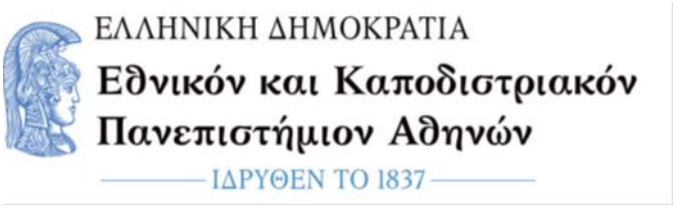


ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΠΟΛΙΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Π.Μ.Σ.: ΠΛΗΡΟΦΟΡΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του/της Άγγελου Ηλία Ιωαννίδη

Α.Μ.: 71151122200010

Ανάλυση Συναισθήματος Ποδοσφαιρικών Tweets με RoBERTa: Συσχέτιση Υποστήριξης Οπαδών και Απόδοσης Ομάδων

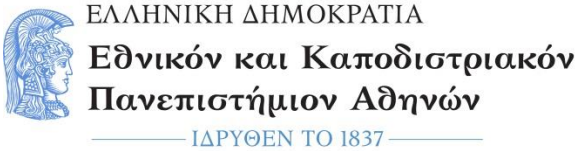
Επιβλέποντες:

α) Χριστίνα Αλεξανδρή

β) Μανώλης Κουμπαράκης

γ) Γίαννης Παναγάκης

Αθήνα, Φεβρουάριος 2024



«Ανάλυση Συναισθήματος Ποδοσφαιρικών Tweets με RoBERTa: Συσχέτιση Υποστήριξης Οπαδών και Απόδοσης Ομάδων»

«ΑΓΓΕΛΟΣ ΙΩΑΝΝΙΔΗΣ»

Επιβλέπων Καθηγητής:
«ΧΡΙΣΤΙΝΑ ΑΛΕΞΑΝΔΡΗ»

Αθήνα, «Φεβρουάριος» «2025»

Η παρούσα διπλωματική εργασία αποτελεί έργο προσωπικής μου προσπάθειας. Για να ολοκληρωθεί και να φτάσει στο επιθυμητό αυτό σημείο απαιτήθηκαν ώρες μελέτης, συγκέντρωσης και συλλογής πληροφοριών.

Ευχαριστώ θερμά όλους όσους με βοήθησαν καθ' όλη την περίοδο εκπόνησης και συγγραφής, δίνοντάς μου κουράγιο και στήριξη. Επίσης, ευχαριστώ τον επιβλέποντα καθηγητή μου για τις πολύτιμες συμβουλές, τις συστάσεις και τις καθοδηγητικές γραμμές που μου έδωσε.

Τέλος, ευχαριστώ την εξεταστική επιτροπή που μου κάνει την τιμή να αξιολογήσει την εργασία μου.

«Ευχαριστίες ή Αφιέρωση»

Περίληψη

Η παρούσα μελέτη αφορά την Ανάλυση Συναισθήματος (Sentiment Analysis) στις αναρτήσεις χρηστών σε μέσα κοινωνικής δικτύωσης (Social Media) στο πεδίο του Ποδοσφαίρου (Football, USA: Soccer). Πρόκειται για αναρτήσεις οπαδών και ποδοσφαιρόφιλων στην πλατφόρμα κοινωνικής δικύωσης «Χ» - πρώην «Twitter» μέγεθος 1.000.000 με δεδομένα και ομάδες από το Πρωτάθλημα English Premier Leauge («Αγγλικό Ποδόσφαιρο»).

Η Ανάλυση Συναισθήματος (Sentiment Analysis) αποτελεί σχετικά πρόσφατο πεδίο εφαρμογών της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP) και της Υπολογιστικής Γλωσσολογίας (Computational Linguistics). Μολονότι η Ανάλυση Συναισθήματος συνήθως εφαρμόζεται σε δεδομένα χρηστών ή/και καταναλωτών προϊόντων και υπηρεσιών, τα τελευταία χρόνια εφαρμόζεται και σε δεδομένα πιο πολύπλοκου περιεχομένου, ακόμη και στον τομέα της Δημοσιογραφίας και της Πολιτικής. Τα δεδομένα από το πεδίο του Ποδοσφαίρου και των ποδοσφαιρόφιλων ή/και οπαδών ποδοσφαιρικών ομάδων μπορούν να χαρακτηριστούν ως σύνθετου και πολύπλοκου περιεχομένου, κυρίως επειδή δεν περιορίζονται σε προκαθορισμένο λεξιολόγιο και ορολογία και επειδή συχνά περιέχουν αρκετά κοινωνιο-γλωσσολογικά στοιχεία.

Στόχο της παρούσας ανάλυσης και μελέτης αποτελεί, κατ' αρχήν, η έρευνα συσχέτισης αποτελεσμάτων ομάδων με συναισθηματικά δεδομένα οπαδών από το Twitter. Ειδικότερα, ερευνάται η δυνατότητα ύπαρξης στοιχείων που ενισχύουν τη σύνδεση μεταξύ οπαδών, ομάδας και αθλητικών επιδόσεων μέσω των social media. Όσον αφορά τα δεδομένα που επεξεργάστηκαν, τα στοιχεία αυτά περιλαμβάνουν τα ποσοστά νίκης, τη συναισθηματική επίδραση, και την ομαδική δυναμική.

Ταυτοχρόνως, στοχεύεται η δημιουργία και αξιολόγηση μοντέλων μηχανικής μάθησης για πρόβλεψη αποτελεσμάτων αγώνων με υψηλή ακρίβεια. και η ανάπτυξη δεικτών για ενσωμάτωση στους αλγόριθμους μηχανικής μάθησης.

Η παρούσα προσέγγιση μπορεί να θεωρηθεί ότι συνδέεται με τις εξής εφαρμογές και οφέλη: βελτίωση στρατηγικής εμπλοκής οπαδών, ενίσχυση ψυχολογίας παικτών και δυναμωμένη σχέση με τους υποστηρικτές. Η έρευνα αποκαλύπτει ισχυρή συσχέτιση ανάμεσα στην ενεργή υποστήριξη των οπαδών και τη βελτίωση των αποτελεσμάτων. Η υποστήριξη αυτή εκτείνεται πέρα από τις εμφανείς εκφράσεις, όπως το χειροκρότημα στο γήπεδο ή τα σχόλια στα κοινωνικά δίκτυα, και συμβάλλει στη δημιουργία μιας διαρκούς σχέσης εμπιστοσύνης και αφοσίωσης.

Επιπλέον, η κατανόηση των συναισθημάτων και των προτιμήσεων του κοινού μέσω της ανάλυσης δεδομένων ανοίγει νέους δρόμους για καινοτόμες στρατηγικές μάρκετινγκ και αλληλεπίδρασης. Η ανάλυση των hashtags, των σχολίων και των συναισθημάτων επιτρέπει την προσαρμογή των εμπειριών στις ανάγκες του κοινού, ενισχύοντας τη σύνδεση με τους υποστηρικτές και την απόδοσή τους στο αγωνιστικό πεδίο.

Λέξεις – Κλειδιά

Ανάλυση Συναισθήματος, X (ex Twitter), Μηχανική Μάθηση, Ποδόσφαιρο

Abstract

The present study concerns the Sentiment Analysis of the posts of users on social media (Social Media) in the field of Football (USA: Soccer). The dataset size 1.000.000 involves posts by fans and football fans on the social media platform "X" - formerly "Twitter", with data and teams from the English Premier League Championship.

Sentiment Analysis is considered to be a relatively recent application field of Natural Language Processing (NLP) and Computational Linguistics. Although Sentiment Analysis is usually applied to data of users and/or consumers of products and services, in recent years, Sentiment Analysis has also been applied to data of a more complex content, even in the field of Journalism and Politics. Data from the field of Football and football fans and/or supporters of football teams can be characterized to be of a complex nature, mainly due to a lower degree of a predefined vocabulary and terminology and a remarkable presence of socio-cultural and socio-linguistic elements.

The aim of the present analysis and study is, primarily, to investigate the correlation of team results with emotional data of fans from Twitter. In particular, we investigate the possibility of elements that strengthen the connection between fans, team and sports performance through social media. In the data concerned, these elements include win percentages, emotional impact, and team dynamics.

Additional targets of the present approach include the creation and evaluation of machine learning models to predict match results with high accuracy and the development of indicators for incorporation into machine learning algorithms.

The present approach can be associated with the following benefits: improved strategic fan engagement, enhanced player psychology and strengthened relationship with fans-supporters. Our research reveals a strong correlation between active fan support and improved results. This active fan support extends beyond the obvious expressions - such as applause during the game or comments on social networks - and helps create a lasting relationship of trust and loyalty.

Understanding audience sentiments and preferences through data analysis is considered to open new avenues for innovative marketing and interaction strategies. The analysis of hashtags, comments and sentiments enables the tailoring of experiences to the needs of the sports fans, enhancing their connection with the players – athlete's performance in the game.

Keywords

Sentiment-Analysis, X (exTwitter), Machine Learning, Football

Περιεχόμενα

Κατάλογος Εικόνων / Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες & Ακρωνύμια

ΔΕ	Διπλωματική Εργασία
ΕΚΠΑ	Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
ΘΕ	Θεματική Ενότητα
ΠΣ	Πρόγραμμα Σπουδών
NLP	Natural Language Processing
CBR	Case Based Reasoning
NN	Neural Network
SVM	Suport Vector Machine
ANN	Artificial Neural Network
MT	Machine Translation
QA	Question Answer
IR	Information Retrivial
HIS	Human Computer Interaction
ASR	Automatic Speech Recognition
TTS	Text to Speech
RF	Random Forest
MLP	Multi Layer Perceptron

1. Εισαγωγή

Η παρούσα μελέτη αφορά την Ανάλυση Συναισθήματος (Sentiment Analysis) στις αναρτήσεις χρηστών σε μέσα κοινωνικής δικτύωσης (Social Media) στο πεδίο του Ποδοσφαίρου (Football, USA: Soccer). Πρόκειται για αναρτήσεις οπαδών και ποδοσφαιρόφιλων στην πλατφόρμα κοινωνικής δικύωσης «X» - πρώην «Twitter «, με δεδομένα και ομάδες από το Πρωτάθλημα English Premier League («Αγγλικό Ποδόσφαιρο»).

