τραβήξουμε δεδομένα από το Twitter API και να παρακάμψουμε τους περιορισμούς που περιγράφονται στην προηγούμενη ενότητα. Το πρόγραμμα έχει δημιουργηθεί και συντηρείται στο GitHub από τον Jefferson Henrique και μερικούς ακόμα χρήστες να συνεισφέρουν.

Ένας πρόσθετος περιορισμός στο Twitter API είναι ότι δεν μπορείς να ανακτήσεις δημοσιεύσεις παλιότερες της εβδομάδας. Ο δημιουργός του script παρατήρησε ότι ένας τρόπος να αποκτήσεις πρόσβαση σε αυτά τα tweets μέσω της ιστοσελίδας αρκεί να κάνεις scroll down για να εμφανιστούν περισσότερες δημοσιεύσεις. Γράφοντας το πρόγραμμα φροντίζει ο αλγόριθμος να μιμηθεί τη συγκεκριμένη συμπεριφορά και έτσι κάμπτει αυτόν τον περιορισμό της διεπαφής.

Αφήνοντας κατά μέρος τη διαδικασία ανάκτησης των αναγνωριστικών και των απαραίτητων αδειών για την εξόρυξη δεδομένων από τη διεπαφή, το script έχει πολύ απλή λειτουργία. Δημιουργείται μια κλάση tweet που περιλαμβάνει όλη τη χρήσιμη πληροφορία. Συγκεκριμένα περιέχει τα εξής πεδία:

- Αναγνωριστικό δημοσίευσης (id)
- Μόνιμος σύνδεσμος (permalink)
- Όνομα χρήστη (username)
- Κείμενο (text)
- Ημερομηνία και Ώρα (date)
- Αναδημοσιεύσεις (retweets)
- Επισήμανση "Μου αρέσει" (favorites)
- Αναφορές σε άλλους χρήστες (mentions)
- Ετικέτες θέματος (hashtags)
- Γεωγραφικό στίγμα (geo)

Παράλληλα επιτρέπει την προσωποποιημένη αναζήτηση παρέχοντας μια σειρά μεταβλητών που μπορούν να ρυθμιστούν από το χρήστη, μεταξύ άλλων η αναζήτηση για δημοσιεύσεις που έγιναν από συγκεκριμένο λογαριασμό, σε συγκεκριμένη χρονική περίοδο, περιέχουν συγκεκριμένο κείμενο ή τμήμα κειμένου και ακόμα περιλαμβάνουν συγκεκριμένες αναφορές και ετικέτες. Δίνεται επίσης η δυνατότητα μέγιστου αριθμού αποτελεσμάτων προς επιστροφή. Όπως θα δούμε και στη συνέχεια, εμείς εφαρμόζουμε αναζήτηση με συγκεκριμένα hashtags και σε συγκεκριμένες χρονικές περιόδους φυσικά.

Ένα ελάττωμα του προγράμματος που δεν κατέστη δυνατό να διορθωθεί είναι ότι αναγνωρίζει περιορισμένο πλήθος χαρακτήρων και όχι το διευρυμένο λατινικό αλφάβητο, στο οποίο περιλαμβάνονται συγκεκριμένοι ιταλικοί χαρακτήρες με τονισμούς που δεν υπάρχουν στην αγγλική γλώσσα. Έτσι υπήρχαν μεμονωμένες περιπτώσεις που το πρόγραμμα συναντούσε τέτοιους χαρακτήρες στο Κείμενο και διέκοπτε τη λειτουργία του. Ταυτόχρονα, υπήρχαν κάποιες λίγες φορές που το Twitter API δεν ήταν αρκούντως συνεργάσιμο και διέκοπτε επίσης τη λειτουργία του script.

Ευτυχώς, στην ευρύτατη πλειοψηφία των παραπάνω περιπτώσεων μια επανάληψη της λειτουργίας ήταν αρκετή για να εξαλειφθεί το πρόβλημα και μόνο τα δεδομένα για μία ημέρα από το παράθυρο των 2 ετών υπέστησαν μια μικρή αλλοίωση. Το ζήτημα ξεπεράστηκε τρέχοντας έναν αλγόριθμο Expectation-Maximization (EM) που προσομοίωσε την κίνηση των δημοσιεύσεων για το επίμαχο δίωρο βάσει της κατανομής των δημοσιεύσεων στις υπόλοιπες 22 ώρες, για τις οποίες είχαμε πλήρη δεδομένα.

4.3 Δεδομένα αναφοράς

Η παρούσα εργασία μελετά την ανίχνευση του ενδιαφέροντος χρηστών των κοινωνικών δικτύων για τηλεοπτικές εκπομπές με βάση τη δραστηριότητά τους στα μέσα κοινωνικής δικτύωσης. Συγκεκριμένα αντλούμε δεδομένα από τις πλατφόρμες Twitter και Google Trends, όπως έχει αναφερθεί και νωρίτερα.

Η εκπομπή στην οποία βασίστηκε σχεδόν το σύνολο των πειραμάτων είναι το ιταλικό talent show με τίτλο *Amici di Maria de Filippi*. Πρόκειται για μια μουσική ακαδημία που φιλοξενεί 20 νεαρά άτομα μεταξύ 18 και 25 ετών που φιλοδοξούν να ξεκινήσουν μια καριέρα στο χώρο ως τραγουδιστές, στιχουργοί ή χορευτές. Το σόου ξεκίνησε το 2001 και έκτοτε προβάλλεται αδιαλείπτως έχοντας συμπληρώσει 17 κύκλους εκπομπών με περίπου 10 επεισόδια ανά κύκλο. Εμείς μελετάμε τις σεζόν 15 και 16 που προβλήθηκαν το 2016 και 2017 με 10 και 9 εβδομαδιαία επεισόδια αντίστοιχα. Η συγκεκριμένη εκπομπή απολαμβάνει μεγάλα νούμερα τηλεθέασης και γενικότερα υψηλή δραστηριότητα στα κοινωνικά δίκτυα, γεγονός που την καθιστά εξαιρετικό αντικείμενο μελέτης.

Ένα show που μελετήθηκε επίσης σε κάποια πειράματα είναι η σατιρική εκπομπή *Le Iene (Οι ύαινες)*. Προβάλλεται από το 1997 και φιλοξενεί κάποιους δημοσιογράφους που σατιρίζουν την πολιτική και μη επικαιρότητα με σκετσάκια και "ειδικά" ρεπορτάζ. Έχει σαφώς πιο περιορισμένο κοινό και δε χρησιμοποιήθηκε σε όλα τα πειράματα, αφού υπήρχαν περιπτώσεις που το δείγμα ήταν αρκετά μικρό για να είναι αξιόπιστο και αποδοτικό.

4.4 Το dataset

Στο δεδομένο σημείο ξεκινάει η ενασχόληση με τα πιο πρακτικά ζητήματα της εργασίας. Σε αυτή την ενότητα θα παρουσιαστούν αναλυτικά τα βήματα που έγιναν για την προετοιμασία των δεδομένων για τα πειράματα.

4.4.1 Ανάκτηση και επεξεργασία των tweets

Η ανάκτηση των tweets έγινε απλά τρέχοντας το script που περιγράφεται στην ενότητα 4.2, με τα αποτελέσματα να αποθηκεύονται σε αρχείο csv. Από όλα τα στοιχεία που τραβάμε από το Twitter API πρέπει αφ' ενός να επιλέξουμε αυτά που μας είναι χρήσιμα κι αφ' ετέρου να δημιουργήσουμε καινούρια πληροφορία που θα αποτελέσει είσοδο στα πειράματα.

Οι παράμετροι που χρησιμοποιήσαμε είναι το χρονικό παράθυρο και η αναζήτηση δημοσιεύσεων με συγκεκριμένο hashtag. Η χρονική διάρκεια της αναζήτησης ήταν οι 2 χρονιές 2016 και 2017 σε παράθυρα του ενός μήνα και 2 ημερών (εξηγείται παρακάτω το γιατί). Αναζητήσαμε δημιοσεύσεις οι οποίες περιέχουν τα hashtags #amici15 και #amici16, που αποτελούν τα official hashtags για τους δύο κύκλους που προβλήθηκαν τις αντίστοιχες χρονιές.

Οι χρησιμοποιούμενες μετρικές αποφασίστηκε να είναι το πλήθος των tweets που πραγματοποιήθηκαν ανά ημέρα και ο αριθμός των μοναδικών χρηστών που δημοσίευσαν tweet καθ' εκάστη μέρα. Ο ακόλουθος κώδικας δημιούργησε τα δεδομένα αυτά, διαβάζοντας ως είσοδο το πλήρες path των αρχείων csv και μερικές αριθμητικές παραμέτρους που θα εξηγηθούν αμέσως.

```
import pandas as pd
import csv

def fix_csv(d, df, fon):
    t = [int(i[-2:]) for i in df.index] #get days from date
    cnt = 0
    fout = open(fon,'w')
    for idx, line in zip(t,df):
        cnt += 1
        fixes = idx - cnt #days missing
        for i in range(fixes):
            fout.write('0\n')
            cnt += 1
        fout.write(str(line)+'\n')
    for i in range(cnt,d): #in case there are missing days at the
end of the month
        fout.write('0\n')
    fout.close()

df = pd.read_csv(fname, usecols = [0, 1, 6, 7], sep=';')

#extract date from date&time field
df['date2'] = df['date'].str.split(' ', 1).str[0]

#only one of this two commands should run each time, if none of
them is commented out only the second one will be executed
df2 = df.groupby(['date2']).size() #get number of tweets per day
df2 = df.groupby(['date2']).username.nunique() #get number of
unique users tweeting per day

#dealing with timezone limitation
if df2.shape[0] == (days+2L):
    df2 = df2[1:-1] #cut the extra 2 days
    flag = False
if flag:
    if int(df2.index[0][-2:]) == prev: #get day of date string
        df2 = df2[1:] #if first extra day has tweets, cut it
    if int(df2.index[-1][-2:]) == 1:
        df2 = df2[:-1] #if second extra day has tweets, cut it
if df2.shape[0] != days: #check if there are missing days/values
    fix_csv(days, df2, fout_name)
else:
    df2.to_csv(fout_name, index=False)
```

Πίνακας 4.1: Κώδικας για τη μέτρηση του πλήθους tweets και μοναδικών χρηστών ανά ημέρα

Ένας ακόμα περιορισμός που μας επέβαλε το Twitter API είναι ότι χρησιμοποιεί μια ενιαία ζώνη ώρας (αντί να προσαρμόζεται στο χρήστη) κι έτσι τα χρονικά παράθυρα δεν ξεκινούσαν και κατέληγαν στα μεσάνυχτα των δοθεισών ημερών, αλλά σε μια ενδιάμεση ώρα. Αυτό είχε ως αποτέλεσμα να επεκτείνουμε κατά 2 ημέρες τα παράθυρα, μία στην αρχή και μία στο τέλος, ώστε να πάρουμε ολοκληρωμένες τις ημέρες που μας ενδιαφέρουν.

Ένα ζήτημα που επιλύει η συνάρτηση fix_csv είναι ότι η μέτρηση των tweets ή των χρηστών γίνεται με ομαδοποίηση κατά την ημερομηνία. Επομένως, αν για κάποια ημέρα δεν έχει υπάρξει δημοσίευση, το πρόγραμμα δε γνωρίζει ότι έπρεπε να υπάρχει σε εκείνο το σημείο μία μέρα με μηδενική δραστηριότητα και δημιουργεί αναντιστοιχία στα δεδομένα.

Τα δύο αυτά ζητήματα συνδυαστικά μας υποχρέωσαν να προσθέσουμε τα if statements για να ελέγχουμε ότι δε χάνεται κάποια μέρα κι αν χάνεται, η fix_csv βρίσκει ποια είναι και συμπληρώνει ένα μηδενικό στη σωστή χρονολογικά θέση. Οι έλεγχοι αυτοί γίνονται με τη χειροκίνητη εισαγωγή του σωστού αριθμού ημερών για τον εκάστοτε μήνα και κάποιων χρήσιμων μεταβλητών για την προσθαφαίρεση. Η αυτοματοποίηση της αρχικοποίησης των μεταβλητών με βάση το μήνα κρίνεται ασύμφορη για τους σκοπούς του συγκεκριμένου προγράμματος.

4.4.2 Δημιουργία του αρχικού dataset

Με βάση τα προηγούμενα έχει ετοιμαστεί το βασικό σύνολο δεδομένων που θα χρησιμοποιηθεί στα πειράματα. Όπως αναλύεται και στο επόμενο κεφάλαιο ότι τα δεδομένα θα αλλάξουν αρκετές φορές μορφή, ωστόσο σε αυτήν τη φάση δημιουργείται ένα dataset αναφοράς, ως εξής (η σειρά των στηλών δεν παίζει κάποιο ρόλο):

- Κάθε γραμμή του set αντιστοιχεί σε ένα επεισόδιο της εκπομπής Amici di Maria de Filippi. Για τη διετία 2016-17 έχουμε συνολικά 19 επεισόδια, άρα και 19 γραμμές ή καλύτερα 19 στοιχεία δεδομένων.
- Η πρώτη στήλη περιλαμβάνει τη συνολική τηλεθέαση που κατέγραψε κάθε επεισόδιο επί του συνόλου των ατόμων που παρακολουθούσαν τηλεόραση εκείνο το διάστημα.
- Η δεύτερη στήλη αντίστοιχα περιέχει το συνολικό αριθμό μοναδικών ατόμων που παρακολούθησαν έστω για 1 λεπτό το συγκεκριμένο επεισόδιο.
- Η τρίτη στήλη περιέχει έναν αριθμό από το Google Trends, ο οποίος αναφέρεται στη σχετική κίνηση στο Google Search για τη συγκεκριμένη εκπομπή, όπως έχουμε περιγράψει και στην αρχή.
- Η τέταρτη στήλη περιλαμβάνει το πλήθος των tweets που αφορούν το συγκεκριμένο επεισόδιο της εκπομπής. Θα εξηγήσουμε στο επόμενο κεφάλαιο πώς υπολογίζουμε ποια tweets αφορούν ποιο επεισόδιο.
- Η πέμπτη στήλη αντίστοιχα περιέχει τον αριθμό των μοναδικών χρηστών που δημοσίευσαν τα παραπάνω tweets.

Στην παρακάτω εικόνα φαίνονται τα όσα περιγράφουμε παραπάνω, με τη σημείωση ότι τα δεδομένα του Google Trends αφορούν την εβδομάδα πριν την προβολή του επεισοδίου και τα δεδομένα από το Twitter την ημέρα προβολής. Φυσικά στα πειράματα έχουν χρησιμοποιηθεί κι άλλοι (χρονικοί) συνδυασμοί δεδομένων.

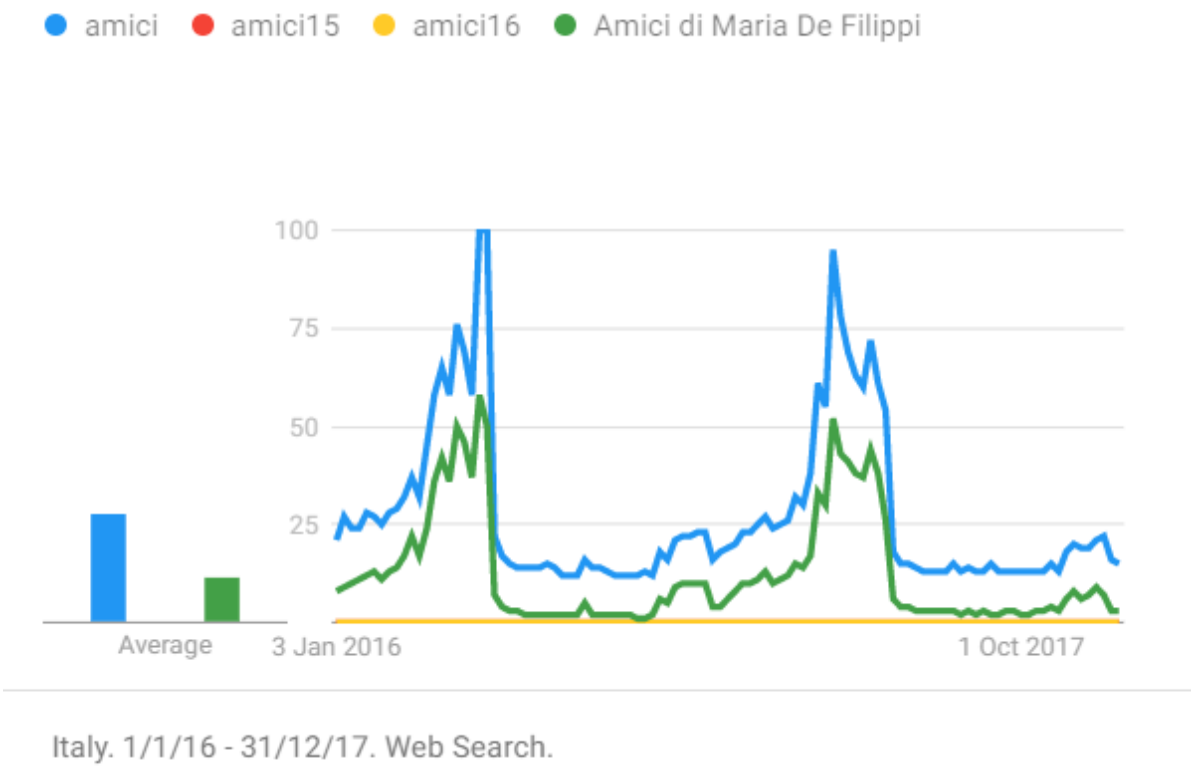
	shareAll	audAll	GTrends	tweets	unqUsers
0	22.3	4609683	46.0	26448	5983
1	21.0	4387338	59.0	21028	4471
2	21.9	4495771	64.0	17689	3942
3	21.8	4485161	57.0	20338	4518
4	23.2	4570204	76.0	17955	4190
5	23.7	4550372	68.0	16325	3625
6	22.3	4400426	57.0	10837	3076
7	24.4	4674471	100.0	18632	4905
8	29.2	5835377	96.0	37017	9195
9	20.7	4064445	38.0	24156	5237
10	20.2	4019710	61.0	15530	3665
11	18.3	3621303	54.0	16566	4074
12	19.7	3793016	95.0	22320	4796
13	21.2	4165198	76.0	16603	3971
14	22.8	4462803	68.0	16765	3684
15	21.4	4214120	64.0	16493	3853
16	19.8	3747570	60.0	11094	2750
17	24.1	4487948	71.0	13490	3214
18	28.5	4821940	62.0	34143	8019

Εικόνα 4.1: Ένα από τα όμοια datasets που χρησιμοποιήθηκαν στα πειράματα

4.4.2.1 Δεδομένα από το Google Trends

Το Google Trends είναι μια πλατφόρμα που μπορεί να δώσει ιδιαίτερα χρήσιμη διαισθητική γνώση και δεδομένα σε συγκεκριμένες περιστάσεις. Ένας από τους στόχους της εργασίας, όπως έχει προαναφερθεί, είναι να επαληθεύσουμε ότι μπορεί να δώσει πολύτιμη πληροφορία και στα audience analytics.