Η Ανάλυση Συναισθήματος (Sentiment Analysis) αποτελεί σχετικά πρόσφατο πεδίο εφαρμογών της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP) και της Υπολογιστικής Γλωσσολογίας (Computational Linguistics). Μολονότι η Ανάλυση Συναισθήματος συνήθως εφαρμόζεται σε δεδομένα χρηστών ή/και καταναλωτών προϊόντων και υπηρεσιών, τα τελευταία χρόνια εφαρμόζεται και σε δεδομένα πιο πολύπλοκου περιεχομένου, ακόμη και στον τομέα της Δημοσιογραφίας και της Πολιτικής. Τα δεδομένα από το πεδίο του Ποδοσφαίρου και των ποδοσφαιρόφιλων ή/και οπαδών ποδοσφαιρικών ομάδων μπορούν να χαρακτηριστούν ως σύνθετου και πολύπλοκου περιεχομένου, κυρίως επειδή δεν περιορίζονται σε προκαθορισμένο λεξιολόγιο και ορολογία και επειδή συχνά περιέχουν αρκετά κοινωνιο-γλωσσολογικά στοιχεία.

Εδώ, το σύνολο (data set) των δεδομένων που επεξεργάστηκε και μελετήθηκε έχει μέγεθος 1.000.562 και η συλλογή πραγματοποιήθηκε απευθείας από την πλατφόρμα κοινωνικής δικύωσης «X» και πιο συγκεκριμένα από την πλατφρόμα του Kaggle: https://www.kaggle.com/datasets/wjia26/epl-teams-twitter-sentiment-dataset/data.

Στόχο της παρούσας ανάλυσης και μελέτης αποτελεί, κατ' αρχήν, η έρευνα συσχέτισης αποτελεσμάτων ομάδων με συναισθηματικά δεδομένα οπαδών από το Twitter. Ειδικότερα, ερευνάται η δυνατότητα ύπαρξης στοιχείων που ενισχύουν τη σύνδεση μεταξύ οπαδών, ομάδας και αθλητικών επιδόσεων μέσω των social media. Όσον αφορά τα δεδομένα που επεξεργάστηκαν, τα στοιχεία αυτά περιλαμβάνουν τα ποσοστά νίκης, τη συναισθηματική επίδραση, και την ομαδική δυναμική.

Ταυτοχρόνως, στοχεύεται η δημιουργία και αξιολόγηση μοντέλων μηχανικής μάθησης για πρόβλεψη αποτελεσμάτων αγώνων με υψηλή ακρίβεια. και η ανάπτυξη δεικτών για ενσωμάτωση στους αλγόριθμους μηχανικής μάθησης.

2. Θεωρητικό Πλαίσιο (Ανάλυση Συναισθηματος και Μηχανική Μαθηση)

2.1 Ανάλυση Συναισθήματος

2.1.1 Τι είναι Αναλυση Συναισθήματος

Η ανάλυση συναισθήματος είναι η διαδικασία αναγνώρισης και ποσοτικοποίησης του συναισθηματικού τόνου που εκφράζεται σε ένα δεδομένο κείμενο. Πρόκειται για τον καθορισμό του εάν ο συγγραφέας ενός κειμένου αισθάνεται θετικά, αρνητικά ή ουδέτερα για ένα συγκεκριμένο θέμα. Με άλλα λογια είναι μια τεχνική που χρησιμοποιείται στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP) για να προσδιορίσει υποκειμενικά στοιχεία που περιέχονται σε ένα κείμενο. Είναι σαν να προσπαθούμε να διαβάσουμε τη διάθεση του συγγραφέα μέσα από τα λόγια του (Jurafsky & Martin, 2000).

Ως πεδίο, έχει αναδειχθεί σε σημαντικό τομέα επεξεργασίας πληροφοριών, λόγω της αυξανόμενης ερευνητικής δραστηριότητας και των ποικίλων ευρημάτων που δημοσιεύονται συνεχώς.Η ανάλυση συναισθημάτων βοηθά στην κατανόηση απόψεων, συναισθημάτων και στάσεων, ενώ οι πρόσφατες τάσεις τονίζουν τη σημασία της προσαρμογής της σε νέες προκλήσεις, όπως η πανδημία COVID-19. Για παράδειγμα, ο Costola και οι συνεργάτες του ανέπτυξαν μοντέλο BERT προσαρμοσμένο στις χρηματοπιστωτικές αγορές, το οποίο αναγνωρίζει το πλαίσιο κάθε λέξης. Παράλληλα, ο Mao και οι συνεργάτες του διερεύνησαν τις προκαταλήψεις των γλωσσικών μοντέλων (PLM) στη συναισθηματική υπολογιστική. Αν και οι περισσότερες μελέτες εστιάζουν σε θέματα όπως η ταξινόμηση συναισθημάτων, τα προκαταρκτικά γλωσσικά μοντέλα και οι διεπαφές συνομιλιών, υπάρχει περιορισμένη έρευνα για την εξέλιξη της ανάλυσης συναισθημάτων σε άλλα πεδία. Έτσι, μια πιο ολιστική προσέγγιση που καλύπτει ανεξερεύνητες πτυχές και προτείνει κατευθύνσεις για μελλοντική έρευνα είναι αναγκαία (Önden et al., 2024).

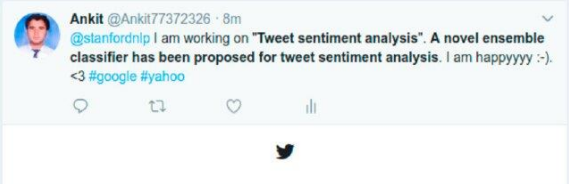
2.1.2 Η Πλατφόρμα Κοινωνικής Δικτύωσης «X» - Πρώην Twitter

Το «X» - πρώην Twitter, μια γνωστή πλατφόρμα κοινωνικής δικτύωσης (social media) με περισσότερους από 300 εκατομμύρια χρήστες, προσφέρει τη δυνατότητα αποστολής μηνυμάτων 140 χαρακτήρων, που ονομάζονται tweets. Κάθε ανάρτηση «tweet» στην πλατφόρμα έχει τη δυνατότητα να περιέχει συνδέσμους, εικόνες και βίντεο. Βασικά στοιχεία περιλαμβάνουν τα handles για δημόσιες αλληλεπιδράσεις, τα hashtags για καλύτερη αναζήτηση και το re-tweeting για κοινοποίηση αναρτήσεων. Οι χρήστες έχουν τη δυνατότητα να ακολουθούν λογαριασμούς που τους ενδιαφέρουν, όπως άτομα, εταιρείες ή οργανώσεις, και να λαμβάνουν ενημερώσεις σε πραγματικό χρόνο για ζητήματα που τους επηρεάζουν. Το «X» - πρώην Twitter δημιουργεί άνω των 500 εκατομμυρίων ενημερώσεων κάθε ημέρα, αποτελώντας μια βασική πηγή περιεχομένου από τους χρήστες του (Larhgotra & Walia, 2024).
Για λόγους σαφήνειας, εφ’ εξής θα αναφερόμαστε στην πλατφόρμα «X» με την αρχική – και γνωστότερη της ονομαστία - ως «Twitter» και στις αναρτήσεις στην πλατφόρμα «X» ως ««tweets».

Τα γενικά χαρακτηριστικά της πλατφόρμας κοινωνικής δικτύωσης «X» - πρώην Twitter μπορούν να συνοψιστούν ως εξής:

- Υπηρεσία: Κοινωνικό δίκτυο όπου οι χρήστες μοιράζονται σύντομες αναρτήσεις (παλιότερα "tweets"), εικόνες, βίντεο και αλληλεπιδρούν με περιεχόμενο.
- Χαρακτηριστικά: Άμεσα μηνύματα, βιντεοκλήσεις/ηχητικές κλήσεις, σελιδοδείκτες, λίστες, κοινότητες, chatbot (Grok), αναζήτηση εργασίας, Community Notes για αξιολόγηση περιεχομένου, και Spaces (λειτουργία κοινωνικού ήχου).
- Επαναπροσδιορισμός: Μετονομάστηκε σε "X" τον Ιούλιο του 2023, με το λογότυπο του πουλιού να αποσύρεται μέχρι τον Μάιο του 2024.

Εικόνα 2.1 Παράδειγμα Tweet



Η ιστορία της της πλατφόρμας κοινωνικής δικτύωσης «X» - πρώην Twitter μπορεί να συνοψιστεί ως εξής:

- Ιδρυτές: Jack Dorsey, Noah Glass, Biz Stone και Evan Williams τον Μάρτιο του 2006.
- Έναρξη: Ιούλιος 2006.
- Ανάπτυξη: Μέχρι το 2012, είχε πάνω από 100 εκατομμύρια χρήστες που παρήγαγαν 340 εκατομμύρια tweets καθημερινά.
- Όριο χαρακτήρων: Αρχικά 140 χαρακτήρες, αυξήθηκε στους 280 το 2017, ενώ το όριο καταργήθηκε για συνδρομητικούς λογαριασμούς το 2023.

Εικόνα 2.2 Παλαιό και Καινούργιο Logo Twitter

Οι βασικές πληροφορίες και στοιχεία της πλατφόρμας κοινωνικής δικτύωσης «X» - πρώην Twitter μπορούν να συνοψιστούν ως εξής:

Ιδιοκτησία & Ηγεσία

- Εξαγορά: Αγοράστηκε από τον Elon Musk τον Οκτώβριο του 2022 έναντι 44 δισεκατομμυρίων δολαρίων· η ιδιοκτησία μεταφέρθηκε στη X Corp. τον Μάρτιο του 2023.
- Ηγεσία: Η Linda Yaccarino ανέλαβε CEO στις 5 Ιουνίου 2023, ενώ ο Musk παρέμεινε πρόεδρος και CTO.
- Μείωση Αξίας: Η αξία της πλατφόρμας εκτιμάται ότι μειώθηκε κατά 72% μέχρι τον Οκτώβριο του 2024 μετά την εξαγορά από τον Musk.

Τάσεις Χρήσης

- Bots: Εκτιμάται ότι το 15% των λογαριασμών (48 εκατομμύρια) είναι bots.
- Εμπλοκή Χρηστών: Το 10% των χρηστών παράγει πάνω από το 80% του περιεχομένου.
- Μείωση Χρήσης: Η παγκόσμια χρήση μειώθηκε κατά 15% μετά την εξαγορά από τον Musk, αν και ο ίδιος ισχυρίστηκε ότι η πλατφόρμα είχε 600 εκατομμύρια χρήστες τον Μάιο του 2024.

Αντιπαραθέσεις

- Παραπληροφόρηση & Λόγος Μίσους: Δέχθηκε κριτική για την αύξηση παραπληροφόρησης και λόγου μίσους μετά την εξαγορά από τον Musk.
- Αμφιλεγόμενα Στατιστικά: Ενώ οι εταιρείες ανάλυσης εφαρμογών αναφέρουν μείωση χρήσης, ο Musk αμφισβητεί τους ισχυρισμούς.
  (Wikipedia contributors, 2025)

1. Ανάλυση συναισθήματος στο Twitter

Η ανάλυση συναισθήματος, γνωστή και ως εξόρυξη απόψεων, αποτελεί ένα σημαντικό εργαλείο για την εξαγωγή υποκειμενικών πληροφοριών από κείμενα. Τα σχόλια και τα tweets συχνά περιέχουν συναισθήματα, τα οποία μπορούν να παρέχουν χρήσιμους δείκτες για πληθώρα εφαρμογών. Σύμφωνα με τις μελέτες, τα συναισθήματα μπορούν να κατηγοριοποιηθούν σε δύο κύριες κατηγορίες: θετικά και αρνητικά συναισθήματα. Η ανάλυση συναισθήματος είναι μία τεχνική επεξεργασίας φυσικής γλώσσας που στοχεύει στην ποσοτικοποίηση μιας εκφρασμένης άποψης ή συναισθήματος μέσα από τα tweets.

Έρευνες έχουν επικεντρωθεί σε θέματα όπως η πρόβλεψη τάσεων στις πωλήσεις, η αξιολόγηση ποιότητας προϊόντων και υπηρεσιών, καθώς και η σύνδεση σχολίων με πολιτικά συναισθήματα (Larhgotra & Walia, 2024). Με τη χρήση τεχνικών βαθιάς μάθησης, που συνδυάζουν ανίχνευση συναισθημάτων και ανακατασκευή συνομιλιών, οι συζητήσεις στο Twitter αποτελούν πλέον πεδίο ενδιαφέροντος. Εταιρείες όπως η Microsoft και η Google έχουν αναπτύξει συστήματα ανάλυσης διάθεσης για επιχειρηματικές και βιομηχανικές εφαρμογές (Larhgotra & Walia, 2024). Τα συστήματα αυτά μετατρέπουν δεδομένα από πηγές όπως blogs, σχόλια και κριτικές σε κείμενα, χρησιμοποιώντας μεθόδους όπως η τοκενοποίηση, η αποκοπή και η αφαίρεση συνδετικών λέξεων, ενώ η επιλογή χαρακτηριστικών είναι κρίσιμη για την αποδοτικότητα (Larhgotra & Walia, 2024). Στοχεύει στην εξαγωγή της πολικότητας (θετικό ή αρνητικό συναίσθημα) και της

υποκειμενικότητας του κειμένου, βασιζόμενη στον τρόπο με τον οποίο οι λέξεις και οι φράσεις εκφράζουν συναισθηματική φόρτιση (Sarlan et al., 2014).

Το Twitter σημαντικό εργαλείο για την εξαγωγή υποκειμενικών πληροφοριών από κείμενα. Το Twitter, με τη φύση του ως πλατφόρμα microblogging, παρέχει μια πλούσια πηγή περιεχομένου που δημιουργείται από χρήστες, αντικατοπτρίζοντας τις απόψεις και τα συναισθήματα σε πραγματικό χρόνο πάνω σε διάφορα θέματα. Λόγω του περιορισμένου μήκους των αναρτήσεων (tweets) — οι οποίες δεν μπορούν να ξεπερνούν τους 280 χαρακτήρες (προηγουμένως 140) — οι χρήστες εκφράζουν τις σκέψεις τους με συντομία, καθιστώντας το Twitter ένα ιδανικό σύνολο δεδομένων για ανάλυση συναισθήματος. Τα tweets μπορούν να αναλυθούν για να εκτιμηθεί η κοινή γνώμη σχετικά με γεγονότα, προϊόντα ή πολιτικές απόψεις. Στοιχεία-κλειδιά του Twitter, όπως τα hashtags, οι αναφορές (mentions), τα retweets και οι απαντήσεις (replies), προσθέτουν πολυπλοκότητα στην ανάλυση συναισθήματος. Τα hashtags συχνά χρησιμοποιούνται για να υποδηλώσουν δημοφιλή θέματα, ενώ οι αναφορές και οι απαντήσεις επιτρέπουν την αλληλεπίδραση μεταξύ των χρηστών, επηρεάζοντας τη ροή του συναισθήματος μέσα στις συνομιλίες. Τα retweets βοηθούν στη διάδοση περιεχομένου, καθιστώντας ορισμένα συναισθήματα πιο έντονα σε μεγαλύτερο δίκτυο χρηστών. Τεχνικές επεξεργασίας φυσικής γλώσσας (NLP), όπως οι μεθόδοι που βασίζονται σε λεξικά και αλγορίθμους μηχανικής μάθησης, χρησιμοποιούνται συνήθως για την κατάταξη των tweets σε θετικά, αρνητικά ή ουδέτερα. Ωστόσο, η ανεπίσημη και συχνά συντομογραφική γλώσσα που χρησιμοποιείται στα tweets, καθώς και η παρουσία emojis, hashtags και συνδέσμων (URLs), μπορεί να δημιουργεί προκλήσεις για την ακριβή ανάλυση συναισθήματος. Παρ' όλες αυτές τις προκλήσεις, η ανάλυση συναισθήματος στο Twitter παραμένει μία από τις πιο αποτελεσματικές μεθόδους για την παρακολούθηση της κοινής γνώμης και την ανίχνευση της εξέλιξης των απόψεων με την πάροδο του χρόνου (Giachanou & Crestani, 2016).

Εικόνα 2.3 Κατανόηση Θετικού & Αρνητικού Tweet

Sentiment	Tweet mention
Positive	Maybe I'm mad but I'm now the proud owner of a potentially #bendy #iPhone6, it's so much bigger than the #4s
	Finally got to see an iPhone 6 today. Not revolutionary at all but it's absolutely gorgeous. (And I want one). #iPhone6
Negative	I'm not sure I want it. It's too big to fit in my back pocket! lol #iphone6
	I'm really disappointed with the #iPhone6. It took them 2 years to change the screen & size. Let down.

(Kim et al., 2016)

2. Επεξεργασία Δεδομένων για Ανάλυση συναισθήματος στο Twitter

Η προεπεξεργασία των δεδομένων αποτελεί απαραίτητο βήμα για τη βελτίωση της απόδοσης των εργαλείων ανάλυσης συναισθήματος, καθώς η εφαρμογή ανάλυσης σε ακατέργαστα tweets μπορεί να οδηγήσει σε χαμηλή ακρίβεια. Τα tweets, ως ακατέργαστα δεδομένα, πρέπει να υποβληθούν σε μια σειρά από προκαταρκτικά βήματα προκειμένου να καθαριστούν και να βελτιωθεί η ποιότητα της ανάλυσης (International Journal of Scientific

Research in Science, Engineering and Technology IJSRSET, 2016b). Η προεπεξεργασία περιλαμβάνει τα εξής στάδια:

- Φιλτράρισμα (Filtering): Στο πρώτο στάδιο, καθαρίζονται τα ακατέργαστα δεδομένα αφαιρώντας στοιχεία που δεν συνεισφέρουν στην ανάλυση συναισθήματος. Αυτά περιλαμβάνουν συνδέσμους URL (π.χ., http://twitter.com), ειδικούς όρους του Twitter, όπως "RT" (ReTweet), ονόματα χρηστών (@username), και emoticons (International Journal of Scientific Research in Science, Engineering and Technology IJSRSET, 2016b).
- Τοκενικοποίηση (Tokenization): Σε αυτό το στάδιο, το κείμενο διασπάται σε μικρότερα τμήματα, συνήθως λέξεις ή φράσεις. Ο διαχωρισμός γίνεται με τη χρήση διαστημάτων ή σημείων στίξης, δημιουργώντας έτσι μια σειρά από "tokens" που αποτελούν τη βάση για περαιτέρω ανάλυση (International Journal of Scientific Research in Science, Engineering and Technology IJSRSET, 2016b).
- Αφαίρεση Λέξεων-Stopwords (Removal of Stopwords): Ακολουθεί η αφαίρεση κοινών λέξεων που δεν προσφέρουν πληροφορίες για το συναίσθημα, όπως άρθρα ("ο", "η", "το") και άλλες λέξεις όπως "σε", "είναι", "αυτό". Η απομάκρυνση αυτών των λέξεων βοηθά στην εστίαση του αλγορίθμου στις σημαντικές λέξεις του κειμένου (International Journal of Scientific Research in Science, Engineering and Technology IJSRSET, 2016b).
- Κατασκευή n-grams (Construction of n-grams): Το τελευταίο στάδιο περιλαμβάνει τη δημιουργία n-grams, δηλαδή ομάδων διαδοχικών λέξεων. Ένα κρίσιμο μέρος αυτού του βήματος είναι η σωστή διαχείριση των λέξεων άρνησης, όπως "όχι" και "μη", οι οποίες συνδέονται με την αμέσως επόμενη ή προηγούμενη λέξη. Για παράδειγμα, η φράση «δεν μου αρέσει η remix μουσική» παράγει n-grams όπως «δεν+μου» και «μου+αρέσει». Η σωστή καταγραφή της άρνησης βελτιώνει την ακρίβεια της ανάλυσης συναισθήματος, καθώς οι λέξεις άρνησης επηρεάζουν την πολικότητα των συναισθημάτων (International Journal of Scientific Research in Science, Engineering and Technology IJSRSET, 2016b).