Μία από τις πιο σημαντικές δυνατότητες που προσφέρει η ιστοσελίδα είναι η δυνατότητα αναζήτησης σε συγκεκριμένο γεωγραφικό χώρο. Εμείς επιλέξαμε φυσικά την Ιταλία, χώρα προβολής των υπό μελέτη εκπομπών. Στην Εικόνα 4.2 γίνεται εμφανής ο τρόπος παρουσίασης των δεδομένων της πλατφόρμας σε χρονοσειρές καθώς και ο τρόπος επιλογής του κατάλληλου όρου για την αναζήτηση, μέσα από τη σύγκριση διαφόρων πιθανών όρων. Τα δεδομένα αυτά είναι εβδομαδιαίας βάσης, μιας και υπάρχουν σχετικοί περιορισμοί για την ανάκτηση ημερήσιων δεδομένων αναζήτησης. Συγκεκριμένα, το μέγιστο παράθυρο αναζήτησης που αποφέρει ημερήσια δεδομένα είναι οι 270. Έπειτα από αυτό το όριο, οι τιμές συμπτύσσονται σε μία τιμή για κάθε εβδομάδα.



Εικόνα 4.2: Παράδειγμα απεικόνισης δεδομένων από το Google Trends

Όπως έχουμε αναφέρει προηγουμένως, τα δεδομένα αυτά αποτελούν σχετικές συχνότητες εμφάνισης των δοθέντων όρων. Ειδικότερα, η Google προφανώς διαθέτει πληροφορίες για τον ακριβή όγκο αναζήτησης κάθε όρου. Αυτές οι πληροφορίες υφίστανται μια δειγματοληψία και τα ληφθέντα νούμερα συγκρίνονται με συνολικούς αριθμούς για κάποια περιοχή και σε κάποιο χρόνο ώστε να παραχθεί η σχετική συχνότητα εμφάνισης για κάθε μέρα ή εβδομάδα. Στη συνέχεια οι αριθμοί αυτοί για όλα τα διαστήματα του παραθύρου αναζήτησης επιδέχονται μια κανονικοποίηση, με το μέγιστο να λαμβάνει την τιμή 100 και τους υπόλοιπους να τίθενται ανάλογα. Το σύνολο τιμών που προκύπτει από την κανονικοποίηση είναι αυτό που προβάλλεται ως πληροφορία στο χρήστη.

Ένα σημαντικό πρόβλημα που προκύπτει από τα παραπάνω και συγκεκριμένα από το κομμάτι της δειγματοληψίας, είναι ότι υπάρχει αναντιστοιχία στα δεδομένα της πλατφόρμας. Η δειγματοληψία που περιγράψαμε γίνεται διαφορετικά για κάθε όρο, για κάθε περιοχή, για κάθε χρονικό παράθυρο. Αυτό έχει ως αποτέλεσμα να παρατηρούνται διαφορές στις τιμές που αποδίδει η ιστοσελίδα, ακόμα κι αν μιλάμε για τον ίδιο όρο, στην ίδια περιοχή και την ίδια μέρα/εβδομάδα, αρκεί να έχει αλλάξει έστω και κατά μία μέρα το παράθυρο αναζήτησης. Δε θα υπήρχε σοβαρό ζήτημα, εάν επρόκειτο απλά για διαφορές της τάξης του 1 ή 2%, ωστόσο έχει παρατηρηθεί πχ σε κυλιόμενο παράθυρο τριών ημερών, οι ίδιες μέρες να έχουν τιμές 50 και 52 αντίστοιχα και με κύλιση μιας μέρας (οπότε και δεν επηρεάζεται το παράθυρο, ούτε το μέγιστο, βάσει του οποίου γίνεται η κανονικοποίηση), να λαμβάνουν τιμές 22 και 51. Πρόκειται για διαφορές που δε δικαιολογούνται στατιστικά και εγείρουν σοβαρά ερωτήματα για την αξιοπιστία των δεδομένων που παρέχει η εταιρεία μέσω της ιστοσελίδας. Είναι απολύτως κατανοητό

Επιστρέφοντας στα της εργασίας, τα δεδομένα αυτά δεν είναι διαθέσιμα μόνο ως γραφήματα, αλλά προσφέρονται και σε αρχείο csv, καθώς και με δυνατότητα ενσωμάτωσης για ιστοσελίδες. Το csv είναι η μορφή που χρησιμοποιήθηκε για την εισαγωγή των δεδομένων στο πρόγραμμά μας. Το μόνο που χρειάστηκε ήταν ένας καθαρισμός του αρχείου, μια και στις πρώτες γραμμές υπάρχουν διάφορες πληροφορίες της αναζήτησης που πραγματοποιήθηκε όπως πχ η γεωγραφική περιοχή. Μετά από αυτά το αρχείο έχει τη δομή "ημερομηνία, αριθμητική τιμή", το οποίο είναι εξαιρετικά εύχρηστο για την εισαγωγή σε πρόγραμμα. Ένα ακόμα λεπτό σημείο ήταν η επιλογή μόνο των εβδομάδων κατά τις οποίες υπήρχε προβολή επεισοδίου. Ωστόσο αποδείχτηκε αρκετά πιο εύκολο από όσο κανείς θα ανέμενε, διότι υπήρχε μεγάλη διαφορά στις τιμές μεταξύ των εβδομάδων που προβλήθηκε επεισόδιο και των υπολοίπων παρά την κανονικοποίηση που συμπτύσσει σημαντικά το εύρος τιμών.

4.4.2.2 Δεδομένα από το Twitter

Έχουμε αναφερθεί εκτενώς τόσο στη διαδικασία ανάκτησης των tweets από τη διεπαφή της πλατφόρμας όσο και στη δομή των ληφθέντων δεδομένων. Σε αυτό το σημείο θα παρουσιάσουμε κάποια στατιστικά στοιχεία για τα δεδομένα μας.

Όπως είπαμε, έχουμε συλλέξει δημοσιεύσεις που αφορούν την εκπομπή *Amici di Maria de Filippi* και συγκεκριμένα περιέχουν το hashtag #amiciXX, με XX τον αιρθμό του κύκλου που προβάλλονται τη συγκεκριμένη περίοδο. Στο πρώτο εξάμηνο του 2016 καταγράφησαν 845072 δημοσιεύσεις που περιείχαν την ετικέτα #amici15, για τον 15ο κύκλο επεισοδίων του talent show. Αντίστοιχα στο δεύτερο εξάμηνο του έτους συγκεντρώθηκαν 103572 tweets χρησιμοποιώντας το #amici16, μιας και οι τηλεοπτικές σεζόν ξεκινούν φθινόπωρο κι όχι στην αρχή του έτους. Για το έτος 2017 τώρα, τα tweets του πρώτου εξαμήνου με την ετικέτα #amici16 μετρήθηκαν 882024 και για το δεύτερο εξάμηνο αντίστοιχα ανακτήσαμε 135288 δημοσιεύσεις που περιελάμβαναν το hashtag #amici17 ενόψει του νέου 17ου κύκλου της εκπομπής.

Ο λόγος για τον οποίο προτιμήθηκε ο συγκεκριμένος διαχωρισμός έχει να κάνει με τον περιορισμό που υπάρχει στα χρονικά παράθυρα από το Google Trends. Έχει σημειωθεί παραπάνω ότι το μέγιστο χρονικό παράθυρο για το οποίο η πλατφόρμα παρέχει ημερήσια δεδομένα παρατηρήθηκε να είναι οι 270 ημέρες. Έπειτα από αυτό το κατώφλι τα δεδομένα παρέχονται σε εβδομαδιαία βάση. Τυπικά δεν υπάρχει κάποιο πρόβλημα με αυτό, ωστόσο παρουσιάζεται ένα πρακτικό ζήτημα που χρήζει ρύθμισης.

Συγκεκριμένα, ένας από τους στόχους της παρούσας διπλωματικής εργασίας είναι η επαλήθευση της καταλληλότητας των δεδομένων που παρέχει το Google Trends για τη μελέτη των ενδιαφερόντων των χρηστών κοινωνικών δικτύων και συγκεκριμένα η ανάλυση της ακροαματικότητας τηλεοπτικών εκπομπών με βάση τη δραστηριότητα των χρηστών. Η ύπαρξη εβδομαδιαίας κλίμακας δεδομένων απέφερε περίπου 40 τιμές, αν αποκλείσουμε την περίοδο του καλοκαιριού που η δραστηριότητα είναι πολύ μειωμένη για να δώσει αξιόπιστα στοιχεία, πόσο μάλλον αν τυχόν σπάγαμε σε μικρότερα παράθυρα τα δεδομένα μας. Γίνεται κατανοητό ότι ένα τέτοιο δείγμα θέτει σε κίνδυνο την αξιοπιστία του αποτελέσματος και για αυτό προτιμήθηκε η λύση των δύο ξεχωριστών εξαμήνων ανά έτος, που αποφέρει περίπου 180 ημερήσιας κλίμακας τιμές και ένα πολύ πιο σημαντικό σε μέγεθος δείγμα υπό μελέτη.

Η συγκεκριμένη επιλογή προσθέτει ένα παράπλευρο πλεονέκτημα στο εγχείρημα μας. Μέσα στο ίδιο ημερολογιακό έτος υπάρχουν δύο εξάμηνα που διαφέρουν "τηλεοπτικά". Όπως εξηγήθηκε και παραπάνω με τις ετικέτες, το τηλεοπτικό έτος ξεκινάει το φθινόπωρο και ολοκληρώνεται την άνοιξη -θεωρώντας καταχρηστικά τηλεοπτικώς νεκρά τα καλοκαίρια- σε αναντιστοιχία με το ημερολογιακό έτος. Σε συνδυασμό με το

γεγονός ότι το επίσημο hashtag της υπό μελέτης εκπομπής περιέχει τον τηλεοπτικό κύκλο κι όχι τη χρονιά προβολής, υπάρχουν δύο διαφορετικές ετικέτες στο ίδιο ημερολογιακό έτος. Η σύγχυση αυτή αντιμετωπίζεται επιτυχώς από το διαχωρισμό σε δύο παράθυρα εξαμηνιαίας διάρκειας.

4.4.2.3 Συγχώνευση αρχείων δεδομένων

Από τα αναφερόμενα στην παραπάνω παράγραφο, συμπεραίνουμε ότι χρειάζεται να συγχωνεύσουμε τα μηνιαία αρχεία που παράγονται από τον κώδικα της Ενότητας 4.4.2.1 ώστε να πάρουμε τα ζητούμενα δεδομένα σε εύρος εξαμήνου. Αυτό επιτυγχάνεται με τις γραμμές κώδικα στον πίνακα που ακολουθεί. Ουσιαστικά αυτό που συμβαίνει είναι ότι ανοίγουμε ένα ένα τα αρχεία που θέλουμε να συγχωνεύσουμε και αντιγράφουμε το περιεχόμενο προσθετικά στο καινούριο αρχείο. Τα νέα αυτά αρχεία χρησιμοποιούνται για τα datasets που χρησιμοποιήθηκαν για την επαλήθευση της καταλληλότητας των δεδομένων της πλατφόρμας Google Trends.

```
fout=open(fout_name,'a') #a for appending
for month in months:
    f = open("Amici_"+month+year+".csv", 'r') #file paths should better be
listed as a whole, this hybrid just made the process easier to me
    #f.next() #skip header (not needed here)
    for line in f:
         fout.write(line)
    f.close() #not necessarily needed
fout.close()
```

Πίνακας 4.2: Κώδικας για τη συγχώνευση δεδομένων μέσω αντιγραφής σε κοινό αρχείο

5. Πειράματα και Αξιολόγηση

Σε αυτό το κεφάλαιο, μετά την ενδελεχή ανάλυση του απαραίτητου θεωρητικού υπόβαθρου και των τεχνικών που χρησιμοποιήθηκαν για τους σκοπούς της παρούσας εργασίας, επεξηγούνται εκτενέστερα κάποια στοιχεία της υλοποίησης, όπως οι τρόποι εύρεσης των κατάλληλων παραμέτρων και εισόδων για τα πειράματα και παρουσιάζονται τα κύρια αποτελέσματα και παρατηρήσεις.

5.1 Πλαίσιο πειραμάτων

Σκοπός της εργασίας είναι η εξαγωγή συμπερασμάτων για το ενδιαφέρον των χρηστών κοινωνικών δικτύων μέσω της δραστηριότητάς τους σε αυτά. Τα συμπεράσματα αυτά θα προκύψουν μέσα από την εκτίμηση της τηλεθέασης και του απόλυτου πλήθους των τηλεθεατών μιας τηλεοπτικής εκπομπής.

Η εκπομπή στην οποία βασίστηκε σχεδόν το σύνολο των πειραμάτων είναι το ιταλικό talent show με τίτλο *Amici di Maria de Filippi*. Πρόκειται για μια μουσική ακαδημία που φιλοξενεί 20 νεαρά άτομα μεταξύ 18 και 25 ετών που φιλοδοξούν να ξεκινήσουν μια καριέρα στο χώρο ως τραγουδιστές, στιχουργοί ή χορευτές. Το σόου ξεκίνησε το 2001 και έκτοτε προβάλλεται αδιαλείπτως έχοντας συμπληρώσει 17 κύκλους εκπομπών με περίπου 10 επεισόδια ανά κύκλο. Εμείς μελετάμε τις σεζόν 15 και 16 που προβλήθηκαν το 2016 και 2017 με 10 και 9 εβδομαδιαία επεισόδια αντίστοιχα. Η συγκεκριμένη εκπομπή απολαμβάνει μεγάλα νούμερα τηλεθέασης και γενικότερα υψηλή δραστηριότητα στα κοινωνικά δίκτυα, γεγονός που την καθιστά εξαιρετικό αντικείμενο μελέτης.

Τα δεδομένα που θα χρησιμοποιήσουμε προέρχονται από τις πλατφόρμες Google Trends και Twitter. Η πρώτη παρέχει κανονικοποιημένες μετρήσεις για τη σχετική συχνότητα αναζήτησης ενός συγκεκριμένου όρου υπό τη μορφή χρονοσειράς. Για το εύρος 2 χρόνων τα δεδομένα αυτά δίνονται σε εβδομαδιαία κλίμακα (gTrends). Από το δημόσιο API του Twitter μπορούμε να κατεβάσουμε και αποθηκεύσουμε το σύνολο των tweets που αφορούν ένα συγεκριμένο hashtag -εδώ το επίσημο της εκπομπής, όπως έχει εξηγηθεί- σε οποιοδήποτε χρονικό εύρος. Επιλέγουμε μηνιαία διαστήματα χωρίς να επηρεάζει κάπως αυτή η απόφαση στην επεξεργασία των δεδομένων. Από τα tweets αυτά υπολογίζουμε το ημερήσιο συνολικό πλήθος δημοσιεύσεων (tweets) και το ημερήσιο πλήθος μοναδικών χρηστών που δημοσίευσαν τα αντίστοιχα tweets (unqUsers). Τα ημερήσια αυτά δεδομένα μπορούν πλέον να ομαδοποιηθούν σε οποιοδήποτε χρονικό διάστημα επιθυμούμε. Αξίζει να αναφερθεί ότι το συνολικό πλήθος των tweets που ανασύρθηκαν ξεπερνάει τα 2 εκατομμύρια.

Τα 3 παραπάνω αποτελούν τα χαρακτηριστικά ή τις εισόδους στα μοντέλα. Οι επιθυμητές έξοδοι είναι 2: το ποσοστό τηλεθέασης και το εκτιμώμενο πλήθος τηλεθεατών που παρακολούθησαν το εκάστοτε επεισόδιο. Αυτές οι 5 ομάδες τιμών αποτελούν το αρχικό dataset, το οποίο θα υφίσταται μεταβολές και μετασχηματισμούς, όπως αναλύεται στα επόμενα. Πολύ συνοπτικά το σύνολο των δεδομένων μας έχει δομηθεί ως εξής:

- Κάθε γραμμή του set αντιστοιχεί σε ένα επεισόδιο της εκπομπής Amici di Maria de Filippi. Για τη διετία 2016-17 έχουμε συνολικά 19 επεισόδια, άρα και 19 γραμμές ή καλύτερα 19 στοιχεία δεδομένων.
- Η πρώτη στήλη περιλαμβάνει τη συνολική τηλεθέαση που κατέγραψε κάθε επεισόδιο επί του συνόλου των ατόμων που παρακολουθούσαν τηλεόραση εκείνο το διάστημα.