Η λήψη αυτών των βημάτων είναι ζωτικής σημασίας για την ακρίβεια και την ποιότητα των αποτελεσμάτων στην ανάλυση συναισθήματος, δεδομένου ότι η γλωσσική πολυπλοκότητα του Twitter απαιτεί εξειδικευμένες μεθόδους προεπεξεργασίας (International Journal of Scientific Research in Science, Engineering and Technology IJSRSET, 2016b).

2.1.3 Οι Βασικές Προσεγγίσεις και Μέθοδοι Επεξεργασίας

1. **Μέθοδος Επεξεργασίας Βασισμένη σε λεξικό**

Η προσέγγιση βασισμένη σε λεξικό είναι μια ευρέως χρησιμοποιούμενη μέθοδος στην ανάλυση συναισθήματος που στηρίζεται σε μια προκαθορισμένη λίστα λέξεων, καθεμία από τις οποίες συνδέεται με ένα συγκεκριμένο συναίσθημα, είτε θετικό είτε αρνητικό. Αυτή η μέθοδος αξιολογεί το συναίσθημα ενός κειμένου αναλύοντας τον συναισθηματικό προσανατολισμό των φράσεων που περιέχει. Η διαδικασία περιλαμβάνει συνήθως αρκετά βασικά βήματα: προεπεξεργασία του κειμένου με αφαίρεση της στίξης, αρχικοποίηση της συνολικής βαθμολογίας πολικότητας, και έλεγχο κάθε λέξης (ή τοκεν) έναντι ενός λεξικού συναισθήματος. Αν μια λέξη βρεθεί στο λεξικό, το αντίστοιχο συναίσθημα προστίθεται στη συνολική βαθμολογία, συμβάλλοντας στην τελική πρόβλεψη του συναισθήματος του κειμένου. Ένα από τα πλεονεκτήματα της προσέγγισης αυτής είναι η απλότητά της και η αποτελεσματικότητά της σε περιπτώσεις όπου τα επισημασμένα δεδομένα είναι περιορισμένα. Ωστόσο, ενδέχεται να μην συλλαμβάνει πάντα τις λεπτές αποχρώσεις των συναισθημάτων που εξαρτώνται από τα συμφραζόμενα, τόσο αποτελεσματικά όσο οι μέθοδοι που βασίζονται στη μάθηση, οι οποίες μπορούν να εκπαιδεύσουν μοντέλα προσαρμοσμένα σε συγκεκριμένες εφαρμογές. Παρά τις αδυναμίες της, η μέθοδος αυτή παραμένει ένα πολύτιμο εργαλείο, ιδιαίτερα σε περιπτώσεις όπου η πρόσβαση σε εκτενή επισημασμένα δεδομένα είναι περιορισμένη (Kolla, 2016).

2. Μηχανική Μάθηση

Στην ανάλυση συναισθημάτων, το στάδιο της ταξινόμησης αξιοποιεί έναν ταξινομητή που έχει εκπαιδευτεί μέσω αλγορίθμων μηχανικής μάθησης. Οι δύο βασικοί τύποι είναι η εποπτευόμενη (supervised) και η μη εποπτευόμενη (unsupervised) μάθηση (Larhgotra & Walia, 2024). Οι πιο βασικοί ταξινομητές είναι οι παρακάτω:

- Πιθανοτικός Ταξινομητής: Προβλέπει κατανομή πιθανοτήτων σε κατηγορίες, αντί να προσδιορίζει απλώς την πιο πιθανή.
- Γραμμικός Ταξινομητής**:** Χρησιμοποιεί γραμμικά μοντέλα πρόβλεψης για να προσδιορίσει την κλάση ενός χαρακτηριστικού. Δημοφιλείς μέθοδοι περιλαμβάνουν τα Νευρωνικά Δίκτυα (NN) και τις Υποστηρικτικές Μηχανές Διανυσμάτων (SVM).
- Ταξινομητής Βασισμένος σε Κανόνες: Χρησιμοποιεί κανόνες τύπου "AN-TOTE" για την ταξινόμηση χαρακτηριστικών σε προκαθορισμένες κατηγορίες.

- Ταξινομητής Δέντρου Αποφάσεων: Διαχωρίζει συνεχώς τον χώρο χαρακτηριστικών σε μικρότερους για την ταξινόμηση ή παλινδρόμηση.

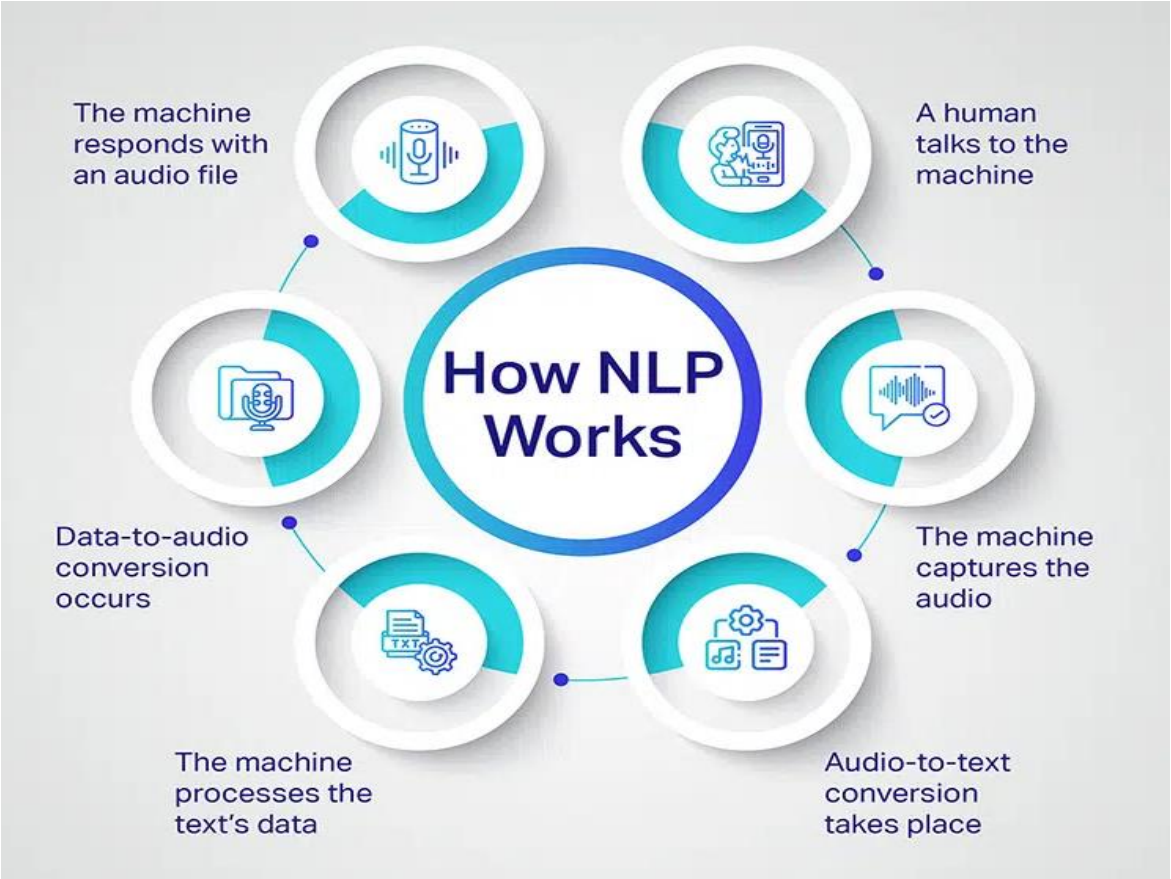
Η εποπτευόμενη μάθηση είναι αποτελεσματική για την ανάλυση συναισθημάτων, αλλά η προετοιμασία δεδομένων απαιτεί ανθρώπινη παρέμβαση. Για την αντιμετώπιση αυτής της πρόκλησης, η μη εποπτευόμενη μάθηση χρησιμοποιεί δείκτες από συναισθηματικά λεξικά για τον προσδιορισμό της πολικότητας (Larhgotra & Walia, 2024).

Οι μέθοδοι μηχανικής μάθησης βασίζονται σε προσεγγίσεις επιβλεπόμενης ταξινόμησης όταν η ανίχνευση συναισθημάτων θεωρείται ως δυαδική, δηλαδή θετική και αρνητική. Αυτή η προσέγγιση χρησιμοποιεί επισημασμένα δεδομένα για να εκπαιδεύσει ταξινομητές. Γίνεται εμφανές ότι τα χαρακτηριστικά του τοπικού συμφραζομένου μιας λέξης είναι απαραίτητο να ληφθούν υπόψη, όπως η άρνηση και η ενίσχυση του συναισθήματος. Η μηχανική μάθηση επιτρέπει στα συστήματα να αναγνωρίζουν και να διακρίνουν συναισθήματα με βάση προηγούμενα δεδομένα, καθιστώντας τα ικανά να προσαρμόζονται σε συγκεκριμένες ανάγκες και να βελτιώνονται με την πάροδο του χρόνου καθώς εκτίθενται σε περισσότερα δεδομένα (Kolla, 2016).

3. Επεξεργασία Φυσικής Γλώσσας (NLP)

Οι μηχανισμοί Επεξεργασίας Φυσικής Γλώσσας (NLP) βασίζονται στη μηχανική μάθηση και, κυρίως, στη στατιστική μάθηση, η οποία χρησιμοποιεί έναν γενικό αλγόριθμο μάθησης μαζί με μια τεράστια ποσότητα δειγμάτων δεδομένων για να μάθει τους κανόνες. Η ανάλυση συναισθήματος έχει αντιμετωπιστεί ως μια εργασία Επεξεργασίας Φυσικής Γλώσσας σε διάφορα επίπεδα. Από την ταξινόμηση σε επίπεδο εγγράφου, έχει επεκταθεί στη διαχείριση σε επίπεδο πρότασης και πρόσφατα σε επίπεδο φράσης. Το NLP είναι ένας τομέας της επιστήμης των υπολογιστών που περιλαμβάνει τη δημιουργία συστημάτων ικανά να εξάγουν νόημα από την ανθρώπινη γλώσσα και να αλληλεπιδρούν με τον πραγματικό κόσμο με έναν διαδραστικό τρόπο (Kolla, 2016).

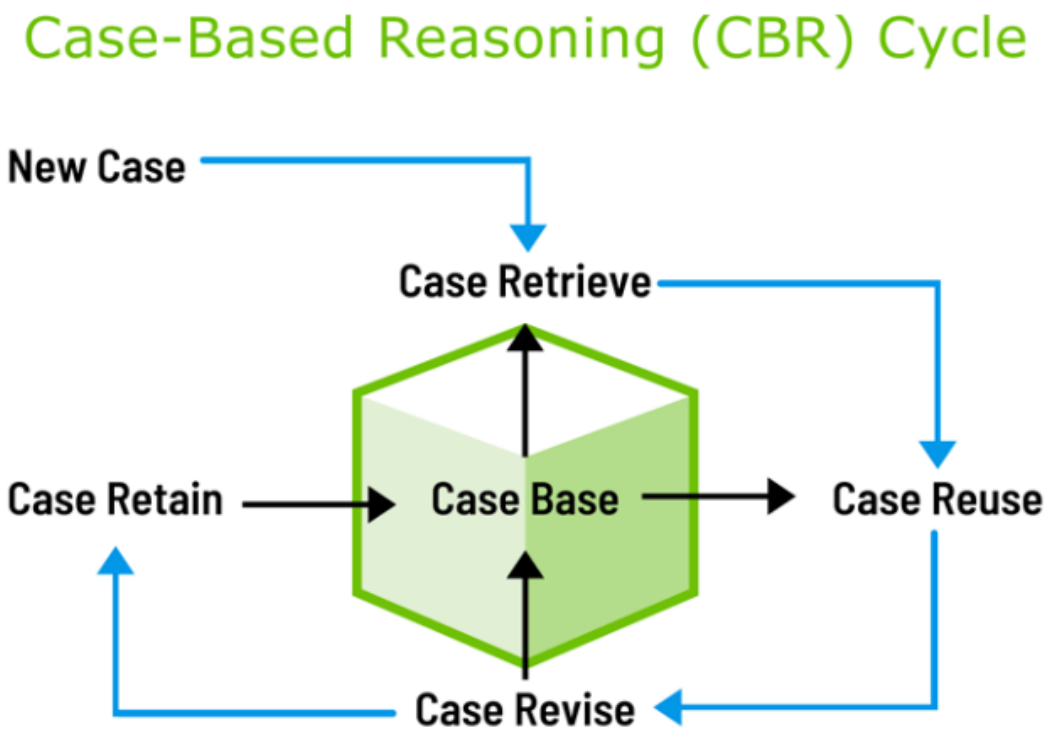
Εικόνα 2.4 Πως Λειτουργεί το Natural Language Processing



4. Μέθοδος Βασισμένη στην Περίπτωση Λογικής (CBR)

Η μέθοδος της Βασισμένης στην Περίπτωση Λογικής (CBR) αποτελεί μια τεχνική υλοποίησης της ανάλυσης συναισθήματος, η οποία ανακαλεί παρελθόντα επιτυχώς λυμένα ζητήματα και χρησιμοποιεί τις λύσεις τους για την επίλυση των τρεχόντων στενά συναφών ζητημάτων. Παρατηρείται ότι μερικά από τα πλεονεκτήματα της χρήσης της είναι ότι δεν χρειάζεται ένα ρητό εξειδικευμένο μοντέλο και έτσι η εξαγωγή γίνεται μια εργασία συλλογής ιστορικού περιποίησης και αυτά τα συστήματα μπορούν να μάθουν αποκτώντας νέες γνώσεις ως περιπτώσεις (Kolla, 2016).

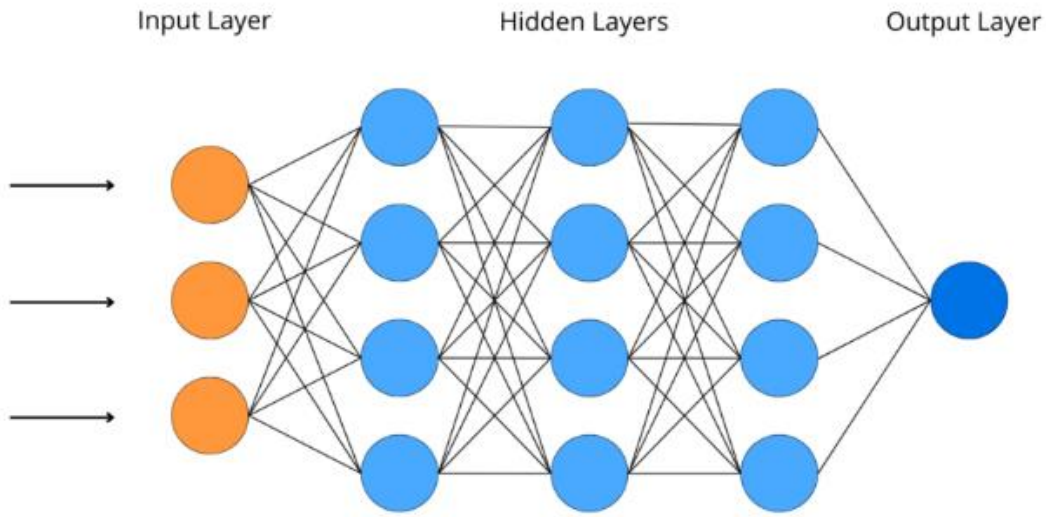
Εικόνα 2.5 Κύκλος Περιπτωσιολογικού Συλλογισμού



5. Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (ANN), γνωστά και ως νευρωνικά δίκτυα, είναι μια τεχνική που συνδέει ομάδες τεχνητών νευρώνων. Θα επεξεργάζεται πληροφορίες χρησιμοποιώντας την προσέγγιση των δικτύων για τον υπολογισμό (Kolla, 2016). Τα νευρωνικά δίκτυα, όπως αποδεικνύεται από τις επιτυχημένες εφαρμογές τους σε διάφορους τομείς, αποτελούν ένα ευέλικτο εργαλείο για την αντιμετώπιση σύνθετων προβλημάτων. Η ικανότητά τους να αναγνωρίζουν μοτίβα, τάσεις και σχέσεις εντός των δεδομένων τα καθιστά ιδιαίτερα κατάλληλα για εργασίες όπως η πρόβλεψη, η πρόγνωση και η ταξινόμηση. Από την ανάλυση της οικονομικής σταθερότητας έως την πρόγνωση του καιρού και τις γεωργικές εφαρμογές, τα νευρωνικά δίκτυα έχουν αποδείξει την αξία τους στην αντιμετώπιση πραγματικών προκλήσεων σε διάφορους κλάδους (Abiodun et al., 2018).

Εικόνα 2.6 Τεχνητά Νευρωνικά Δίκτυα



6. Υβριδική Προσέγγιση

Η υβριδική μέθοδος συνδυάζει τεχνικές που βασίζονται στο λεξικό και στη μηχανική μάθηση. Έρευνες δείχνουν ότι αυτή η μεθοδολογία βελτιστοποιεί την απόδοση της ταξινόμησης.