- Η δεύτερη στήλη αντίστοιχα περιέχει το συνολικό αριθμό μοναδικών ατόμων που παρακολούθησαν έστω για 1 λεπτό το συγκεκριμένο επεισόδιο.
- Η τρίτη στήλη περιέχει έναν αριθμό από το Google Trends, ο οποίος αναφέρεται στη σχετική κίνηση στο Google Search για τη συγκεκριμένη εκπομπή, όπως έχουμε περιγράψει και στην αρχή.
- Η τέταρτη στήλη περιλαμβάνει το πλήθος των tweets που αφορούν το συγκεκριμένο επεισόδιο της εκπομπής. Θα εξηγήσουμε στο επόμενο κεφάλαιο πώς υπολογίζουμε ποια tweets αφορούν ποιο επεισόδιο.
- Η πέμπτη στήλη αντίστοιχα περιέχει τον αριθμό των μοναδικών χρηστών που δημοσίευσαν τα παραπάνω tweets.

5.2 Πείραμα 1: Συσχέτιση δεδομένων Twitter - Google Trends

Τα κοινωνικά δίκτυα αποτελούν αντικείμενο μελέτης από επιστήμονες διαφόρων κλάδων για πολλές δεκαετίες. Με την ανάπτυξη του διαδικτύου έγινε ένα ξέσπασμα που οδήγησε στη δημιουργία διαφόρων μέσων κοινωνικής δικτύωσης, τα οποία με τη σειρά τους γνωρίζουν μεγάλη άνοιξη και αριθμούν εκατοντάδες εκατομμύρια ενεργούς χρήστες. Τα δημοφιλέστερα εξ αυτών είναι το Facebook, το Twitter και το Instagram. Τα δύο πρώτα αποτελούν σημαντική πηγή πληροφορίας, καθώς εκτός από φωτογραφίες επιτρέπονται και δημοσιεύσεις κειμένου, συζητήσεις κλπ. Είναι επομένως λογικό να έχουν αξιοποιηθεί τα δεδομένα που προσφέρουν σε ποικίλες ερευνητικές δραστηριότητες και πειράματα.

Ένας από τους στόχους της παρούσας εργασίας είναι να συμπεριλαμβάνονται και τα Google Trends δίπλα στα ονόματα των παραπάνω μέσων κοινωνικών δικτύων τουλάχιστον σε ό,τι έχει να κάνει με στατιστικά και εκτιμήσεις τηλεθέασης και να χρησιμοποιούνται τα δεδομένα που προσφέρει η πλατφόρμα για τον όγκο αναζήτησης συγκεκριμένων όρων για το συνδυασμό με δεδομένα που προέρχονται από Twitter (και Facebook) Αυτό θα επιτευχθεί μέσω επαλήθευσης για την καταλληλότητα της συγκεκριμένης υπηρεσίας.

Η παραπάνω "καταλληλότητα" αποφασίστηκε να μετρηθεί με τη μετρική της συσχέτισης (correlation). Συγκεκριμένα στο πείραμα χρησιμοποιήθηκε το Pearson's correlation coefficient ή δισδιάστατη συσχέτιση. Πρόκειται για έναν αριθμό μεταξύ -1 και 1, όπου το 1 επιδεικνύει πλήρη γραμμική συσχέτιση των δύο μεταβλητών, το -1 αντιστρόφως ανάλογη γραμμική συσχέτιση και το 0 καμία συσχέτιση των δύο σειρών δεδομένων. Κοινώς, όσο μεγαλύτερη η απόλυτη τιμή, τόσο μεγαλύτερη η συσχέτιση των 2 μεταβλητών. Η φόρμουλα υπολογισμού του συντελεστή αυτού είναι η εξής:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

όπου σ_A η διασπορά μιας μεταβλητής Α, μ_A η μέση τιμή μιας μεταβλητής Α και $E[A]$ η αναμενόμενη τιμή για την Α.

Σε συνδυασμό με το συντελεστή αυτό, υπολογίζεται και ένας δείκτης σημαντικότητας της μηδενικής υπόθεσης, ο οποίος υπολογίζει την πιθανότητα το παραχθέν αποτέλεσμα να βρέθηκε από σύμπτωση και να μην υπάρχει πραγματικά συσχέτιση μεταξύ των δύο μεταβλητών.

5.2.1 Το πείραμα

Για τους σκοπούς της εργασίας συγκεντρώσαμε τα ημερήσια δεδομένα που αφορούν την τηλεοπτική εκπομπή *Amici di Maria de Filippi* για το πρώτο (181 μέρες) και το δεύτερο (184 μέρες) εξάμηνο του 2017, για τους λόγους που έχουν αναλυθεί προηγουμένως.

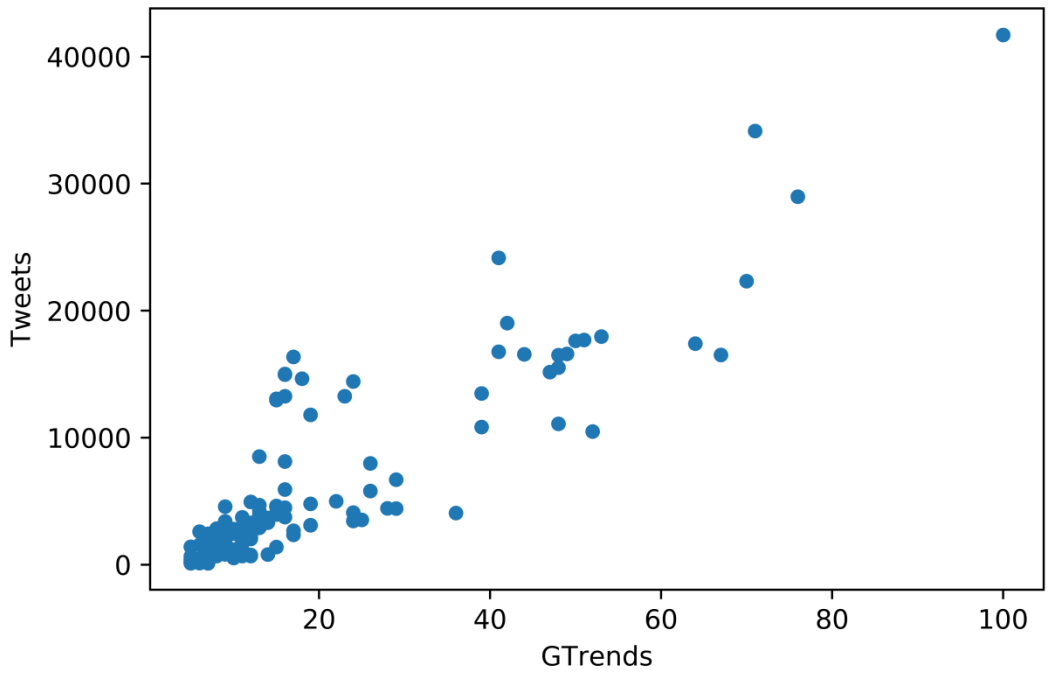
Έχοντας επίσης προεπεξεργαστεί τα δεδομένα, χρειάζονται ελάχιστες γραμμές κώδικα για να πάρουμε τη συσχέτιση των δύο συνόλων δεδομένων, οι οποίες παρουσιάζονται στον παρακάτω πίνακα.

```
df = pd.DataFrame()
df['GTrends'] = pd.read_csv(Gname)
df['Tweets'] = pd.read_csv(Tname)
df.corr(method='pearson')
df.plot(x='GTrends', y='Tweets', kind='scatter')
```

Πίνακας 5.1: Κώδικας python για την εύρεση συσχέτισης μεταξύ δύο μεταβλητών

5.2.2 Αποτελέσματα

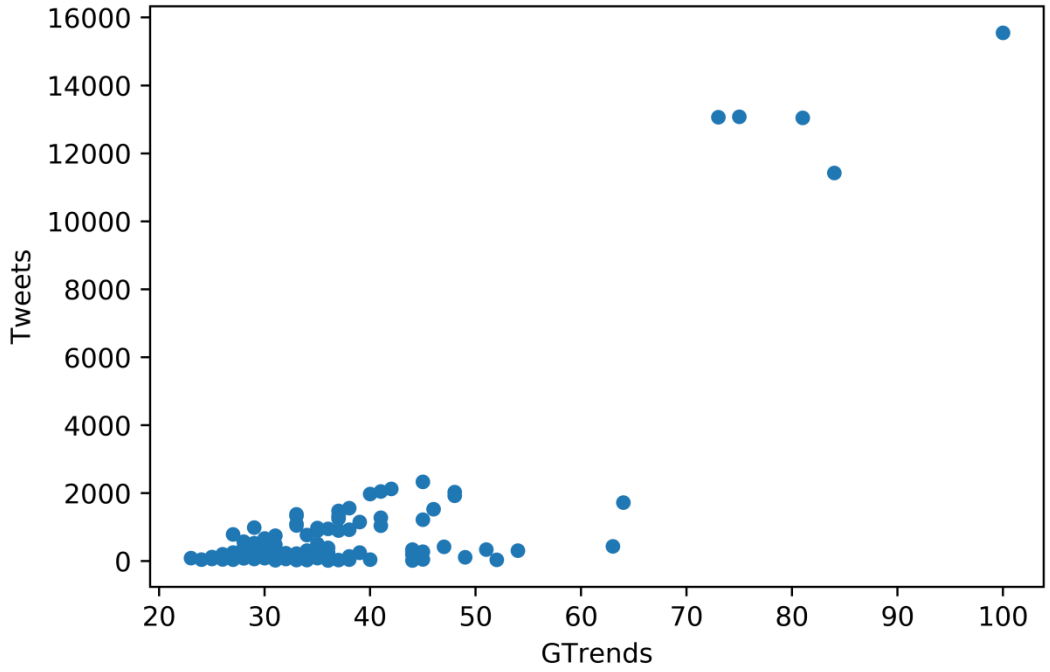
Για το πρώτο εξάμηνο του 2017 μετρήθηκε συντελεστής 0,893 και δείκτης σημαντικότητας της τάξης του 10^{-32}. Έχοντας εξασφαλίσει την 1-1 αντιστοιχία των δεδομένων από το Google Trends και το Twitter, το αποτέλεσμα αυτό μας επιβεβαιώνει ότι υπάρχει ισχυρή γραμμική συσχέτιση μεταξύ των δύο συνόλων δεδομένων. Παράλληλα, το χαμηλότατο σκορ σημαντικότητας ουσιαστικά αποκλείει την πιθανότητα σύμπτωσης της γραμμικότητας στη συναρτησιακή σχέση των δεδομένων. Στην Εικόνα 5.1 γίνεται αναπαράσταση των σημείων που δημιουργούνται αν βάλουμε ως συντεταγμένες τα δεδομένα από τα δύο μέσα. Εκεί μπορούμε πράγματι να δούμε ότι υπάρχει μια αρκετά ισχυρή συσχέτιση γραμμικής φύσεως, αφού τα σημεία βρίσκονται πάνω και γύρω από μια ευθεία γραμμή και μάλιστα με μικρή σχετικά διασπορά.



Εικόνα 5.1: Συσχέτιση των δεδομένων Google Trends και Twitter για το 1ο εξάμηνο του 2017

Για το δεύτερο εξάμηνο του 2017 μετρήθηκε συντελεστής 0,816 και δείκτης σημαντικότητας της τάξης του 10^{-30}. Ο χαμηλότερος συντελεστής που εμφανίζεται στο

δεύτερο εξάμηνο μπορεί να ερμηνευτεί από το γεγονός ότι σε αυτά τα δεδομένα περιλαμβάνονται οι καλοκαιρινοί μήνες που η δραστηριότητα (ειδικά στο Twitter) αγγίζει το μηδέν. Στην Εικόνα 5.2 είναι εμφανές ότι οι υπάρχει σημαντικός αριθμός τέτοιων σημείων. Ωστόσο, ακόμα κι έτσι διαφαίνεται μια τάση να προσομοιωθεί μια ευθεία γραμμή και σίγουρα τα δεδομένα δεν εμφανίζουν έναν τυχαίο διασκορπισμό στο χώρο, γεγονός που θα έδειχνε μη συσχέτιση των 2 συνόλων.



Εικόνα 5.2: Συσχέτιση των δεδομένων Google Trends και Twitter για το 2ο εξάμηνο του 2017

Τα ευρήματα του πειράματος αυτού είναι τα αναμενόμενα και επιθυμητά: Τα δεδομένα από Google Trends και Twitter για το ίδιο θέμα και χρονικό παράθυρο έχουν ισχυρή γραμμική συσχέτιση μεταξύ τους και αυτό σημαίνει ότι μπορούμε να χρησιμοποιήσουμε την πλατφόρμα σε επόμενες ερευνητικές δραστηριότητες.

5.3 Πείραμα 2: Εξαγωγή στατιστικών τηλεθέασης βάσει των δεδομένων Twitter και Google Trends

Ο βασικός σκοπός της εργασίας είναι η χρήση μηχανικής μάθησης για την εξαγωγή αποτελεσμάτων αναφορικά με την εκτίμηση της τηλεθέασης προγραμμάτων. Το παρόν πείραμα αποτελεί τη διαδικασία ανάκτησης των προαναφερθέντων αποτελεσμάτων. Αρχικά, θα πρέπει να προσδιορίσουμε εκ των προτέρων κάποιες παραμέτρους με σκοπό να μειώσουμε τον αριθμό των εκτελέσεων και των διαφορετικών στιγμιοτύπων. Έπειτα ακολουθεί το κυρίως μέρος του πειράματος με την εφαρμογή των διαφορετικών αλγορίθμων παλινδρόμησης -όπως έχει περιγραφεί στις προηγούμενες ενότητες- για την εξαγωγή μοντέλου εκτίμησης τηλεθέασης και την επαλήθευση των αποτελεσμάτων από το γενετικό αλγόριθμο.

5.3.1 Εύρεση του βέλτιστου χρονικού παραθύρου πειραμάτων

Πιθανότατα το πιο δύσκολο κομμάτι σε κάθε εργασία είναι ο προσδιορισμός των κατάλληλων παραμέτρων που θα δοθούν ως είσοδοι στο πρόγραμμα που υλοποιείται, ούτως ώστε να επιτευχθεί το βέλτιστο αποτέλεσμα. Έτσι και εδώ υπάρχουν πολλά στοιχεία που επηρεάζουν τη λειτουργία του μοντέλου και εν τέλει την αποδοτικότητα των αλγορίθμων για την εκτίμηση της τηλεθέασης, τα οποία έχουν αρκετές διαφορετικές τιμές ή κλάσεις τιμών που σημαίνει αρκετά πειράματα για την εύρεση της βέλτιστης επιλογής.

Μια εξίσου μεγάλη δυσκολία είναι να βρεθεί η σωστή σειρά προσδιορισμού των παραμέτρων. Το πρόβλημα αυτό απορρέει από το γεγονός ότι κάθε παράμετρος αποτελεί ουσιαστικά τη ρίζα ενός δέντρου με κλάδους τις δυνατές επιλογές για αυτή την παράμετρο και υποδέντρα για κάθε συνδυασμό παραμέτρων. Εύκολα συμπεραίνει κανείς ότι για κάθε επιπλέον παράμετρο του μοντέλου, αυξάνονται εκθετικά τα πραγματοποιούμενα πειράματα και κατά συνέπεια ο συνολικός φόρτος δουλειάς. Συνεπώς, όχι απλά χρειάζεται να προσδιοριστούν μία-μία οι παράμετροι για να μειωθεί ο φόρτος και ο χρόνος δουλειάς, αλλά επιβάλλεται η όλη διαδικασία να γίνει αποδοτικά για να αποφέρει τα μέγιστα οφέλη τόσο σε χρόνο όσο φυσικά και στα αποτελέσματα καθαυτά.

Η παράμετρος που υποδείχθηκε ως αυτή που χρειάζεται να οριστικοποιηθεί εξαρχής είναι το παράθυρο, μέσα στο οποίο θα επιλέξουμε να λάβουμε υπόψη τα δεδομένα τόσο από Google Trends όσο και από το Twitter. Μια σειρά πειραμάτων απέδωσε το καταλληλότερο διάστημα για καθεμία από τις δύο πλατφόρμες που χρησιμοποιούμε στην παρούσα έρευνα.

5.3.1.1 Google Trends

Για την ιστοσελίδα τα πράγματα είναι αρκετά απλά στη συγκεκριμένη παράμετρο. Ο λόγος είναι ότι η πλατφόρμα παρέχει τα εβδομαδιαία δεδομένα με μοναδικό τρόπο, ο οποίος είναι να μετράει την εβδομάδα από την Κυριακή μέχρι το επόμενο Σάββατο. Δεδομένου ότι τα επεισόδια της εκπομπής *Amici di Maria de Filippi* προβάλλονται ημέρα Σάββατο, μας δίνει δύο επιλογές για τα δεδομένα του Google Trends: η πρώτη είναι να επιλέγονται τα δεδομένα που αφορούν την εβδομάδα που ολοκληρώνεται την ημέρα που εκπέμπει το επεισόδιο (πρότερη αναζήτηση) και η δεύτερη είναι να επιλέγονται τα δεδομένα που αφορούν την εβδομάδα μετά το επεισόδια (κατοπινή αναζήτηση).

Πρότερη αναζήτηση	Κατοπινή αναζήτηση
22.542%	25.979%
16.158%	17.042%
13.295%	15.611%
6.953%	8.647%

Πίνακας 5.2: Σύγκριση του μέσου ποσοστιαίου σφάλματος στις δύο επιλογές χρονικού παραθύρου των Google Trends για διάφορες τιμές των υπόλοιπων παραμέτρων

Είναι προφανές ότι τα παραπάνω αποτελέσματα δεν είναι βέλτιστα, ούτε φυσικά αποτελούν κατ' ανάγκη αντιπροσωπευτικές επιδόσεις του τελικού μοντέλου, παρότι έχουμε κάποια ενθαρρυντικά αποτελέσματα. Πρόκειται για μετρήσεις για τις δύο επιλογές παραθύρου στο Google Trends χρησιμοποιώντας διάφορους συνδυασμούς τιμών για τις υπόλοιπες παραμέτρους. Το αποτέλεσμα είναι **καθολικά υπέρ της πρότερης αναζήτησης**, μιας και δίνει σημαντικά χαμηλότερο μέσο ποσοστιαίο σφάλμα σε κάθε πείραμα, για όλους τους χρησιμοποιούμενους συνδυασμούς των υπολοίπων παραμέτρων.