2.1.4 Χαρακτηριστικά Παραδείγματα Εφαρμογών

1. Μηχανική (Αυτόματη) Μετάφραση

Η μηχανική μετάφραση (Machine Translation - MT) είναι μια βασική εφαρμογή της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP), αξιοποιώντας προχωρημένα μοντέλα όπως η αρχιτεκτονική encoder-decoder, η οποία αναπτύχθηκε αρχικά για τη μηχανική μετάφραση και έχει γίνει θεμελιώδης σε διάφορα έργα NLP. Η MT αντιμετωπίζει προκλήσεις λόγω δομικών και λεξιλογικών αποκλίσεων μεταξύ των γλωσσών, με την τυπολογία γλωσσών να εξετάζει αυτές τις διαφορές. Τα δίκτυα encoder-decoder, συμπεριλαμβανομένων των transformers, αποτελούνται από έναν κωδικοποιητήπου δημιουργεί μια συμφραστική αναπαράσταση της εισόδου και έναν αποκωδικοποιητή που παράγει το μεταφρασμένο αποτέλεσμα. Οι μηχανισμοί διασταυρούμενης προσοχής (cross-attention) στους transformers επιτρέπουν στον αποκωδικοποιητή να έχει πρόσβαση σε όλες τις καταστάσεις του κωδικοποιητή,

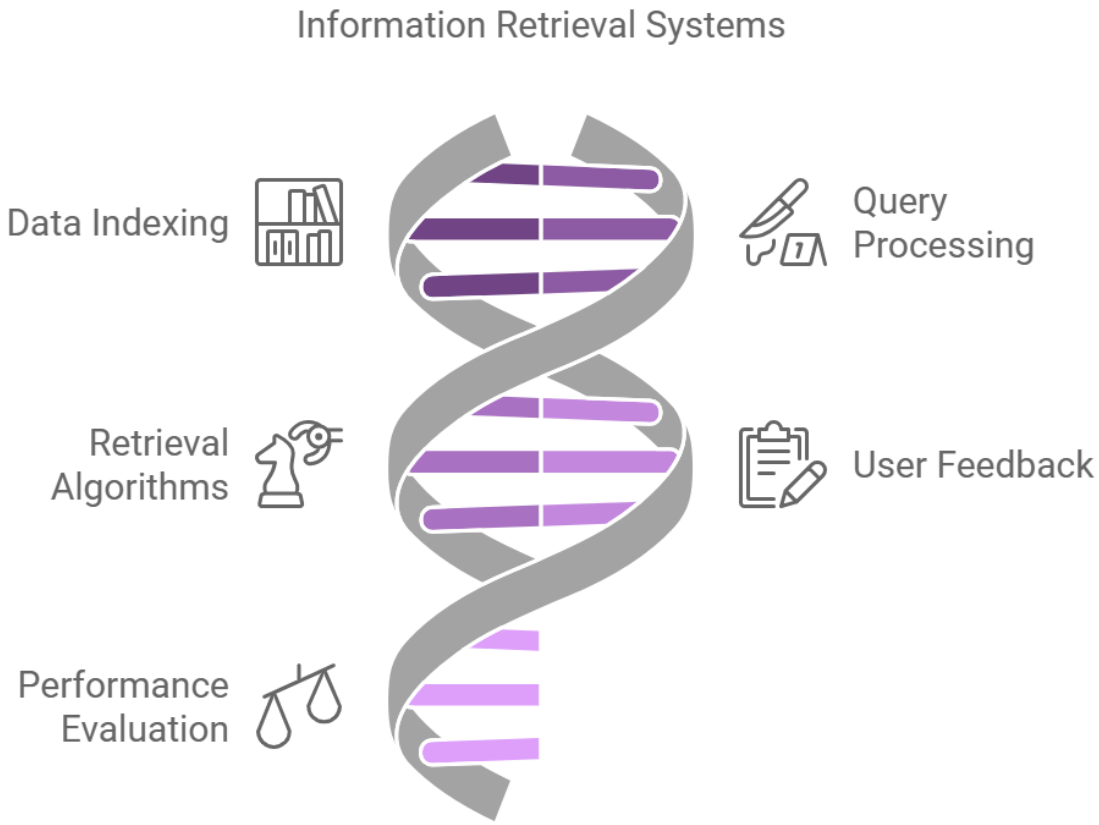
βελτιώνοντας την ακρίβεια της μετάφρασης. Τα μοντέλα MT εκπαιδεύονται σε παράλληλα σώματα κειμένων, γνωστά ως bitexts, και μπορούν επίσης να χρησιμοποιήσουν την αντίστροφη μετάφραση (backtranslation) για τη δημιουργία συνθετικών bitexts από μονογλωσσικά δεδομένα. Η ποιότητα της MT αξιολογείται με βάση την επάρκεια και τη φυσικότητα, με την ανθρώπινη αξιολόγηση να θεωρείται το χρυσό πρότυπο, αν και αυτόματες μετρικές όπως το η ομοιότητα ενσωμάτωσης (embedding similarity) χρησιμοποιούνται επίσης ευρέως (Jurafsky & Martin, 2000).

2. Ανάκτηση Πληροφοριών και Απάντηση Ερωτήσεων

Τα συστήματα ανάκτησης πληροφοριών και απάντησης ερωτήσεων (QA) έχουν σχεδιαστεί για να ικανοποιούν τις ανθρώπινες ανάγκες πληροφόρησης, δεδομένου ότι μεγάλο μέρος της γνώσης είναι κωδικοποιημένο σε κείμενο. Από τις πρώτες δεκαετίες της ύπαρξης των υπολογιστών, υπήρχαν προσπάθειες να απαντώνται ερωτήσεις, όπως οι στατιστικές του μπέιζμπολ στις αρχές του 1960. Σήμερα, οι σύγχρονες μηχανές αναζήτησης έχουν ενσωματώσει μοντέλα γλώσσας, τα οποία είναι εκπαιδευμένα να απαντούν ερωτήσεις με ακρίβεια και ταχύτητα, θολώνοντας τα όρια μεταξύ αναζήτησης και απάντησης ερωτήσεων. Τα συστήματα QA συχνά εστιάζουν σε ερωτήσεις που αφορούν γεγονότα ή λογική και μπορούν να απαντηθούν με απλές πληροφορίες (Jurafsky & Martin, 2000).

Η ανάκτηση πληροφοριών (IR) αναφέρεται στην ανάκτηση όλων των ειδών των μέσων με βάση τις ανάγκες πληροφόρησης των χρηστών, και το σύστημα IR αναφέρεται συχνά ως μηχανή αναζήτησης. Η διαδικασία αυτή περιλαμβάνει την αναζήτηση εγγράφων από μια συλλογή με βάση ένα ερώτημα του χρήστη. Η αξιολόγηση των συστημάτων ανάκτησης γίνεται μέσω μετρικών όπως η ακρίβεια και η ανάκληση, ενώ τα συστήματα QA χρησιμοποιούν την αρχιτεκτονική retriever/reader. Στο στάδιο του retriever, ένα σύστημα IR επιστρέφει μια σειρά από έγγραφα βάσει ενός ερωτήματος, ενώ στο στάδιο του reader, ένα μεγάλο γλωσσικό μοντέλο δημιουργεί μια απάντηση βασισμένη σε αυτά τα έγγραφα. Η απόδοση αυτών των συστημάτων αξιολογείται μέσω ακριβούς αντιστοίχισης, βαθμολογίας F1 ή μέσης αναδρομικής κατάταξης (mean reciprocal rank) (Jurafsky & Martin, 2000).

Εικόνα 2.7 Ανάκτηση Πληροφοριών



3. Διαλογικά Συστήματα– Chatbots

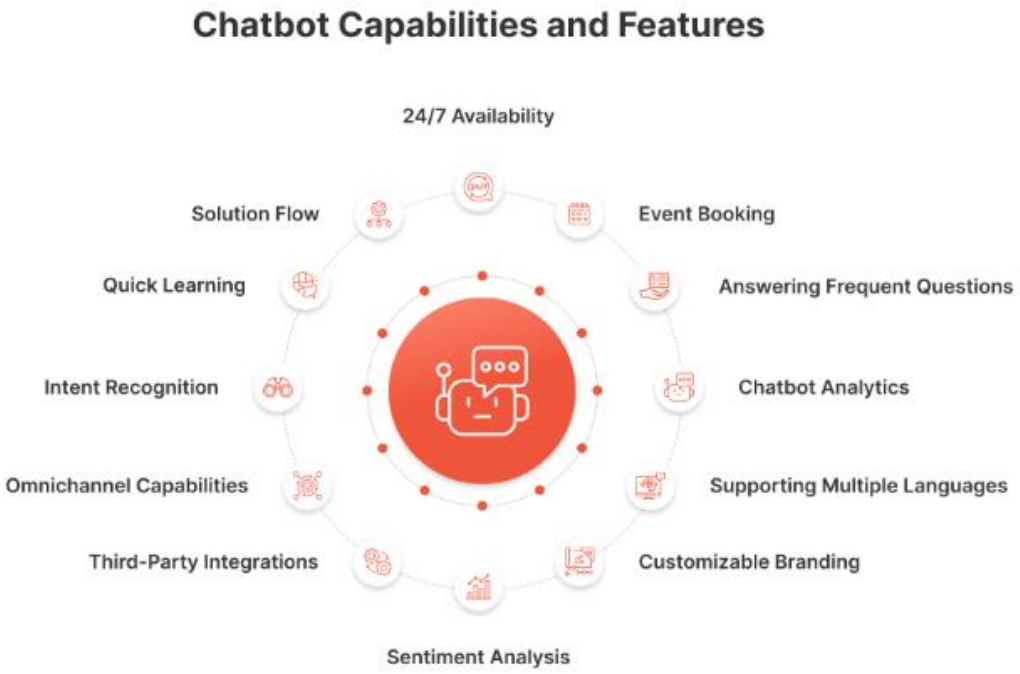
Τα διαλογικά συστήματα και τα chatbots είναι εφαρμογές επεξεργασίας γλώσσας που επιτρέπουν την αλληλεπίδραση μεταξύ ανθρώπων και υπολογιστών μέσω συνομιλίας (Jurafsky & Martin, 2000). Η ανθρώπινη συνομιλία είναι μια πολύπλοκη διαδικασία που απαιτεί συντονισμό και αλληλοκατανόηση, καθιστώντας σημαντική την κατανόηση του τρόπου με τον οποίο οι άνθρωποι επικοινωνούν πριν σχεδιαστεί ένα σύστημα διαλόγου. Τα συστήματα αυτά πρέπει να κατανοούν πότε να αρχίσουν και να σταματήσουν να μιλούν, να ανιχνεύουν το τέλος των προτάσεων και να αντιλαμβάνονται τις διακοπές ή τις διορθώσεις του χρήστη (Jurafsky & Martin, 2000).