Το αποτέλεσμα είναι εν μέρει αναμενόμενο και δικαιολογείται από το γεγονός ότι αυξημένο ενδιαφέρον για μια εκπομπή πριν την προβολή μπορεί να οδηγήσει τελικά σε αυξημένη ακροαματικότητα του επόμενου επεισοδίου. Από την άλλη ενδιαφέρον θα είχε ως αποτέλεσμα να παρατηρηθεί μια αντιστρόφως ανάλογη σχέση ανάμεσα στον όγκο αναζήτησης της εκπομπής στο Google μετά την εκπομπή και στα νούμερα που αυτή απέφερε, ωστόσο κάτι τέτοιο δεν παρατηρήθηκε στα πραγματοποιούμενα πειράματα. Μια τέτοια παρατήρηση δεν απορρίπτεται και αντιθέτως κρατείται για μελλοντική εργασία.

Ως συμπέρασμα, μπορούμε να πούμε ότι έχουμε την πρώτη οριστικοποιημένη παράμετρο: Στα επόμενα πειράματα χρησιμοποιούμε τα δεδομένα του Google Trends για τις αναζητήσεις που πραγματοποιούνται **όλη την εβδομάδα πριν και ανήμερα της εκπομπής**.

5.3.1.2 Twitter

Για το κοινωνικό αυτό δίκτυο τα πράγματα γίνονται λίγο πιο δύσκολα. Από τη στιγμή που έχουμε τα ημερήσια δεδομένα για τον αριθμό των δημοσιεύσεων και τους μοναδικούς χρήστες που δραστηριοποιούνται στο Twitter σχετικά με την εκπομπή, μπορούμε να χρησιμοποιήσουμε πάρα πολλά διαφορετικά παράθυρα για να αποσπάσουμε πληροφορία. Επίσης, υπάρχει η δυνατότητα να εφαρμόσουμε διαφορετικά βάρη στις επιλεχθείσες ημέρες, ώστε πχ η ημέρα προβολής να έχει μεγαλύτερη βαρύτητα στο άθροισμα από τις υπόλοιπες, μιας και είναι εξαιρετικά πιθανό να δημοσιεύει κάποιος σχετικά με την εκπομπή ενόσω την παρακολουθεί.

Όπως και στα Google Trends, μπορούμε να χωρίσουμε τα δεδομένα μας σε παράθυρα της μιας εβδομάδας και να δώσουμε ως είσοδο στο μοντέλο το συνολικό άθροισμα. Φυσικά δεν υπάρχει ο ίδιος περιορισμός με πριν και μπορούμε να μετακυλήσουμε το παράθυρο κατά το δοκούν. Μια προφανής επιλογή είναι να τοποθετήσουμε την ημέρα προβολής στο μέσο κάθε βδομάδας και να αθροίζουμε τη δραστηριότητα των τριών προηγούμενων και τριών επόμενων ημερών αντίστοιχα.

Εμφανίζοντας τα ημερήσια δεδομένα σε γράφημα παίρνουμε μια πολύ ενδιαφέρουσα πληροφορία προς την κατεύθυνση επιλογής γραφήματος. Ενώ κατά τους μήνες προβολής της εκπομπής υπάρχει συνολικά υψηλή δραστηριότητα, παρατηρούμε ότι οι ημέρες προβολής και η επόμενη αυτών εμφανίζουν ένα ιδιαίτερο ξέσπασμα. Κι αν για την ημέρα προβολής είναι απόλυτα φυσιολογικό, η μεγάλη διαφορά στα νούμερα της ακριβώς επόμενης ημέρας σε σύγκριση με τις υπόλοιπες της εβδομάδες είναι ένα άκρως ενδιαφέρον εύρημα για τον προσανατολισμό των πειραμάτων.

Πράγματι η παρατήρηση αυτή επιβεβαιώθηκε πειραματικά, μιας και η εισαγωγή των επιπλέων ημερών δε βελτίωσε καθόλου την απόδοση του συγκεκριμένου feature. Επομένως, θα επικεντρώσουμε τα πειράματά μας πάλι σε δύο επιλογές: στη μέτρηση δημοσιεύσεων μόνο κατά την ημέρα προβολής (1) και την αντίστοιχη μέτρηση για την ημέρα προβολής και την επόμενή της (2).

Αξίζει να σημειωθεί ότι τα δεδομένα της επόμενης από την προβολή ημέρας έχουν προστεθεί στο μοντέλο με συντελεστή βάρους 0,5. Η χρήση του συντελεστή κρίθηκε απαραίτητη μετά τις δοκιμές που έγιναν και προκρίθηκε η επιλογή του 0,5 ως το βέλτιστο ποσοστό συμμετοχής.

1	2
23.868%	23.668%
11.626%	10.521%
8.047%	8.989%
7.958%	8.847%
8.016%	7.616%
7.911%	9.068%

Πίνακας 5.3: Μέσο ποσοστιαίο σφάλμα του χαρακτηριστικού *tweets* για τις δύο επιλογές χρονικού παραθύρου

Πειραματιζόμενοι με τα δεδομένα των δημοσιεύσεων για 1 και 2 ημέρες αντίστοιχα καταλήξαμε στον παραπάνω πίνακα. Τα αναφερόμενα ποσοστά είναι -όπως πάντα- το μέσο ποσοστιαίο σφάλμα της εκτίμησης του ποσοστού τηλεθέασης της τηλεοπτικής εκπομπής *Amici di Maria de Filippi*. Παρατηρούμε ότι όσο κατεβαίνουμε σε ποσοστά -άρα χρησιμοποιούμε και καλύτερα μοντέλα- το νούμερο 1 έχει σαφώς καλύτερη απόδοση, παρότι το αναμενόμενο ήταν οι 2 ημέρες να παρέχουν πιο αντιπροσωπευτικό δείγμα της ακροαματικότητας. Επομένως, θα χρησιμοποιήσουμε τα δεδομένα μόνο της ημέρας της εκπομπής.

1	2
23.668%	23.279%
10.537%	10%
8.047%	9.068%
8.684%	8.763%
8.016%	7.295%
7.026%	7.111%

Πίνακας 5.4: Μέσο ποσοστιαίο σφάλμα του χαρακτηριστικού *unqUsers* για τις δύο επιλογές χρονικού παραθύρου

Αναφορικά με τους μοναδικούς χρήστες, υπάρχει πάλι μία μικρή τάση υπέρ των δεδομένων μίας ημέρας στα καλύτερα μοντέλα και παρόλο που υπάρχει και ένα πείραμα που προκρίνει την επιλογή των 2 ημερών, θα προκριθεί κι εδώ **η επιλογή των ημερήσιων δεδομένων**. Οι λόγοι της επιλογής είναι δύο: τα δεδομένα της μίας ημέρας είναι τα απλούστερα που μπορούμε να έχουμε σε κάθε περίπτωση και εφόσον δεν υπάρχει πρόσθετο κέρδος, δεν υπάρχει και λόγος να προβούμε σε μια πιο σύνθετη επιλογή μεταβλητής.

5.3.2 Αξιολόγηση με τεχνικές παλινδρόμησης

Όπως έχει ήδη γίνει κατανοητό και από τη θεωρητική ανάλυση, ο αλγόριθμος μηχανικής μάθησης που ταιριάζει στα δεδομένα και κυρίως στην επιθυμητή (συνεχή) έξοδο του μοντέλου μας είναι η παλινδρόμηση. Κατά τη διάρκεια των πειραμάτων εξερευνήθηκαν διαφορετικές εκδοχές του αλγορίθμου, στην προσπάθεια να προσομοιώσουμε το δυνατόν καλύτερα την ακροαματικότητα των επεισοδίων.

Συγκεκριμένα σε αυτή την ενότητα θα δούμε τους αλγορίθμους της γραμμικής παλινδρόμησης, της πολυωνυμικής παλινδρόμησης και της γενικής μη γραμμικής παλινδρόμησης, καθώς και μια υβριδική υλοποίηση κάνοντας χρήση μεθόδων γενετικών αλγορίθμων.

5.3.2.1 Cross-Validation

Προτού προχωρήσουμε στα πειράματα, αξίζει να αναλύσουμε τη μεθοδολογία που ακολουθείται για την εκτίμηση της ακροαματικότητας. Η μέθοδος που χρησιμοποιούμε για την εκτίμηση ονομάζεται cross-validation. Όπως μαρτυρά το όνομα, πρόκειται για μια τεχνική εξακρίβωσης του αποτελέσματος μέσω διασταύρωσης των εκτιμήσεων.

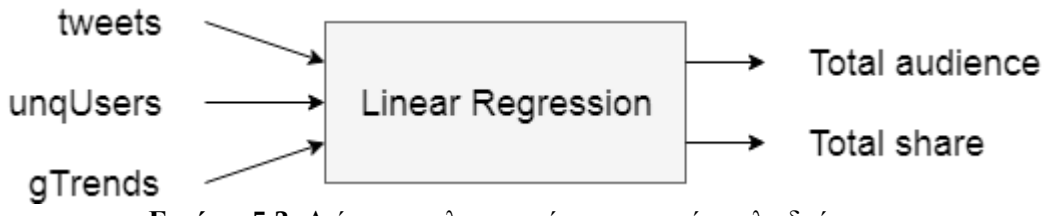
Το μικρό σύνολο δεδομένων δε μας επιτρέπει να κάνουμε το διαχωρισμό σε σύνολο εκπαίδευσης και ελέγχου αντίστοιχα, επομένως η παραπάνω επιλογή αποτελεί μονόδρομο. Ουσιαστικά πρόκειται για μια επαναληπτική διαδικασία διαχωρισμού του dataset σε train και test για την εκτίμηση των παραμέτρων. Οι τελικές παράμετροι υπολογίζονται παίρνοντας τον μέσο όρο των παραμέτρων που προκύπτουν από το cross-validation. Υπάρχουν εξαντλητικές και δειγματοληπτικές διαδικασίες. Οι δεύτερες επιταχύνουν τον αλγόριθμο, μιας και δε δοκιμάζονται όλοι οι πιθανοί διαχωρισμοί, αλλά το μικρό μέγεθος του dataset μας επιτρέπει να τρέχουμε τον εξαντλητικό αλγόριθμο.

Ο αλγόριθμος που εφαρμόζουμε ονομάζεται Leave-One-Out Cross-Validation. Πρόκειται για μια ειδική περίπτωση του αλγορίθμου Leave-P-Out, στον οποίο δημιουργούνται όλοι οι συνδυασμοί p δεδομένων και σε κάθε γύρο του αλγορίθμου ένα σύνολο p δεδομένων αποτελεί το test set και τα υπόλοιπα το train set. Από τη θεωρία Συνδυαστικής γνωρίζουμε ότι για p=1 ελαχιστοποιείται ο αριθμός των συνδυασμών και συγκεκριμένα παίρνει την τιμή n, όπου n το μέγεθος του dataset. Έτσι λοιπόν, ο συγκεκριμένος αλγόριθμος ουσιαστικά συνίσταται στην εκτίμηση της τιμής για ένα στοιχείο των δεδομένων βάσει όλων των υπολοίπων. Για τα 19 υπό μελέτη επεισόδια θα γίνουν 19 εκτελέσεις του αλγορίθμου και αντίστοιχα 19 εκτιμήσεις της ακροαματικότητας, μία τη φορά. Το τελικό υπερεπίπεδο που προσομοιώνει τα δεδομένα μας θα δοθεί από το μέσο όρο των παραμέτρων που προκύπτουν από κάθε εκτίμηση.

Με τον παραπάνω αλγόριθμο επιτυγχάνουμε τη μέγιστη δυνατή εξακρίβωση της αποδοτικότητας του μοντέλου χωρίς να μας προβληματίζει το ομολογουμένως μικρό μέγεθος του dataset.

5.3.2.2 Γραμμική παλινδρόμηση

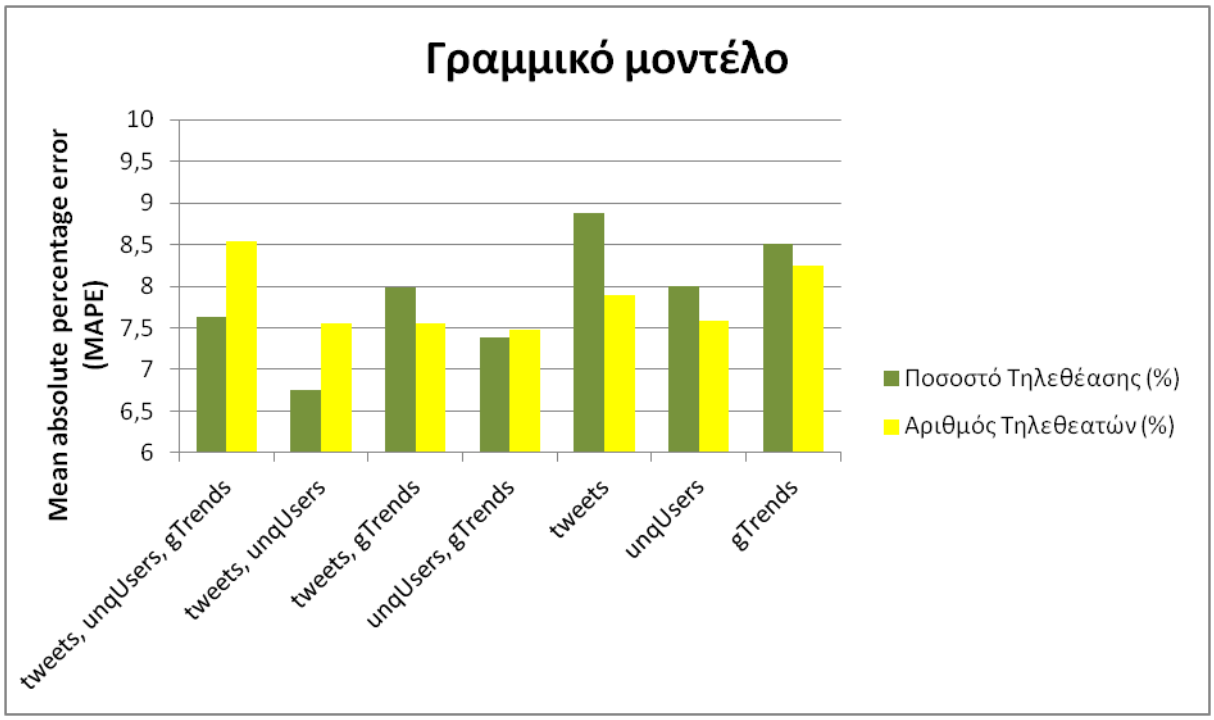
Πρόκειται για το πιο απλό από τα μοντέλα παλινδρόμησης πολλών εισόδων, ενώ ταυτόχρονα είναι αυτό πάνω στο οποίο ουσιαστικά χτίζονται και λειτουργούν και όλα τα υπόλοιπα. Όπως έχουμε περιγράψει, αντιστοιχεί τα δεδομένα εισόδου σε μια συνάρτηση εξόδου γραμμική ως προς κάθε μεταβλητή και βάρος. Μαθηματικά, αυτό σημαίνει ότι ο αλγόριθμος υπολογίζει την ευθεία (για μια μεταβλητή εισόδου), το επίπεδο (για δύο) ή το υπερεπίπεδο (για τρεις και άνω) που ελαχιστοποιεί τη συνολική απόσταση των εισόδων.



Εικόνα 5.3: Διάγραμμα λειτουργίας γραμμικής παλινδρόμησης

Οι μεταβλητές εισόδου είναι τρεις: Ο συνολικός όγκος δημοσιεύσεων που αντιστοιχεί με βάση την ανάλυση της προηγούμενης παραγράφου στο εκάστοτε επεισόδιο (tweets), ο αριθμός μοναδικών χρηστών που δημοσίευσαν σχετικά με την εκπομπή στο ίδιο

διάστημα (unqUsers) και τα δεδομένα από το Google Trends (gTrends), όπως έχουν αναλυθεί. Αυτές οι τρεις μεταβλητές εξετάζονται σε όλα τα μοντέλα προσπαθώντας να ανακαλύψουμε το βέλτιστο συνδυασμό τους για την εκτίμηση της ακροαματικότητας των τηλεοπτικών εκπομπών. Τα αποτελέσματα των πειραμάτων με επιθυμητή έξοδο τόσο τον αριθμό των μοναδικών τηλεθεατών όσο και το ποσοστό τηλεθέασης του εκάστοτε επεισοδίου παρουσιάζονται στο γράφημα της Εικόνας 5.3.



Εικόνα 5.4: Μέσο απόλυτο ποσοστιαίο σφάλμα για όλους τους δυνατούς συνδυασμούς εισόδων - Γραμμικό μοντέλο

Μια πολύ σημαντική παρατήρηση που προκύπτει από τα πειράματα αυτά είναι ότι τα tweets και οι users συνδυάζονται καλά μεταξύ τους και μάλιστα με αρκετά μικρότερο ποσοστό σφάλματος για την περίπτωση του ποσοστού τηλεθέασης. Θεωρητικά δεν περιμέναμε να συμβεί κάτι τέτοιο μιας και πρόκειται για δύο μεταβλητές με μεγάλη συσχέτιση μεταξύ τους -της τάξης του 0,98- και συνεπώς φαίνεται πολύ δύσκολο να μπορούν να προσφέρουν διαφορετική πληροφορία. Όμως υπάρχει εξήγηση, αρκεί να αναλογιστούμε τι συμβαίνει όταν περνάμε από τη μία διάσταση στις δύο. Πράγματι, είναι λογικό δύο μεταβλητές που συσχετίζονται γραμμικά να απλώνονται καλύτερα στο χώρο από κοινού, από ό,τι η καθεμία ξεχωριστά στο επίπεδο. Επομένως, το επίπεδο που προσεγγίζει τα δεδομένα στις δύο διαστάσεις προσαρμόζεται καλύτερα στα δεδομένα από τις αντίστοιχες ευθείες του κάθε χαρακτηριστικού.

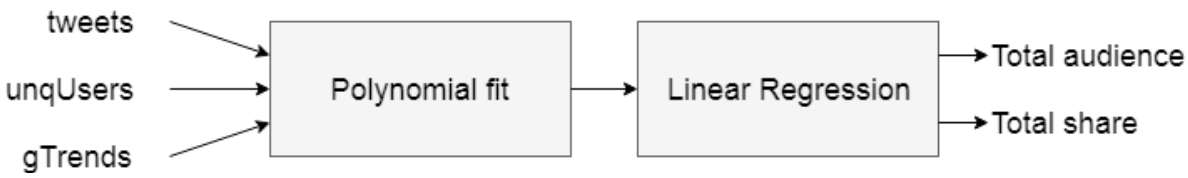
Παρατηρήθηκε επίσης ότι τα δεδομένα από το Google Trends επιφέρουν μείωση του σφάλματος περίπου μία μονάδα στην περίπτωση της εκτίμησης του ποσοστού τηλεθέασης και αποδεικνύουν για μία ακόμα φορά ότι αποτελούν δεδομένα αξιόπιστα και σημαντικά για ερευνητικούς σκοπούς.

Επιστρέφοντας στο πρώτο συμπέρασμα, οι δύο πανομοιότυπες μεταβλητές καταφέρνουν να προσφέρουν από κοινού επιπλέον πληροφορία και βελτιστοποίηση στα μοντέλα, αλλά δείχνουν να δυσχεραίνουν την αποδοτικότητα του μοντέλου όταν

χρησιμοποιούνται μαζί με τα δεδομένα από το Google Trends. Από την προηγούμενη χωρική ανάλυση δε δικαιολογείται η αύξηση που παρατηρούμε στη χρήση και των τριών μεταβλητών μαζί (πάνω αριστερά τιμή στους πίνακες) σε σχέση με τη χρήση οποιασδήποτε από τις δύο μεταβλητές από κοινού με τα δεδομένα του Google Trends.

Το παραπάνω φαινόμενο ονομάζεται "κατάρα της διαστατικότητας". Ο όρος αυτός, που εισήχθη πρώτη φορά το 1961 από τον Richard E. Bellman, αναφέρεται στην ανάλυση δεδομένων πολλών μεταβλητών καθώς αυξάνεται η διάσταση. Συγκεκριμένα, ο ισχυρισμός είναι ότι για δεδομένο αριθμό δειγμάτων υπάρχει μία μέγιστη διάσταση-ορόσημο, πάνω από την οποία μειώνεται η απόδοση του μοντέλου και μάλιστα πρόκειται στη συνήθη περίπτωση για αρκετά μικρό αριθμό. Έτσι και στα πειράματά μας, τα δισδιάστατα μοντέλα φαίνεται να δουλεύουν επαρκώς και -κοντά- στο βέλτιστο.

5.3.2.3 Πολυωνυμική παλινδρόμηση

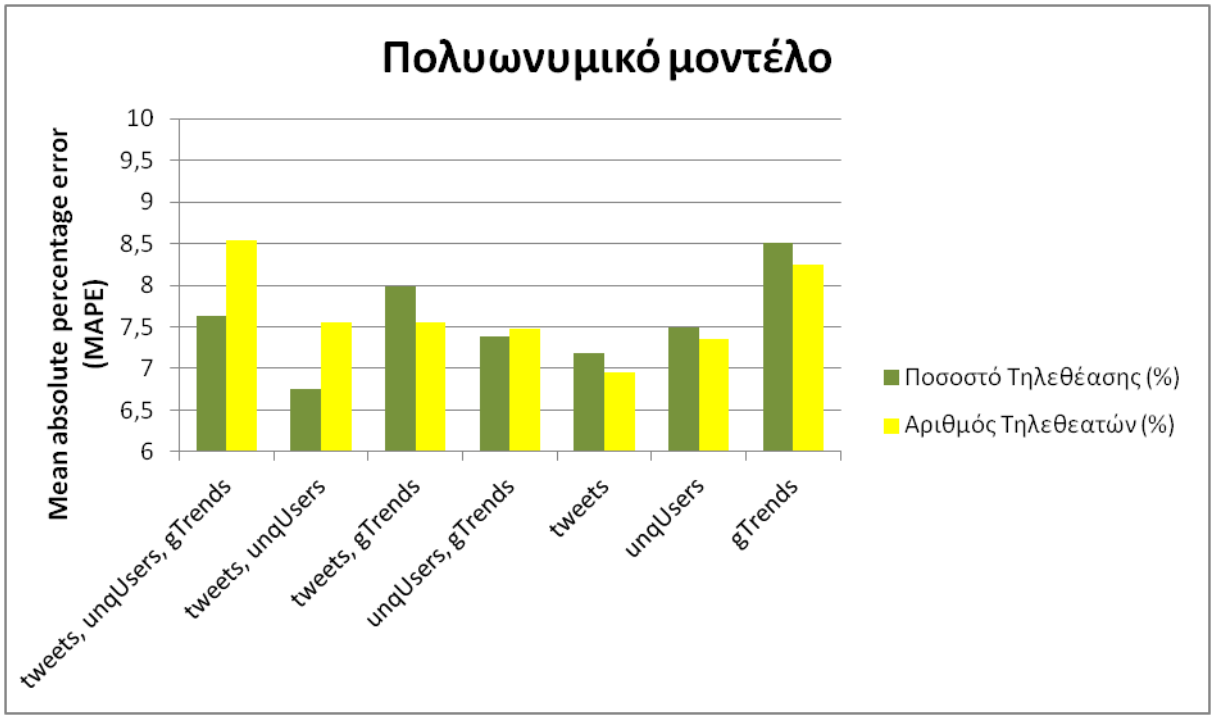


Εικόνα 5.5: Διάγραμμα λειτουργίας πολυωνυμικής παλινδρόμησης

Το μοντέλο αυτό σχετίζεται άμεσα με το αντίστοιχο γραμμικό. Για την ακρίβεια, σε θέμα υλοποίησης δεν υπάρχει καμία απολύτως διαφορά. Η διαφορά έγκειται στο ότι πλέον οι μεταβλητές y συσχετίζονται με κάποιο πολυώνυμο των ανεξάρτητων μεταβλητών x κι όχι με τις ίδιες τις τιμές απαραίτητα. Για παράδειγμα, για δύο μεταβλητές x, z και μέγιστο βαθμό 2 προσπαθούμε να εκτιμήσουμε το πολυωνυμικό μοντέλο που δίνεται από τον τύπο

$$y = \beta_{00} + \beta_{10}x + \beta_{01}z + \beta_{20}x^2 + \beta_{11}xz + \beta_{02}z^2 + \varepsilon.$$

Για περισσότερες μεταβλητές και βαθμό έχουμε όλα τα δυνατά πεπλεγμένα πολυώνυμα μέχρι n βαθμού σε πλήρη αντιστοιχία με το προηγούμενο. Στο παρακάτω γράφημα βλέπουμε τις καλύτερες εκτιμήσεις, πάντα βάσει μέσου ποσοστιαίου σφάλματος.



Εικόνα 5.6: Μέσο απόλυτο ποσοστιαίο σφάλμα για όλους τους δυνατούς συνδυασμούς εισόδων - Πολυωνυμικό μοντέλο

Μια σημαντική παρατήρηση είναι ότι στους περισσότερους συνδυασμούς εισόδων η βέλτιστη επίδοση ή ακριβέστερα το ελάχιστο μέσο ποσοστιαίο σφάλμα επιτυγχάνεται για βαθμό πολυωνυμικής συνάρτησης ίσο με 1. Παρατηρείται δηλαδή ότι το βέλτιστο μοντέλο πολυωνυμικής παλινδρόμησης είναι το γραμμικό, το οποίο ακριβώς μελετήσαμε και στην προηγούμενη υποενότητα.

Το παραπάνω αποτέλεσμα δεν είναι ακριβώς αναπάντεχο. Μολονότι θα θέλαμε όσο ανεβάζουμε την πολυπλοκότητα των μοντέλων να επιτυγχάνονται καλύτερα αποτελέσματα, αυτό δεν είναι κάτι που μπορεί να εξασφαλιστεί απλά και μόνο επειδή αυξάνουμε το βαθμό του πολυωνύμου. Το γεγονός, δηλαδή, ότι αναζητούμε μια ακριβέστερη από τη γραμμική συνάρτηση ως εκτίμηση δε συνεπάγεται μεγαλύτερη απόδοση, γιατί πολύ απλά ενδέχεται να μην υπάρχει καλύτερη καμπύλη εκτίμησης της κατανομής των δεδομένων από το υπερεπίπεδο.

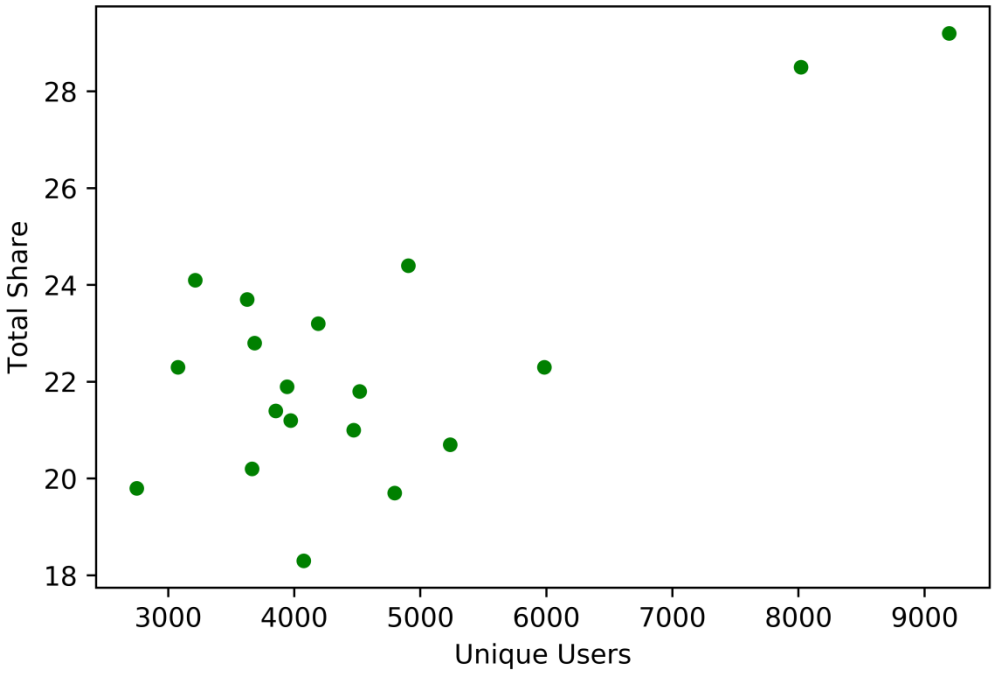
Παρατηρείται φυσικά ότι οι περιπτώσεις στις οποίες επικρατεί το γραμμικό μοντέλο των πολυωνυμικών προσεγγίσεων δίνουν ακριβώς την ίδια απόδοση. Τυπικά η υλοποίηση διαφέρει από τη στιγμή που στα τελευταία πειράματα δημιουργείται μια πολυωνυμική συνάρτηση, ωστόσο πρακτικά και κατ' ουσίαν δημιουργείται το ίδιο μοντέλο για βαθμό 1. Συγκεκριμένα, η υλοποίηση της γραμμικής παλινδρόμησης στη γλώσσα Python υπολογίζει τη μεροληψία (σταθερός όρος) χρησιμοποιώντας τον κλειστό τύπο από τον οποίον υπολογίζεται στη στατιστική μέθοδο των ελαχίστων τετραγώνων. Έτσι, το γραμμικό μοντέλο παίρνει την εκτίμηση της μεροληψίας χωρίς τη διαδικασία της προσαρμογής στα δεδομένα. Στο πολυωνυμικό μοντέλο η μεροληψία αποτελεί μέλος του συστήματος μηχανικής μάθησης. Κατά τη μετατροπή της εισόδου -όπως εξηγήθηκε στην αρχή της υποενότητας- παράγεται μία στήλη που αποτελείται μόνο από '1' και δίνεται μαζί με τις υπόλοιπες εισόδους. Αυτοί είναι οι όροι μεροληψίας για κάθε στοιχείο του συνόλου δεδομένων. Αποδεικνύεται λοιπόν πειραματικά ότι ο κλειστός τύπος της μεροληψίας για τον αλγόριθμο των ελαχίστων τετραγώνων είναι ακριβώς αυτό που μαθαίνει το μοντέλο μηχανικής μάθησης κατά το στάδιο της προσαρμογής.

5.3.2.4 Μη γραμμική παλινδρόμηση

Κάθε σύνολο δεδομένων που δεν μπορεί να συσχετιστεί με κάποια από τις δύο παραπάνω κατηγορίες μπορεί να αναλυθεί με ένα μη γραμμικό μοντέλο. Η γραμμικότητα αυτή συνήθως αντιστοιχεί στον τρόπο υπολογισμού των βαρών β, αλλά εν γένει ισχύει ο παραπάνω ορισμός. Όπως σχολιάστηκε κατά την παρουσίαση ένας τρόπος να χειριστούμε δεδομένα που φαίνεται να αντιστοιχούν σε τέτοια μοντέλα είναι η γραμμικοποίηση. Στη συγκεκριμένη περίπτωση θα μετασχηματίσουμε τα δεδομένα εισόδου ώστε να μπορούν να αποτελέσουν αντιπροσωπευτική είσοδο και να έχει μια φυσιολογική απόδοση το γραμμικό μοντέλο.

Ο τρόπος με τον οποίο επιτυγχάνεται η γραμμικοποίηση είναι με τη μελέτη και ανάλυση των δεδομένων για την εύρεση μοτίβων μέσα στο σύνολο δεδομένων και ιδανικά έναν κλειστό τύπο που δίνει τη συνάρτηση της εξόδου από τα δεδομένα εισόδου. Ένας απλός τρόπος να αποσπαστεί πληροφορία από τα δεδομένα εισόδου είναι η γραφική τους απεικόνιση σε σχέση με την επιθυμητή έξοδο.

Στο παρακάτω στιγμιότυπο έχουμε απεικονίσει τα ποσοστά τηλεθέασης που σημείωσε η εκπομπή *Amici di Maria de Filippi* συναρτήσει του αριθμού των μοναδικών χρηστών που δραστηριοποιήθηκαν στο Twitter δημοσιεύοντας "τιτιβίσματα" με την ετικέτα #amici, όπως έχουμε αναλύσει στα προηγούμενα. Παρακάτω θα βρείτε το ίδιο στιγμιότυπο, αλλά συμπεριλαμβάνει πλέον και την αρχή των αξόνων για να μπορούμε να αντιληφθούμε πλήρως την κατανομή των σημείων σε σχέση με το επίπεδο.



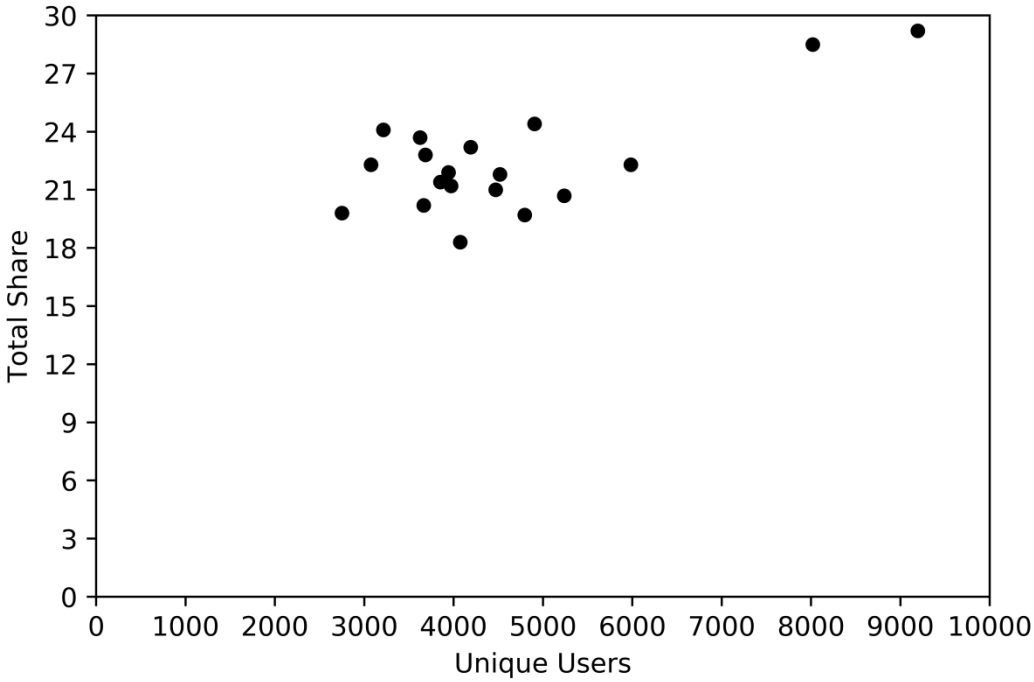
Εικόνα 5.7: Σχέση ποσοστού τηλεθέασης και αριθμού μεμονωμένων χρηστών Twitter που δημοσιεύουν σχετικά με το *Amici di Maria de Filippi*

Η πρώτη παρατήρηση είναι ότι τα δεδομένα έχουν μεν μια συνάφεια και μια συσχέτιση, ωστόσο υπάρχει σχετικά μεγάλη διασπορά από μια υποτιθέμενη ευθεία που τα αντιστοιχίζει γραμμικά. Αυτό μας δείχνει ότι έχει νόημα η ανάλυση που επιχειρούμε σε

αυτή την υποενότητα, μιας και τα δεδομένα μας δεν έχουν τόσο έντονο το στοιχείο της γραμμικότητας ή έστω της "πολυωνυμικότητας" που υποθέταμε μέχρι τώρα.