Τα chatbots είναι σχεδιασμένα να προσομοιώνουν ανεπίσημες ανθρώπινες συνομιλίες. Εκπαιδεύονται σε μεγάλα σύνολα δεδομένων, συμπεριλαμβανομένων δεδομένων συνομιλίας από τα κοινωνικά μέσα, τα οποία συχνά απαιτούν φιλτράρισμα για την απομάκρυνση τοξικών στοιχείων. Η εκπαίδευση των chatbots μπορεί να περιλαμβάνει

τεχνικές fine-tuning για τη βελτίωση της ποιότητας της συνομιλίας και την ασφάλεια, αποφεύγοντας επικίνδυνες προτάσεις. Σύγχρονα chatbots χρησιμοποιούν επίσης μηχανές αναζήτησης για να παράγουν απαντήσεις, με το σύστημα να προσομοιώνει μια "ψευδοσυνομιλία" με το εργαλείο αναζήτησης (Jurafsky & Martin, 2000).

Τα διαλογικά συστήματα, όπως τα task-based συστήματα, χρησιμοποιούνται για την επίλυση συγκεκριμένων εργασιών όπως κρατήσεις ή αγορές. Τα περισσότερα εμπορικά συστήματα διαλόγου χρησιμοποιούν αρχιτεκτονικές βασισμένες σε πλαίσια (frame-based), όπου το σύστημα πρέπει να συμπληρώσει πεδία μέσω ερωτήσεων προς τον χρήστη. Η αρχιτεκτονική διαλόγου-state προσθέτει πλουσιότερες αναπαραστάσεις και πιο εξελιγμένους αλγόριθμους για τη διαχείριση των πράξεων διαλόγου του χρήστη και τη δημιουργία απαντήσεων από το σύστημα, με αρχές από την αλληλεπίδραση ανθρώπου-υπολογιστή (HCI) να είναι κρίσιμες στο σχεδιασμό τους (Jurafsky & Martin, 2000).

Εικόνα 2.8 Δυνατότητες Chatbots



4. Αναγνώριση Ομιλίας και Κείμενο σε Ομιλία

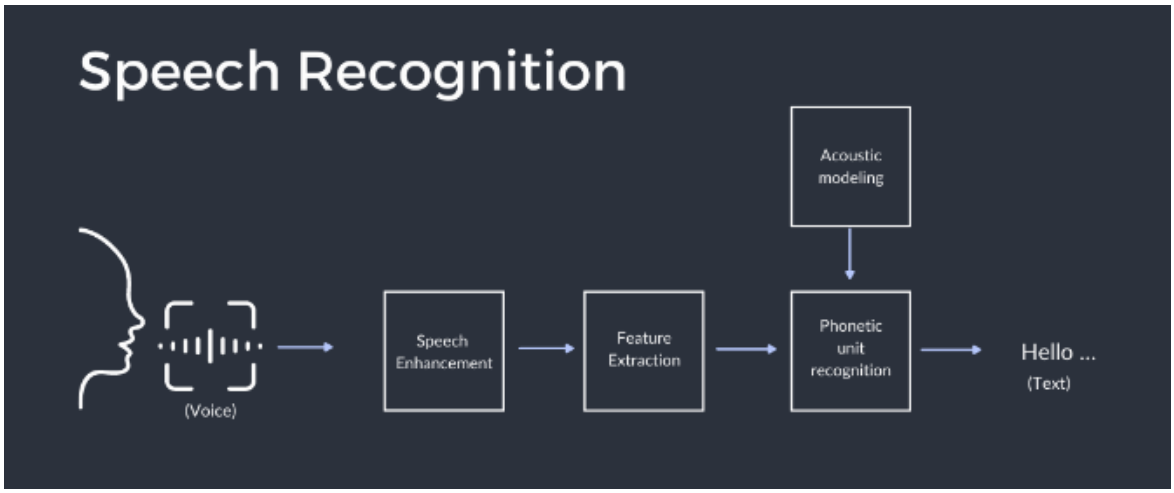
Η αναγνώριση ομιλίας και η μετατροπή κειμένου σε ομιλία είναι δύο από τις πιο σημαντικές εφαρμογές στην επεξεργασία γλώσσας από υπολογιστές. Η κατανόηση της προφορικής γλώσσας ή τουλάχιστον η μεταγραφή των λέξεων σε γραπτό κείμενο ήταν από τους πρώτους στόχους της επεξεργασίας γλώσσας. Αν και η αυτόματη μεταγραφή ομιλίας από οποιονδήποτε ομιλητή σε οποιοδήποτε περιβάλλον παραμένει ένα άλυτο πρόβλημα, η τεχνολογία αναγνώρισης ομιλίας (Automatic Speech Recognition - ASR) έχει εξελιχθεί αρκετά ώστε να είναι πλέον εφαρμόσιμη σε πολλές πρακτικές εργασίες, όπως η επικοινωνία με έξυπνες συσκευές, οι βοηθοί προσωπικής χρήσης, οι εφαρμογές τηλεφωνίας, και τα σύνθετα διαλογικά συστήματα (Jurafsky & Martin, 2000).

Η αναγνώριση ομιλίας (ASR) έχει διάφορες παραμέτρους, όπως το μέγεθος του λεξιλογίου που μπορεί να αναγνωρίσει. Ορισμένα ASR συστήματα επιτυγχάνουν πολύ υψηλή ακρίβεια σε εργασίες με περιορισμένο λεξιλόγιο, όπως η αναγνώριση ναι-όχι ή η αναγνώριση ψηφίων, ενώ πιο ανοιχτές εργασίες, όπως η μεταγραφή συνομιλιών με λεξιλόγια που φτάνουν τις 60.000 λέξεις, είναι πολύ πιο δύσκολες. Η είσοδος σε ένα σύστημα αναγνώρισης ομιλίας είναι μια σειρά από ακουστικά κύματα, τα οποία αναλύονται και μετατρέπονται σε μια φασματική αναπαράσταση, όπως το log mel φάσμα (Jurafsky & Martin, 2000).

Τα συστήματα μετατροπής κειμένου σε ομιλία (TTS) έχουν ως στόχο να μετατρέψουν σειρές γραμμάτων σε ηχητικά κύματα, μια τεχνολογία που είναι σημαντική για εφαρμογές όπως τα διαλογικά συστήματα, τα παιχνίδια και η εκπαίδευση. Όπως και τα συστήματα ASR, τα συστήματα TTS βασίζονται στην αρχιτεκτονική encoder-decoder και αξιολογούνται από ανθρώπινους ακροατές (Jurafsky & Martin, 2000).

Τα μοντέλα αναγνώρισης ομιλίας χρησιμοποιούν συνήθως δύο κοινές προσεγγίσεις: το μοντέλο encoder-decoder με προσοχή και τα μοντέλα που βασίζονται στη συνάρτηση απώλειας CTC. Τα συστήματα TTS απαιτούν πρώτα μια κανονικοποίηση κειμένου για τη διαχείριση αριθμών, συντομογραφιών και άλλων μη τυποποιημένων λέξεων, και αξιολογούνται μέσω τεστ με ανθρώπινους ακροατές, όπως το mean opinion score (MOS) (Jurafsky & Martin, 2000).

Εικόνα 2.9 Αναγνώριση Ομιλίας



2.1.5 Παράδειγμα Ανάλυση Συναισθήματος

Η συγκεκριμένη μελέτη εξάγει ανταγωνιστική γνώση από τα μέσα κοινωνικής δικτύωσης, συγκρίνοντας τις αντιλήψεις των καταναλωτών και τις πωλήσεις δύο αντίπαλων κατασκευαστών smartphone (Kim et al., 2016).

Συγκεκριμένα, διεξήχθη μελέτη περίπτωσης, αναλύοντας 229.948 tweets που αφορούσαν το iPhone6 και το GalaxyS5 για διάστημα τεσσάρων μηνών, χρησιμοποιώντας επεξεργασία φυσικής γλώσσας (NLP), ανάλυση συναισθημάτων με βάση λεξικό και ταξινόμηση προθέσεων αγοράς (Kim et al., 2016).

Στα συμπεράσματα της μελέτης συμπεριλαμβάνεται η διαπίστωση ότι τα στοιχεία από τα κοινωνικά δίκτυα αποκαλύπτουν στρατηγική πληροφόρηση, αντανακλώντας τις αδυναμίες στην ηγεσία της αγοράς και τις προθέσεις των καταναλωτών. Το κενό στις κοινωνικές τοποθετήσεις ευθυγραμμίζεται στενά με το κενό αποστολών, αποδεικνύοντας ότι η δημόσια γνώμη μπορεί να φωτίσει τις διακυμάνσεις στις πωλήσεις. Η επεξεργασία δεδομένων από τις πλατφόρμες κοινωνικής δικτύωσης δίνει τη δυνατότητα στις εταιρείες να εποπτεύουν τις γνώμες των καταναλωτών σχετικά με τα προϊόντα τους και τα προϊόντα των ανταγωνιστών, να εκτιμούν τις πωλήσεις και να τροποποιούν τις στρατηγικές τους ώστε να αντιμετωπίσουν αδυναμίες (Kim et al., 2016).

2.2 Μηχανική Μάθηση

Στην παρούσα εργασία, χρησιμοποιήθηκαν τρεις διαφορετικοί αλγόριθμοι μηχανικής μάθησης, ο Random Forest Classifier ο Multi-layer Perceptron Classifier και ο XGBoost, για την ανάλυση και επεξεργασία των δεδομένων. Παρακάτω παρουσιάζεται μια συνοπτική περιγραφή αυτών των αλγορίθμων, καθώς και τα βασικά χαρακτηριστικά και πλεονεκτήματα που τους καθιστούν κατάλληλους για τις απαιτήσεις της εργασίας.