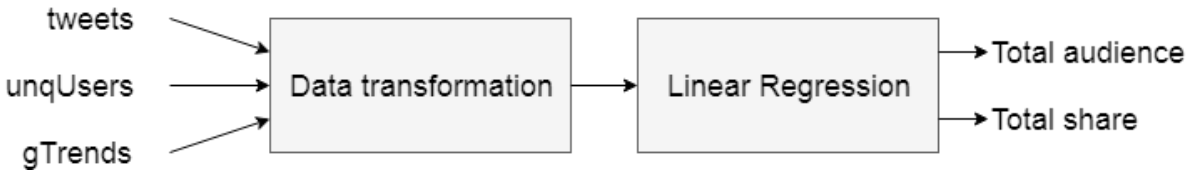
Μπορούμε να παρατηρήσουμε ακόμα ότι η συνάρτηση που ενδεχομένως προσαρμόζεται καλύτερα στα δεδομένα μας είναι κοίλη, σχηματίζει δηλαδή ένα "γόνατο" όσο αυξάνονται οι τιμές και φαίνεται να συγκλίνει σε κάποια τιμή. Το συμπέρασμα αυτό είναι απόλυτα λογικό, καθώς η μεταβλητή του άξονα y είναι φραγμένη ως ποσοστό. Το θεωρητικό φράγμα είναι το 100, αλλά πρακτικά αυτό βρίσκεται πολύ χαμηλότερα, διότι μιλάμε για ποσοστό τηλεθέασης που είναι στατιστικά και ουσιαστικά απίθανο να "τερματίσει".

Σε επίπεδο γραφήματος, κοίλες είναι οι συναρτήσεις που έχουν αυτό το "γόνατο" και στρέφονται προς τα κάτω, δηλαδή συγκλίνουν σε κάποια ευθεία παράλληλα με τον x άξονα. Αλγεβρικά, τέτοιες συναρτήσεις είναι όσες έχουν δεύτερη παράγωγο αρνητική σε όλο το πεδίο ορισμού τους. Επηρεαζόμενοι και από τα δύο στιγμιότυπα, μας έρχονται δύο τέτοιες συναρτήσεις στο νου: η λογαριθμική (log) και η τετραγωνική ρίζα (sqrt). Κατά προφανή τρόπο αποδεικνύεται και αλγεβρικά η διαίσθηση ότι είναι κοίλες συναρτήσεις. Αυτές οι δύο συναρτήσεις είναι που θα μελετηθούν στην παρούσα υποενότητα.

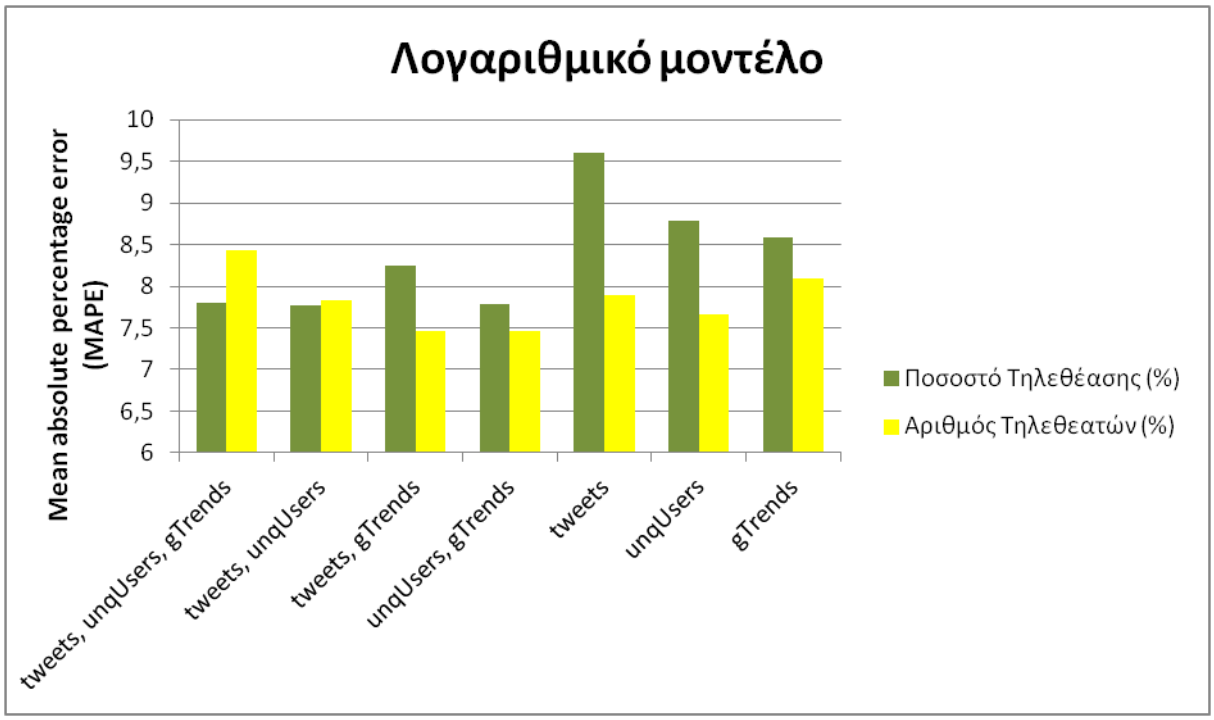


Εικόνα 5.8: Το γράφημα της Εικόνας 5.7 σε σχέση και με την αρχή των αξόνων

Στην Εικόνα 5.8 είναι ακόμα πιο ξεκάθαρη η ανάγκη μετατροπής των δεδομένων με κάποια κοίλη συνάρτηση, καθώς η διασπορά στο χώρο φαίνεται μικρότερη και είναι σαφής η μη γραμμικότητα των δεδομένων. Προσθέτουμε λοιπόν στο μεγάλο dataset έξι ακόμα στήλες, που είναι οι μετασχηματισμοί των δεδομένων από το Twitter σε λογαριθμική κλίμακα και στην τετραγωνική ρίζα, για να συνεχίσουμε τα πειράματα.

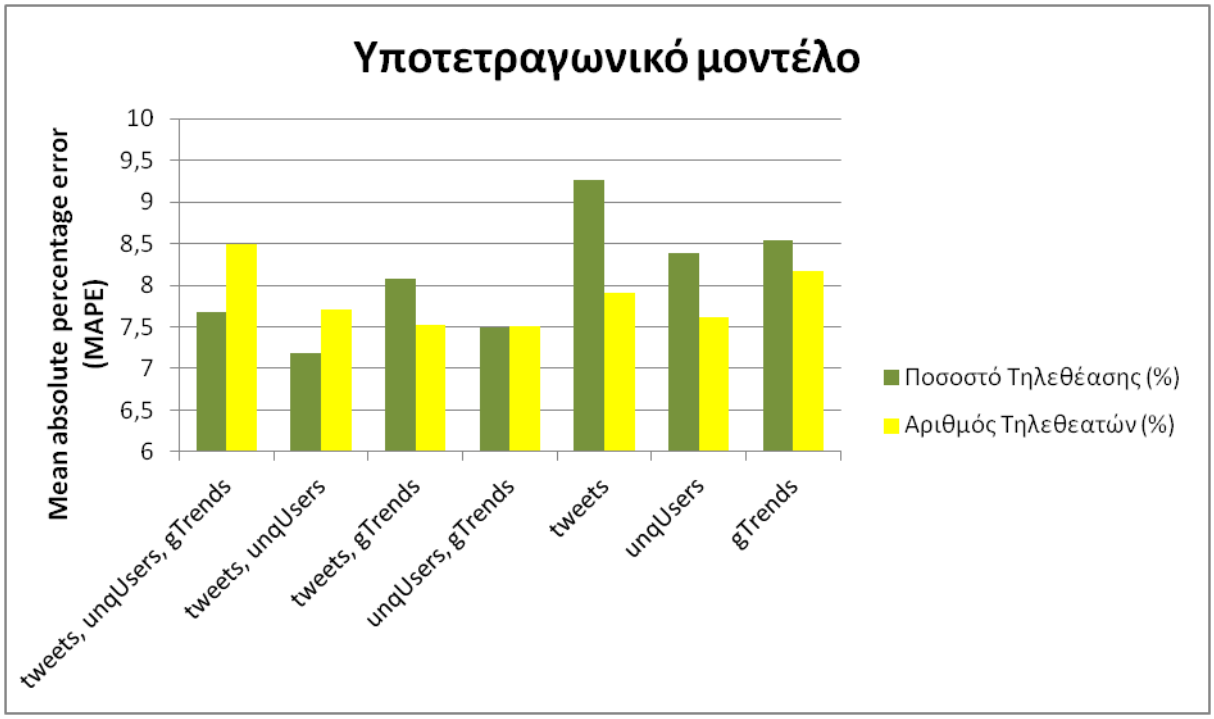


Εικόνα 5.9: Διάγραμμα λειτουργίας μη γραμμικού μοντέλου με μετατροπή των δεδομένων



Εικόνα 5.10: Μέσο απόλυτο ποσοστιαίο σφάλμα για όλους τους δυνατούς συνδυασμούς εισόδων μετά τη μετατροπή στη λογαριθμική κλίμακα - Γραμμικό μοντέλο

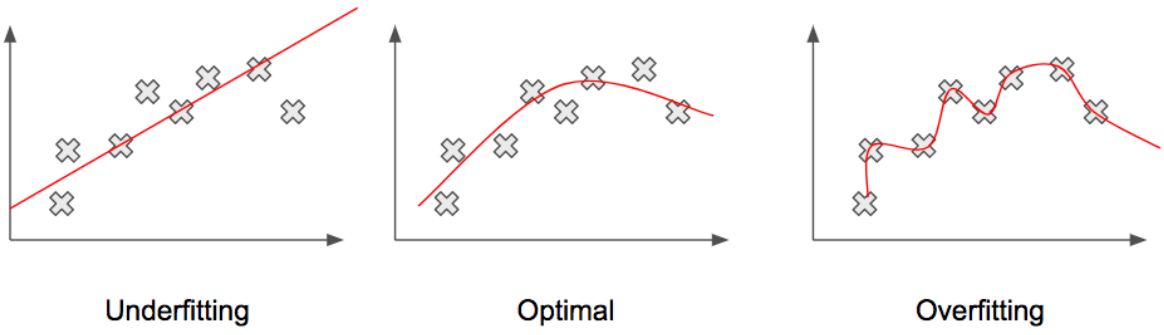
Παρατηρούμε ότι δεν επιτυγχάνουμε κάποια βελτίωση με τη λογαριθμική κλίμακα. Υπάρχει μόνο μια μικρή βελτίωση στην εκτίμηση από τα ατομικά χαρακτηριστικά, η οποία ωστόσο δεν είναι το βασικό αντικείμενο μας. Αντιθέτως, στα πολυδιάστατα μοντέλα υπάρχει μια μικρή αύξηση στο μέσο σφάλμα, γεγονός που μας υποδεικνύει ότι η λογαριθμική δεν είναι η συνάρτηση που ψάχναμε για να μετασχηματίσουμε τα δεδομένα μας. Σε αυτό το σημείο απεικονίζουμε και την αντίστοιχη ανάλυση για την τετραγωνική ρίζα.



Εικόνα 5.11: Μέσο απόλυτο ποσοστιαίο σφάλμα για όλους τους δυνατούς συνδυασμούς εισόδων μετά τη μετατροπή σε τετραγωνική ρίζα - Γραμμικό μοντέλο

Παρά το γεγονός ότι σε θεωρητικό επίπεδο ξεκινάμε από καλύτερη αφετηρία, προσπαθώντας να προσομοιώσουμε όσο καλύτερα μπορούμε τις αρχικές μας μεταβλητές, ακόμα και μετασχηματίζοντάς τις βάσει της κατανομής τους στο χώρο, το τελικό αποτέλεσμα δεν απέδωσε τα αναμενόμενα. Εμφανίζει παρόμοιες επιδόσεις με το γραμμικό μοντέλο και υστερεί από το πολυωνυμικό που είναι σαφώς ανώτερο από τα προαναφερθέντα.

Στη σύγκριση των δύο μετασχηματισμών, ο λογαριθμικός απέδωσε ελάχιστα καλύτερα στην εκτίμηση του αριθμού των τηλεθεατών, ενώ η ρίζα ήταν κατά τι ανώτερη στο ποσοστό τηλεθέασης. Αν εφαρμόσουμε στα δύο μετασχηματισμένα datasets και από το πολυωνυμικό μοντέλο, παρατηρούμε και στις δύο περιπτώσεις συναρτήσεων ότι το πολυωνυμικό μοντέλο δε βελτιώνει καθόλου την απόδοση και η καλύτερη επίδοση ταυτίζεται με το γραμμικό μοντέλο για όλους τους συνδυασμούς. Μόνο στα ατομικά χαρακτηριστικά βελτιώνει την απόδοση του γραμμικού μοντέλου και μάλιστα μόνο για τις δύο κατηγορίες δεδομένων του Twitter. Τα δεδομένα από το Google Trends επιμένουν γραμμικά.

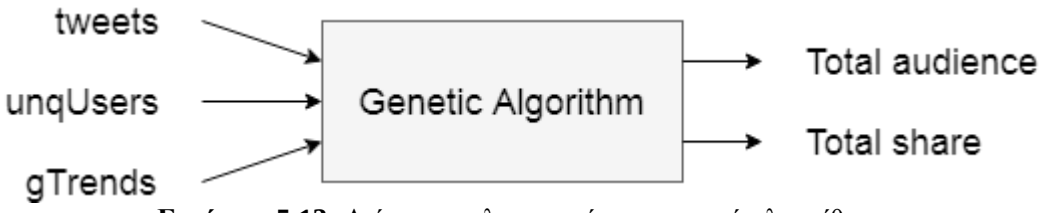


Εικόνα 5.12: Παράδειγμα λειτουργίας μοντέλου από το γενικό στο ειδικό

Η βελτίωση που εμφανίζεται στα ατομικά χαρακτηριστικά και παρουσιάζεται σε σχετικά μεγάλες δυνάμεις -από τρίτη και άνω- δεν είναι τίποτα άλλο από το φαινόμενο του overfitting. Όταν εφαρμόζουμε όλο και πιο περίπλοκα μοντέλα για να εκτιμήσουμε μια μεταβλητή, είναι πολύ λογικό ο αλγόριθμος να "υπερπροσαρμοστεί" στα δεδομένα και να θέλει να κάνει όσο πιο περισσότερες ακριβείς εκτιμήσεις μπορεί. Το πρόβλημα έγκειται στο ότι πλέον παύει να μαθαίνει από τα δεδομένα και προσπαθεί να τα απομνημονεύσει και να αναπαράξει.

Ένα τέτοιο μοντέλο δεν έχει πρακτική χρησιμότητα καθώς δεν μπορεί να γενικευτεί και να χρησιμοποιηθεί σε διαφορετικά δεδομένα, μιας και έχει προσαρμοστεί υπερβολικά στα υπάρχοντα. Αντίστοιχα και το underfitting είναι πολύ σημαντικό πρόβλημα, μιας και υποδεικνύει ένα μοντέλο που είναι πολύ γενικό και δεν έχει καταφέρει να προσαρμοστεί στα δεδομένα και εν γένει στο πρόβλημα. Στην Εικόνα 5.12 γίνεται οπτικοποίηση της παραπάνω εξήγησης σε ένα τυχαίο dataset.

5.3.3 Προσέγγιση με γενετικούς αλγορίθμους



Εικόνα 5.13: Διάγραμμα λειτουργίας γενετικού αλγορίθμου

Ένας γενετικός αλγόριθμος είναι μια στοχαστική διαδικασία μέσα από την οποία παράγεται μια προσεγγιστική απάντηση για το προς επίλυση πρόβλημα. Στην προκειμένη περίπτωση προσομοιώνεται η λειτουργία της γραμμικής παλινδρόμησης. Το όνομα αυτής της κλάσης αλγορίθμων προέρχεται από το γεγονός ότι επιχειρούν ευριστική αναζήτηση μοντελοποιώντας τον τρόπο με τον οποίο επιβιώνουν και εξελίσσονται γενετικά τα είδη, σύμφωνα με την επικρατούσα θεωρία του Δαρβίνου. Ένας γενετικός αλγόριθμος αποτελείται από τα εξής δομικά συστατικά:

- Συνθήκη τερματισμού: Κάποιο κριτήριο ικανό να μας υποδείξει πότε έχουμε μια αποδεκτή λύση. Μπορεί να περιέχει μεταξύ άλλων όριο στον αριθμό των γενεών που θα παραχθούν και κάποιο φράγμα/κατώφλι για το σφάλμα ή τη μετρική απόδοσης του αλγορίθμου.
- Επιλογή: Η διαδικασία προσδιορισμού του πληθυσμού από τον οποίο θα παραχθεί η επόμενη γενιά γονιδιωμάτων.
- Παραλλαγή: Η διαδικασία επιβολής αλλαγών στα επιλεχθέντα άτομα για να διαφοροποιηθεί η συμπεριφορά τους. Συνήθεις τεχνικές είναι η μετάλλαξη και η διασταύρωση.

Πρακτικά αυτά τα τρία στοιχεία είναι ότι χρειάζεται ένας γενετικός αλγόριθμος. Πρώτα, γίνεται μια αρχικοποίηση του πληθυσμού σε τυχαίες τιμές, έπειτα αξιολογείται ως προς την υγεία και μετά μπαίνουμε στο κυρίως σώμα του (επαναληπτικού) αλγορίθμου. Επιλέγεται ένα τμήμα του πληθυσμού με βάση την ορισμένη διαδικασία, μεταβάλλεται με τις προαναφερθείσες τεχνικές και έτσι παράγονται οι απόγονοι. Ακολουθεί αξιολόγηση του πληθυσμού και έλεγχος της συνθήκης τερματισμού. Αν δεν έχει ικανοποιηθεί στο σύνολό της, περνάμε στην επόμενη γενιά.

Σε πρώτη φάση φτιάχνουμε δύο συναρτήσεις για την τυχαία αρχικοποίηση των ατόμων και του πρώτου πληθυσμού. Έπειτα φτιάχνουμε μια συνάρτηση που ελέγχει την "υγεία" ενός ατόμου, δηλαδή πόσο ταιριάζει στα δεδομένα και την αναμενόμενη έξοδο. Σε αυτή τη συνάρτηση υπολογίζονται όλες οι μετρικές απόδοσης και σφαλμάτων. Έπειτα αυτή η συνάρτηση δίνει τα παραπάνω δεδομένα στη συνάρτηση αξιολόγησης, η οποία επιλέγει τα καταλληλότερα άτομα για να παραλλαχθούν. Η λογική που χρησιμοποιείται για την επιλογή είναι η ελαχιστοποίηση των σφαλμάτων.

Όπως είπαμε, τα δεδομένα μετασχηματίζονται ώστε να αναπαρίστανται με δυαδικές συμβολοσειρές. Οι τεχνικές μεταβολής που επιλέγονται είναι αυτές της μετάλλαξης και της διασταύρωσης. Στη μετάλλαξη επιλέγεται αρχικά ένας αριθμός ατόμων και έπειτα σε αυτά ένας αριθμός γονιδίων που αντιστρέφονται. Κατά τη διασταύρωση επιλέγονται τα δύο καλύτερα άτομα και εκτελείται ο αλγόριθμος της ομοιόμορφης διασταύρωσης με ακριβώς τα μισά γονίδια από κάθε άτομο. Στη συνέχεια επανεκτιμάται η υγεία του συστήματος και ελέγχεται το κριτήριο τερματισμού. Αν δεν ικανοποιούνται οι απαιτήσεις που θέσαμε και υπάρχουν ακόμα διαθέσιμες γενεές, ο αλγόριθμος συνεχίζει την ευριστική αναζήτηση για τη λύση.

Το ξεχωριστό στοιχείο που έχει αυτή την κλάση αλγορίθμων είναι ότι λειτουργούν τελείως στοχαστικά. Αυτό διευκολύνει την αποφυγή τοπικών ελαχίστων που μπορεί να υπάρχουν στη συνάρτηση σφαλμάτων. Επίσης το "τρικ" με τους μεγάλους πληθυσμούς από τους οποίους επιλέγονται τα ισχυρότερα άτομα για τη διαδικασία της εξέλιξης εξασφαλίζει σε μεγάλο βαθμό ότι δε θα χαθεί μια καλή λύση από αυτή την (ψευδο)τυχαία επιλογή ατόμων.

Στον πίνακα της επόμενης σελίδας μπορείτε να δείτε τα αποτελέσματα για το μέσο ποσοστιαίο σφάλμα για διάφορες εκτελέσεις με διαφορετικές παραμέτρους από τις παραπάνω, όπου φαίνεται η δυνατότητα των γενετικών αλγορίθμων να παράγουν μια πολύ καλή λύση χωρίς να έχουν τη μαθητική ικανότητα άλλων αλγορίθμων μηχανικής μάθησης, αλλά με εξαιρετικές ευριστικές μεθόδους.

Εικόνα 5.14: Μέσο απόλυτο ποσοστιαίο σφάλμα γενετικού αλγορίθμου για διάφορους συνδυασμούς παραμέτρων εισόδου

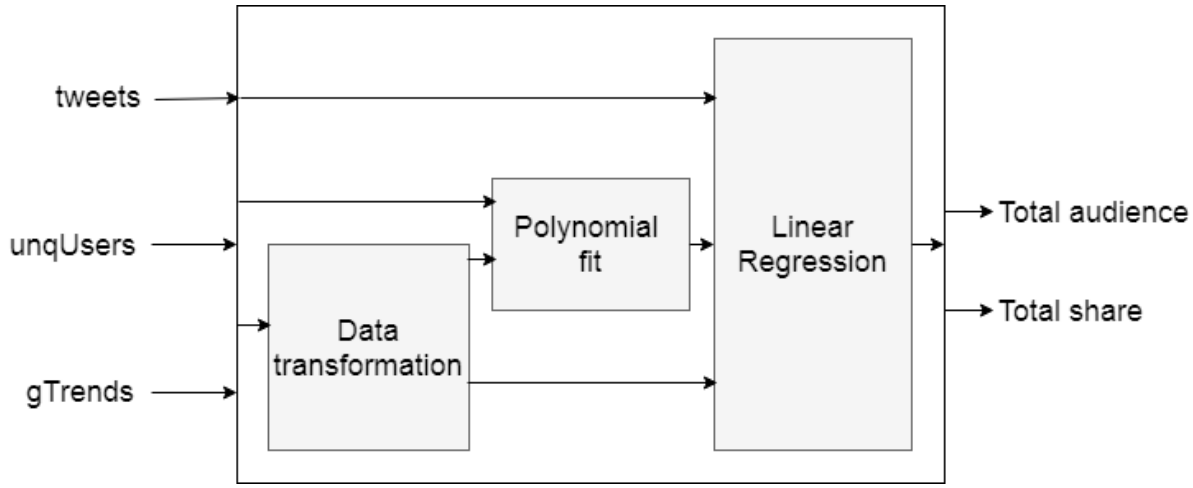
Δε στεκόμαστε περισσότερο στα αποτελέσματα, μιας και η μόνη τους χρησιμότητα είναι να επαληθεύσουν την αποδοτικότητα των μοντέλων που εφαρμόσαμε στις προηγούμενες υποενότητες και να ορίσουμε ενδεχομένως κάποια προσεγγιστικά κάτω φράγματα από τις χαμηλότερες τιμές του συγκεκριμένου πειράματος.

	MPE
0	7.989
1	7.379
2	8.868
3	7.711
4	8.189
5	8.279
6	9.484
7	9.300
8	7.216
9	7.184
10	7.521
11	7.642
12	7.400
13	7.768
14	8.553
15	8.858
16	6.747
17	7.768
18	8.211
19	9.611

Εικόνα 5.15: Μέσο απόλυτο ποσοστιαίο σφάλμα για διάφορες εκτελέσεις του υβριδικού μοντέλου

5.3.4 Υβριδικό μοντέλο

Το τελευταίο στάδιο των πειραμάτων είναι το λεγόμενο υβριδικό μοντέλο. Ουσιαστικά, προσπαθούμε να συνδυάσουμε τις διάφορες κατηγορίες δεδομένων που δημιουργήθηκαν τόσο στο στάδιο αναζήτησης του κατάλληλου χρονικού παραθύρου. Για να επιτευχθεί αυτό παίρνουμε όλες τις δυνατές δυάδες στηλών από το μεγάλο dataset στο οποίο έχουμε αποθηκεύσει κάθε μία μεταβλητή από όλα τα πειράματα και εκτελούμε γραμμική παλινδρόμηση με κάθε δυνατή δυάδα ως είσοδο. Στην προηγούμενη σελίδα παρουσιάζονται ενδεικτικά 20 από τις εξόδους του προγράμματος -πάντα ως μέσο ποσοστιαίο σφάλμα-, αφού ο αριθμός των συγκεκριμένων πειραμάτων είναι τριψήφιος και απαγορευτικός για οποιαδήποτε κομψή παρουσίαση του πλήρους αποτελέσματος.



Εικόνα 5.16: Διάγραμμα λειτουργίας υβριδικού μοντέλου. Επιλέγεται κάθε φορά μία από τις δυνατές διαδρομές μεταξύ των άκρων.

Αρκετά ενδιαφέρον είναι το γεγονός ότι κάποια από τα πιο αποδοτικά μοντέλα περιλαμβάνουν ως μεταβλητή εισόδου την κλάση 2 των μοναδικών χρηστών, δηλαδή μια κατηγορία που απορρίφθηκε από την επιλογή παραμέτρων παρότι μάλιστα απέδιδε ελαφρώς καλύτερα σε κάποια πειράματα. Αυτές οι δύο λέξεις (ελαφρώς, κάποια) είναι και το κλειδί της απόρριψης. Δεδομένου ότι δεν υπάρχει σαφής ένδειξη για διαρκή βέλτιστη απόδοση προτιμήθηκαν τα δεδομένα ενός απλούστερου άρα και γενικότερου μοντέλου. Τα ακριβή συμπεράσματα παρουσιάζονται στην επόμενη ενότητα.

5.4 Σύνοψη πειραματικών ευρημάτων

Στο πρώτο πείραμα επιχειρήθηκε μια ανάλυση συνάφειας μεταξύ ημερήσιων δεδομένων που προέρχονται από το Google Trends για την αναζήτηση όρων σχετικών με την εκπομπή *Amici di Maria de Filippi* και του πλήθους δημοσιεύσεων ανά ημέρα στο Twitter, ομαδοποιώντας τα δεδομένα ανά εξάμηνο για το έτος 2017. Για το πρώτο και το δεύτερο εξάμηνο σημειώθηκε συσχέτιση της τάξης του 0,89 και 0,82 αντίστοιχα. Τα παραπάνω αποτελέσματα αποδεικνύουν **ισχυρή γραμμική συσχέτιση** μεταξύ των δύο κατηγοριών δεδομένων. Αυτό πρακτικά σημαίνει ότι τα δεδομένα της πλατφόρμας της Google είναι αξιόπιστα και μπορούν να αποτελέσουν τόσο συμπλήρωμα όσο και εναλλακτική για πληθώρα δεδομένων από κοινωνικά δίκτυα σε ερευνητικές δραστηριότητες σχετικές με την εκτίμηση τηλεθέασης.

Το δεύτερο πείραμα είναι και το κυρίως θέμα της παρούσας εργασίας, καθώς περιλαμβάνει την αναζήτηση και μοντελοποίηση του κατάλληλου αλγορίθμου για την εκτίμηση της τηλεθέασης τηλεοπτικών εκπομπών βάσει δεδομένων κοινωνικών δικτύων. Αρχικά, χρειάστηκε να προσδιοριστούν τα χρονικά παράθυρα που θα χρησιμοποιηθούν για τις μετρήσεις, δεδομένου ότι τα επεισόδια της επιλεγμένης εκπομπής είναι εβδομαδιαία.

Όσον αφορά το Google Trends, είναι δεδομένη η αρχή και το τέλος της βδομάδας στην οποία αντιστοιχεί η αποδιδόμενη τιμή και δεν μπορεί να μετακυλήσει κατά 1-2 μέρες. Από τη στιγμή που η ημέρα εκπομπών (Κυριακή) πέφτει πάνω ακριβώς στο σύνορο, υπήρχαν δύο επιλογές: η εβδομάδα πριν + η ημέρα προβολής και η εβδομάδα ακριβώς μετά την προβολή. Τα σχετικά πειράματα υπέδειξαν **την εβδομάδα πριν και ανήμερα της προβολής της εκπομπής** ως την καταλληλότερη για την εκτίμηση της τηλεθέασης.

Το αντίστοιχο πείραμα για τα δεδομένα από το Twitter ήταν λίγο πιο σύνθετο μια και υπάρχουν ημερήσια δεδομένα από την πλατφόρμα, επομένως οι συνδυασμοί είναι πολύ

περισσότεροι. Μια απλή απεικόνιση των ημερήσιων δεδομένων σε μορφή χρονοσειράς υποδεικνύει το γεγονός ότι ο κύριος όγκος δημοσιεύσεων στο Twitter γίνεται την ημέρα της προβολής επεισοδίου, καθώς και την επόμενη αυτής. Συνεπώς, τα πειράματα περιορίζονται μεταξύ των δεδομένων από την ημέρα της προβολής και τον συνδυασμό και των 2 ημερών. Τα **δεδομένα της μίας ημέρας** ήταν ελαφρώς πιο αποδοτικά και παρόλο που δεν είχαν την υπεροχή σε όλα τα πειράματα, επιλέγονται **και για λόγους απλότητας και δυνατότητας γενίκευσης** του μοντέλου.

Έχοντας ρυθμίσει τις επιλογές χρονικού παραθύρου, εκτελούμε τα πειράματα για τους διαφορετικούς αλγόριθμους παλινδρόμησης για να εκτιμήσουμε το **ποσοστό τηλεθέασης** καθώς και το **πλήθος τηλεθεατών** κάθε επεισοδίου. Για το **γραμμικό μοντέλο** το μέσο απόλυτο ποσοστιαίο σφάλμα κυμάνθηκε μεταξύ **6,7-8,8%** και **7,5-8,5%** αντίστοιχα. Στο **πολυωνυμικό μοντέλο** οι τιμές βρέθηκαν μεταξύ **6,7-8,5%** και **6,9-8,5%** αντίστοιχα. Για τα μη γραμμικά δεδομένα τα σφάλματα υπολογίστηκαν μεταξύ **7,1-9,6%** και **7,6-8,5%** αντίστοιχα στη γραμμική προσέγγιση και για την πολυωνυμική μεταξύ **6,1-9,6%** και **6,4-8,5%** αντίστοιχα.

Τα αποτελέσματα υποδεικνύουν ότι το **γραμμικό μοντέλο 2 διαστάσεων** υπερτερεί στην πλειοψηφία των πειραμάτων και υστερεί για δέκατα στα υπόλοιπα. Συγκεκριμένα, οι δύο μεταβλητές που αναδεικνύονται ως οι αποδοτικότερες για την **εκτίμηση του συνολικού αριθμού τηλεθεατών της εκπομπής είναι τα δεδομένα από το Google Trends σε συνδυασμό με τον αριθμό μοναδικών χρηστών** που δημοσίευσαν tweet σχετικό με την εκπομπή την ημέρα προβολής. Για την **εκτίμηση του τηλεμεριδίου απέδωσε καλύτερα η δυάδα αριθμού μοναδικών χρηστών και αριθμού δημοσιεύσεων στο Twitter**, πάντα αναφερόμενοι αποκλειστικά στην ημέρα προβολής.

Τα παραπάνω αποτελέσματα επαληθεύτηκαν από το πείραμα της προσέγγισης μέσω γενετικού αλγορίθμου, όπου το μέσο ποσοστιαίο σφάλμα κυμάνθηκε μεταξύ **6,1** και **11%**, παρέχοντας τα αντίστοιχα φράγματα για τη βέλτιστη λειτουργία του μοντέλου.

6. Επίλογος

Στο τελευταίο κεφάλαιο της εργασίας συνοψίζουμε τα σημαντικότερα αποτελέσματα της εργασίας και αναφέρουμε τις μελλοντικές δυνατότητες που δίνει η παρούσα δουλειά, τόσο ως δείγμα των επερχομένων ερευνητικών δραστηριοτήτων του γράφοντος όσο κι ως προτάσεις προς άπαντες τους ενδιαφερόμενους για την ανάλυση συμπεριφοράς χρηστών μέσων κοινωνικής δικτύωσης γενικά και την ανάλυση ακροαματικότητας μέσω κοινωνικών δικτύων ειδικότερα.

6.1 Σύνοψη και συμπεράσματα

Είναι εμφανής η διαρκώς αυξανόμενη παρουσία των κοινωνικών δικτύων και ιδιαίτερα των μέσων κοινωνικής δικτύωσης στην καθημερινή μας ζωή τα τελευταία χρόνια, απόρροια της ευκολίας και της μαζικότητας που παρέχουν στο χρήστη σε τομείς όπως η επικοινωνία και η ενημέρωση. Το ξέσπασμα αυτό δεν αφήνει αδιάφορη την επιστημονική κοινότητα και αναζητούνται συνεχώς τρόποι ωφέλειας από αυτή την ανοιχτή και πλούσια πηγή δεδομένων ως προς τη μελέτη διαφόρων πτυχών της ανθρώπινης συμπεριφοράς, αλλά και εφαρμογές όπως το στοχευμένο μάρκετινγκ και η επιχειρησιακή ανάλυση. Η αξιοποίηση δεδομένων από κοινωνικές πλατφόρμες αποτελεί ένα σχετικά καινούριο επιστημονικό τομέα, ωστόσο η πληθώρα πληροφορίας που διαθέτει η πλατφόρμα μπορεί να αποτελέσει καταλύτη για την εξερεύνηση διαφόρων κοινωνικών φαινομένων.

Το αντικείμενο της παρούσης διπλωματικής εργασίας είναι η μελέτη και εξαγωγή συμπερασμάτων για το ενδιαφέρον των χρηστών για διάφορες τηλεοπτικές εκπομπές βασιζόμενοι στη δραστηριότητα τους στα κοινωνικά δίκτυα. Συγκεκριμένα, αντλούνται στοιχεία από τις πλατφόρμες Twitter και Google Trends. Έχουν πραγματοποιηθεί αρκετές έρευνες τα τελευταία χρόνια βασισμένες σε δεδομένα από το Twitter για ανάλυση σχετική με ακροαματικότητα. Η διαφοροποίηση σε σχέση με άλλες προσπάθειες έγκειται στην καινοτόμο αξιοποίηση της δεύτερης πλατφόρμας.

Επικεντρωθήκαμε στο συνδυασμό των δεδομένων από το Twitter και το Google Trends για την εκτίμηση της ακροαματικότητας τηλεοπτικών εκπομπών. Για τους σκοπούς της μελέτης μας χρησιμοποιούνται δεδομένα από το Ιταλικό reality show "Amici di Maria de Filippi". Η προσέγγιση είναι ανεξάρτητη της γλώσσας και μπορεί να γενικευτεί για οποιαδήποτε περιοχή παρουσιάζει σχετικό ενδιαφέρον. Για την μοντελοποίηση και αυτοματοποίηση χρησιμοποιείται η γλώσσα προγραμματισμού Python.

Πρώτος στόχος ήταν να εξετάσουμε κατά πόσον τα δεδομένα από το Google Trends είναι αξιόπιστα και προσφέρονται για τους σκοπούς της έρευνάς μας. Τα υψηλότατα ποσοστά συνάφειας της ημερήσιας σχετικής αναζήτησης στο Google Search Engine με τον αντίστοιχο αριθμό δημοσιεύσεων στο Twitter αποδεικνύουν τον παραπάνω ισχυρισμό και μας επιτρέπουν τη συνέχιση των πειραμάτων με την εν λόγω πληροφορία.

Έπειτα προχωρήσαμε στην αναζήτηση του καταλληλότερου μοντέλου για την εκτίμηση του αριθμού των τηλεθεατών κάποιου επεισοδίου και του αντίστοιχου μεριδίου τηλεθέασης με τη χρήση τεχνικών παλινδρόμησης. Μελετήθηκαν τα μοντέλα της γραμμικής, της πολυωνυμικής και της εν γένει μη γραμμικής παλινδρόμησης μέσω μετασχηματισμού των δεδομένων εισόδου, καθώς και ένα υβρίδιο των προηγουμένων με μεικτά δεδομένα ως είσοδο στο γραμμικό μοντέλο. Προτού εισαχθούν τα δεδομένα στα μοντέλα, εξερευνήθηκαν διάφορες κατηγορίες παραμέτρων για την εύρεση των ιδανικών χρονικών παραθύρων λήψης δεδομένων. Μετρική αξιολόγησης των μοντέλων είναι το μέσο ποσοστιαίο σφάλμα.

Τα αποτελέσματα υποδεικνύουν ότι το γραμμικό μοντέλο 2 διαστάσεων υπερτερεί στην πλειοψηφία των πειραμάτων και υστερεί για δέκατα στα υπόλοιπα. Συγκεκριμένα, οι δύο μεταβλητές που αναδεικνύονται ως οι αποδοτικότερες για την εκτίμηση του συνολικού αριθμού τηλεθεατών της εκπομπής είναι τα δεδομένα από το Google Trends σε συνδυασμό με τον αριθμό μοναδικών χρηστών που δημοσίευσαν tweet σχετικό με την εκπομπή την ημέρα προβολής. Για την εκτίμηση του τηλεμεριδίου απέδωσε καλύτερα η δυάδα αριθμού μοναδικών χρηστών και αριθμού δημοσιεύσεων στο Twitter, πάντα αναφερόμενοι αποκλειστικά στην ημέρα προβολής.

Τόσο κατά την εφαρμογή των μη γραμμικών μοντέλων, όσο και του υβριδικού, παρατηρήθηκαν ελαφρώς χαμηλότερες τιμές σφάλματος σε κάποια μεμονωμένα πειράματα. Οι παραπάνω αποδόσεις οφείλονται στο φαινόμενο του overfitting και ουδόλως προσφέρονται ως χρήσιμη πληροφορία στην κατεύθυνση της γενίκευσης του παραπάνω μοντέλου. Το γραμμικό μοντέλο με δεδομένα μίας ημέρας είναι όχι μόνο το απλούστερο και γενικότερο μοντέλο που μπορούμε να δημιουργήσουμε, αλλά κρίθηκε και το πλέον αποτελεσματικό στη συντριπτική πλειοψηφία των πειραμάτων.

Το παραπάνω αποτέλεσμα και ο βέλτιστος χαρακτήρας της λύσης-μοντέλου αποτιμήθηκε επίσης με μια προσέγγιση της γραμμικής παλινδρόμησης από γενετικό αλγόριθμο. Η προσέγγιση αυτή χρησιμοποιήθηκε ως ένα εργαλείο γρήγορης -προσεγγιστικής- εκτίμησης των αποτελεσμάτων και εύρεσης φραγμάτων για το σφάλμα και επαλήθευσε πλήρως τα παραπάνω συμπεράσματα. Πιο συγκεκριμένα, το μέσο ποσοστιαίο σφάλμα κυμάνθηκε σε τιμές γύρω από το 7,5%. Για εκπομπές μέσου τηλεμεριδίου της τάξης του 20% αυτό σημαίνει σφάλμα περίπου 1,5 ποσοστιαίας μονάδας. Παρότι δεν είναι ιδανικό, είναι αποδεκτό δεδομένων δύο συνθηκών:

- Του μικρού δείγματος, μόλις 19 επεισοδίων
- Της ύπαρξης μιας έκτοπης τιμής, η οποία εμφανίζει σφάλμα της τάξης του 30% σε όλα τα πειράματα. Πρόκειται για επεισόδιο με εξαιρετικά υψηλή δραστηριότητα στο Twitter, η οποία δεν αποτυπώθηκε στην τηλεθέασή του. Πιθανές αιτίες είναι η ύπαρξη κάποιου αξιοπερίεργου συμβάντος σε ζωντανό χρόνο, το οποίο διαδόθηκε και σχολιάστηκε χωρίς να υπάρχει ο ανάλογος αριθμός "μαρτύρων" ή η ταυτόχρονη προβολή κάποιου έκτακτου μουσικού, αθλητικού ή άλλου γεγονότος, το οποίο επικράτησε σε τηλεμερίδιο.

Off-the-record εκτελέσεις που έγιναν εξαιρώντας τα δεδομένα του συγκεκριμένου επεισοδίου έριξαν το σφάλμα στα επίπεδα του 5%, ήτοι μικρότερο της μίας ποσοστιαίας μονάδας.

Κατά τη διάρκεια των πειραμάτων επιχειρήθηκε η εξαγωγή δημογραφικών στοιχείων μέσα από την εκτίμηση τηλεμεριδίου και αριθμού θεατών για συγκεκριμένες ηλικιακές ομάδες και με διάκριση φύλου, χωρίς ωστόσο να προκύπτει κάποιο σαφές συμπέρασμα από τα αποτελέσματα.

6.2 Μελλοντικές επεκτάσεις

Η παρούσα μελέτη δίνει χώρο για πολλές πιθανές επεκτάσεις. Κάποιες από αυτές τις επεκτάσεις χρησιμεύουν στην περαιτέρω αξιολόγηση και βελτίωση της παρούσας προσέγγισης εκτίμησης ακροαματικότητας τηλεοπτικών εκπομπών και άλλες έχουν άλλους στόχους, όπως την ανίχνευση δημογραφικών χαρακτηριστικών των χρηστών κοινωνικών δικτύων.

Ένας προφανής τρόπος αξιολόγησης της παρούσας προσέγγισης είναι η χρησιμοποίηση μιας πολύ μεγαλύτερης βάσης δεδομένων τόσο σε αριθμό επεισοδίων ή/και

εκπομπών όσο και σε δεδομένα από Twitter και Google Trends, για να γίνει βέβαιο ότι η απόδοση που καταγράφηκε στην παρούσα μελέτη μπορεί να γενικευτεί και μάλιστα με άκρως πιθανή βελτίωση λόγω του μεγαλύτερου μεγέθους πληροφορίας που θα έχουμε στη διάθεση μας για τα επόμενα πειράματα.

Μια εναλλακτική πρόταση στην ίδια ερευνητική περιοχή είναι η χρήση διαφορετικών μεθόδων μηχανικής μάθησης όπως τα δέντρα αποφάσεων ή διαφορετικές ρυθμίσεις σε επίπεδο παραμέτρων. όπως η χρήση των προηγούμενων τιμών ως κάποιας μορφής μεροληψία. Ιδιαίτερο ενδιαφέρον σε αυτόν τον τομέα παρουσιάζει η επιδραστικότητα των δραστήριων χρηστών, η οποία μπορεί να μετρηθεί από παράγοντες όπως οι ακόλουθοι.

Φυσικά, ανοιχτό πεδίο έρευνας αποτελεί η εξαγωγή δημογραφικών στοιχείων από τη δραστηριότητα των χρηστών, όπως η ηλικία, το φύλο και ενδεχομένως ο τόπος διαμονής.

Βιβλιογραφία

About Twitter - https://about.twitter.com/

Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B (2012) Twitter improves seasonal influenza prediction. HEALTHINF, In, pp 61–70

Aldrich, J. (1998). "Doing Least Squares: Perspectives from Gauss and Yule". International Statistical Review. 66 (1): pp. 61–81.

Arias, M., Arratia, A. & Xuriguera, R., "Forecasting with twitter data, special issue on social web mining." ACM Transactions on Intelligent Systems and Technology (TIST), 51, 2013.

Askitas N., Zimmermann K.F., Google econometrics and unemployment forecasting, Applied Economics Quarterly, 55 (2) (2009), pp. 107-120

Asur S, Huberman BA (2010) Predicting the future with social media. CoRR abs/1003.5699. http://arxiv.org/abs/1003.5699

Bethea, R. M.; Duran, B. S.; Boullion, T. L. (1985). "Statistical Methods for Engineers and Scientists.", New York: Marcel Dekker. ISBN 0-8247-7227-X.

Bollen J, Mao H, Zeng XJ (2011) Twitter mood predicts the stock market. Journal of computational Science 2(1) CoRR abs/1010.3003.

Botta F, Moat HS, Preis T (2015) Quantifying crowd size with mobile phone and twitter data. R Soc open sci 2:150162.

Bühlmann, Peter; van de Geer, Sara (2011). "Statistics for High-Dimensional Data: Methods, Theory and Applications." Springer. ISBN 9783642201929.

Bulut, L. & Dogan, C. (2018). Google Trends and Structural Exchange Rate Models for Turkish Lira–US Dollar Exchange Rate. Review of Middle East Economics and Finance, 14(2).

Cawley, Gavin C.; Talbot, Nicola L. C. (2010). "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", 11. Journal of Machine Learning Research: 2079–2107.

Chauhan A, Kummamuru K, Toshniwal D (2016) Prediction of places of visit using tweets. Knowl Inf Syst: 1–22

Chiang, C.L (2003), "Statistical methods of analysis, World Scientific", ISBN 981-238-310-7, p. 274

Choi H., Varian H. Predicting the present with google trends Econ. Rec., 88 (2012), pp. 2-9

Christensen, Ronald (2002). "Plane Answers to Complex Questions: The Theory of Linear Models" (Third ed.). New York: Springer. ISBN 0-387-95361-2.

Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). "Applied multiple regression/correlation analysis for the behavioral sciences." (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates

Crisci A., et. al, "Predicting TV programme audience by using twitter based metrics", Multimedia Tools and Applications Journal, Vol. 77, No. 10, pp. 12203–12232, May 2018.

Fan, Jianqing (1996). "Local Polynomial Modelling and Its Applications: From linear regression to nonlinear regression". Monographs on Statistics and Applied Probability. Chapman & Hall/CRC. ISBN 0-412-98321-4.

Genetic Algorithm for Linear Regression

Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S. and Brilliant L., "Detecting influenza epidemics using search engine query data", Nature, Vol . 457, pp. 1012-1014, Feb. 2009.

Goel, Sharad et al "Predicting consumer behavior with Web search." Proceedings of the National Academy of Sciences 107.41 (2010): 17486-17490.

Goldberg, David (1989). "Genetic Algorithms in Search, Optimization and Machine Learning." Reading, MA: Addison-Wesley Professional. ISBN 978-0201157673.

Google Trends Help - https://support.google.com/trends/

Grasso V, Crisci A, Nesi P, Pantaleo G, Zaza I, Gozzini B. Public crowd-sensing of heat-waves by social media data. 16th EMS annual meeting & 11th European conference on applied climatology (ECAC), 12–16September 2016 | Trieste, Italy, CE2/AM3, Delivery and communication of impact based forecasts and risk based warnings

Grasso V, Zaza I, Zabini F, Pantaleo G, Nesi P, Crisci A (2016) Weather events identification in social media streams: tools to detect their evidence in twitter. PeerJ preprints 4:e2241v1.

Hassani, H. & Silva, E.S., "Forecasting with big data: a review.", Annals of Data Science, 2(1), pp. 5–19, 2015.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2009). "The Elements of Statistical Learning" (second ed.). Springer-Verlag. ISBN 978-0-387-84858-7.

Hsieh W-T et al (2013) Predicting tv audience rating with social media. Sixth International Joint Conference on Natural Language Processing

https://www.scipy.org/

Jun, Seung-Pyo & Yoo, Hyoung Sun & Choi, San, "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications", Technological Forecasting and Social Change, Elsevier, vol. 130(C), pages 69-87, 2018.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S., "Social media? Get serious! Understanding the functional building blocks of social media.", Business Horizons, 54(3), pp.241-251, 2011.

Kim, Sungil and Heeyoung Kim (2016). "A new metric of absolute percentage error for intermittent demand forecasts." International Journal of Forecasting, volume 32 issue 3, pp. 669-679.

Kohavi, Ron, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", International Joint Conference on Artificial Intelligence, 1995

Koza, John (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA: MIT Press. ISBN 978-0262111706.

Leskovec J (2011) Social media analytics: tracking, modeling and predicting the flow of information through networks. Proceedings of the 20th international conference companion on world wide web. ACM

Lu Y, Kruger R, Thom D, Wang F, Koch S, Ertl T, Maciejewski R. Integrating predictive analytics and social media. In Visual Analytics Science and Technology (VAST), 2014 I.E. conference on 2014 Oct 25. IEEE, p 193–202

Machine Learning Crash Course by Google.

Madlberger L, Almansour A. Predictions based on Twitter—A critical view on the research process. InData and Software Engineering (ICODSE), 2014 International Conference on 2014 Nov 26. IEEE, p 1–6

Makridakis, Spyros. "Accuracy measures: theoretical and practical concerns." International Journal of Forecasting 9.4 (1993): pp. 527-529

Mathieu Rouaud, 2013: "Probability, Statistics and Estimation" Chapter 2: Linear Regression, Linear Regression with Error Bars and Nonlinear Regression.

Mavragani A., Tsagarakis K.P., YES or NO: Predicting the 2015 GReferendum results using Google Trends, Technol. Forecast. Soc. Chang., 109 (2016), pp. 1-5

Metrics To Evaluate Machine Learning Algorithms in Python

Mishne G, Glance N (2006) Predicting movie sales from blogger sentiment. AAAI 2006 spring symposium on computational approaches to Analysing weblogs

Molteni L. and Ponce De Leon J., "Forecasting with twitter data: an application to Usa Tv series audience", Int. Journal of Design & Nature and Ecodynamics", Vol. 11, No. 3, pp. 220–229, Jul. 2016.

Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2018 (in millions) - Statista

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proc. of 4th ICWSM. AAAI Press, p 122–129

Official website - Google Trends

Official website - Twitter

Paul MJ, Dredze M (2011) You are what you tweet: Analysing twitter for public health. Proc. of ICWSM, In

Pedhazur, Elazar J (1982). "Multiple regression in behavioral research: Explanation and prediction" (2nd ed.). New York: Holt, Rinehart and Winston. ISBN 0-03-041760-0.

Project Jupyter

Python 2D plotting library - Matplotlib

Reddy ASS, Kasat P, Jain A (2012) Box-office opening prediction of movies based on hype analysis through data mining. Int J Comput Appl 56(1)

Regression in Python - scikit-learn

Russell, Stuart J.; Norvig, Peter (2010). "Artificial Intelligence: A Modern Approach" (Third ed.). Prentice Hall. ISBN 9780136042594.

Schootman M, Toor A, Cavazos-Rehg P, et al The utility of Google Trends data to examine interest in cancer screening BMJ Open 2015;5:e006678.

Scientific Computing in Python - NumPy

Seber, G. A. F.; Wild, C. J. (1989). "Nonlinear Regression.", New York: John Wiley and Sons. ISBN 0471617601.

Sen A., Srivastava M., "Regression Analysis — Theory, Methods, and Applications", Springer-Verlag, Berlin, 2011

Sereday S. and Cui J., "Using machine learning to predict future tv ratings", Data Science, Nielsen, Vol. 1, No. 3, pp. 3-12, Feb. 2017.

Sikdar S, Adali S, Amin M, Abdelzaher T, Chan KL, Cho JH, Kang B, O'Donovan J. Finding true and credible information on Twitter. In Information Fusion (FUSION), 2014 17th International Conference on 2014 Jul 7. IEEE, p 1–8

Sinha S, Dyer C, Gimpel K, Smith NA. Predicting the NFL Using Twitter. arXiv:1310.6998v1 [cs.SI] 25 Oct 2013

Sitaram A, Huberman BA (2010) Predicting the future with social media. In Social Computing Lab, HP Labs, Palo Alto

Sommerdijk B, Sanders E, van den Bosh A. Can Tweets Predict TV Ratings? The International Conference on Language Resources and Evaluation is organised by ELRA biennially with the support of institutions and organisations involved in HLT

Tofallis (2015). "A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation", Journal of the Operational Research Society, 66(8), pp. 1352-1362

Tumasjan A, Sprenger T, Sandner PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proc. of 4th ICWSM. AAAI Press, p 178–185

Varma, Sudhir; Simon, Richard (2006). "Bias in error estimation when using cross-validation for model selection". BMC Bioinformatics. 7: 91.

Wakamiya S, Lee R, Sumiya K (2011) Towards better TV viewing rates: exploiting crowd's media life logs over twitter for TV rating. Proceedings of the 5th international conference on ubiquitous information management and communication. ACM

Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from twitter posts. In: Social computing. Behavioural - Cultural Modeling and Prediction. Springer, Berlin Heidelberg, pp 231–238

Willmott, C. J.; Matsuura, K. (January 2006). "On the use of dimentioned measures of error to evaluate the performance of spatial interpolators". International Journal of Geographical Information Science. 20: pp. 89–102.

Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2, ISBN 9789812834119

YangJing Long, "Human age estimation by metric learning for regression problems", Proc. International Conference on Computer Analysis of Images and Patterns: pp.74–82, 2009