1. Random Forest Classifier

Ο ταξινομητής Random Forest (RF) είναι μια μέθοδος συνόλου για ταξινόμηση που δημιουργεί πολλαπλά δέντρα απόφασης χρησιμοποιώντας τυχαία επιλεγμένα υποσύνολα δείγματος εκπαίδευσης και χαρακτηριστικών. Συνδυάζοντας τα αποτελέσματα αυτών των δέντρων («τυχαία δάση» - random forests), ο RF ταξινομητής επιτυγχάνει ακριβείς και αξιόπιστες προβλέψεις, και είναι ιδιαίτερα δημοφιλής στον τομέα της απομακρυσμένης ανίχνευσης λόγω της ικανότητάς του να διαχειρίζεται δεδομένα υψηλής διάστασης και πολυδιάστατους συνδυασμούς χαρακτηριστικών (Belgiu & Drăguţ, 2016).

Η λειτουργία των «τυχαίων δασων» - random forests μπορεί να συνοψιστεί ως εξής:

- Δειγματοληψία Bootstrap: Το σύνολο δεδομένων εκπαίδευσης δειγματοληπτείται τυχαία με επαναλήψεις για να δημιουργηθούν πολλά δείγματα bootstrap. Κάθε δείγμα bootstrap χρησιμοποιείται για να εκπαιδεύσει ένα ξεχωριστό δέντρο απόφασης.
- Τυχαία Επιλογή Χαρακτηριστικών: Σε κάθε κόμβο κάθε δέντρου απόφασης, επιλέγεται ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτή η τυχαιότητα βοηθά στη μείωση της συσχέτισης μεταξύ των δέντρων και προλαμβάνει την υπερβολική προσαρμογή (overfitting).
- Ανάπτυξη Δέντρων: Κάθε δέντρο απόφασης αναπτύσσεται στην πλήρη του έκταση χωρίς περικοπή (pruning).
- Πρόβλεψη: Για να γίνουν προβλέψεις, κάθε δέντρο στο δάσος ψηφίζει. Η πιο συχνή κατηγορία μεταξύ όλων των ψήφων επιλέγεται ως η τελική πρόβλεψη.

Τα κύρια χαρακτηριστικά και πλεονεκτήματα των «τυχαίων δασων» - random forests μπορούν να συνοψιστούν ως εξής:

- Μείωση Υπερβολικής Προσαρμογής (Overfitting): Η τυχαιότητα στην επιλογή χαρακτηριστικών και στη δειγματοληψία bootstrap βοηθά στην πρόληψη της υπερβολικής προσαρμογής, καθιστώντας τα τυχαία δάση πιο ανθεκτικά σε θόρυβο.
- Βελτιωμένη Ακρίβεια: Συνδυάζοντας πολλά δέντρα απόφασης, τα τυχαία δάση συχνά επιτυγχάνουν μεγαλύτερη ακρίβεια από τα μεμονωμένα δέντρα.
- Σημασία Χαρακτηριστικών: Τα τυχαία δάση μπορούν να παρέχουν μια εκτίμηση της σημασίας κάθε χαρακτηριστικού στη διαδικασία πρόβλεψης.
- Ανθεκτικότητα: Τα τυχαία δάση είναι σχετικά ανθεκτικά σε ακραία δεδομένα και θόρυβο.
- Ταχύτητα: Σε σύγκριση με μεθόδους όπως η ενίσχυση (boosting), τα τυχαία δάση συνήθως είναι πιο γρήγορα στην εκπαίδευση και πρόβλεψη.

Οι περιορισμοί των «τυχαίων δασων» - random forests μπορούν να συνοψιστούν ως εξής:

- Ερμηνεία: Τα τυχαία δάση μπορεί να είναι δύσκολο να ερμηνευτούν λόγω της πολυπλοκότητας του συνόλου. Η κατανόηση των ακριβών μηχανισμών με τους οποίους το δάσος κάνει προβλέψεις μπορεί να είναι δύσκολη.
- Υπολογιστικό Κόστος: Για πολύ μεγάλα σύνολα δεδομένων ή πολύπλοκα μοντέλα, η εκπαίδευση των τυχαίων δασών μπορεί να είναι υπολογιστικά δαπανηρή.

Συνοπτικά, τα τυχαία δάση είναι μια ισχυρή και ευέλικτη μέθοδος συνόλου που προσφέρει αρκετά πλεονεκτήματα, όπως βελτιωμένη ακρίβεια, ανθεκτικότητα και σημασία χαρακτηριστικών. Χρησιμοποιούνται ευρέως σε διάφορες εφαρμογές, από ταξινόμηση και παλινδρόμηση μέχρι ανίχνευση ανωμαλιών και συστήματα σύστασης (Breiman, 2001).

2. Multi-layer Perceptron Classifier

Ο πολυεπίπεδος αντιληπτής (multilayer perceptron, MLP) είναι ο πιο γνωστός και ευρέως χρησιμοποιούμενος τύπος νευρωνικού δικτύου. Στις περισσότερες περιπτώσεις, τα σήματα μεταδίδονται εντός του δικτύου σε μία μόνο κατεύθυνση: από την είσοδο προς την έξοδο. Σε αυτήν την αρχιτεκτονική, η οποία ονομάζεται "feedforward" (προοδευτική), δεν υπάρχει βρόχος—η έξοδος κάθε νευρώνα δεν επηρεάζει τον ίδιο τον νευρώνα. Τα επίπεδα που δεν είναι άμεσα συνδεδεμένα με το περιβάλλον ονομάζονται κρυφά επίπεδα (hidden layers). Υπάρχουν επίσης δίκτυα ανατροφοδότησης (feedback networks), τα οποία μπορούν να μεταδίδουν σήματα σε αμφότερες τις κατευθύνσεις λόγω αντιδράσεων εντός του δικτύου.

Αυτά τα δίκτυα είναι πολύ ισχυρά και μπορούν να είναι εξαιρετικά περίπλοκα. Είναι δυναμικά και αλλάζουν συνεχώς την κατάστασή τους μέχρι να φτάσουν σε μια κατάσταση ισορροπίας, με την αναζήτηση μιας νέας ισορροπίας να συμβαίνει με κάθε αλλαγή εισόδου (Popescu et al., 2009).

Τα βασικά στοιχεία της δομής δικτύου Multi Layer Perceptron μπορούν να συνοψιστούν ως εξής:

- Επίπεδο Εισόδου: Εδώ εισάγονται τα δεδομένα προς επεξεργασία. Η εισαγωγή πολλών επιπέδων στο MLP καθορίστηκε από την ανάγκη αύξησης της πολυπλοκότητας των περιοχών απόφασης. Όπως αναφέρθηκε, ένας αντιληπτής με ένα μόνο επίπεδο και μία είσοδο παράγει περιοχές απόφασης υπό τη μορφή ημισπατών. Με την προσθήκη ενός άλλου επιπέδου, κάθε νευρώνας λειτουργεί ως ένας τυπικός αντιληπτής για τις εξόδους των νευρώνων του προηγούμενου επιπέδου, επιτρέποντας στην έξοδο του δικτύου να εκτιμήσει κοίλες περιοχές απόφασης, που προκύπτουν από την τομή των ημισπατών που δημιουργούνται από τους νευρώνες. Με τη σειρά του, ένας τριών επιπέδων αντιληπτής μπορεί να δημιουργήσει αυθαίρετες περιοχές απόφασης (Popescu et al., 2009).
- Κρυφά Επίπεδα: Αυτά τα επίπεδα περιλαμβάνουν νευρώνες που εκτελούν μη γραμμικούς μετασχηματισμούς στα δεδομένα. Ο αριθμός των κρυφών επιπέδων και νευρώνων σε αυτά τα επίπεδα μπορεί να ποικίλει και επηρεάζει την απόδοση του δικτύου (Zhao et al., 2015).
- Επίπεδο Εξόδου: Παράγει την τελική πρόβλεψη ή ταξινόμηση βασισμένη στην επεξεργασία που έγινε στα προηγούμενα επίπεδα (Zhao et al., 2015).

Όσον αφορά τη συνάρτηση ενεργοποίησης των νευρώνων, έχει διαπιστωθεί ότι τα πολυεπίπεδα δίκτυα δεν προσφέρουν αύξηση της υπολογιστικής δύναμης σε σύγκριση με δίκτυα ενός επιπέδου, αν οι συναρτήσεις ενεργοποίησης είναι γραμμικές. Αυτό συμβαίνει γιατί μια γραμμική συνάρτηση γραμμικών συναρτήσεων παραμένει γραμμική (Popescu et al., 2009).

2.1 Εκπαίδευση και Επικύρωση Δεδομένων

Κατά τη διάρκεια της εκπαίδευσης, χρησιμοποιούνται επικυρωμένα δεδομένα που δεν συμμετέχουν στη διαδικασία εκπαίδευσης για να ελέγχεται η γενίκευση του μοντέλου. Η

εκπαίδευση σταματά όταν η απόδοση στο σύνολο επικύρωσης δεν βελτιώνεται περαιτέρω, προκειμένου να αποφευχθεί η υπερβολική εξάσκηση. Ο ταξινομητής MLP είναι ιδιαίτερα χρήσιμος για προβλήματα όπου απαιτούνται σύνθετες σχέσεις μεταξύ χαρακτηριστικών, όπως στην πιστοποίηση πιστώσεων και σε άλλες εφαρμογές μηχανικής μάθησης (Zhao et al., 2015).

3. XGBoost

Ο XGBoost, που σημαίνει eXtreme Gradient Boosting, είναι ένας προηγμένος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται ευρέως για την επίλυση διαφόρων προβλημάτων δεδομένων. Βασίζεται στο πλαίσιο του gradient boosting και υλοποιεί εξελιγμένα αλγορίθμους μάθησης δένδρων. Ο XGBoost προσφέρει μία τεχνική ενίσχυσης δένδρων (tree boosting) που είναι γνωστή επίσης ως GBDT (Gradient Boosted Decision Trees) ή GBM (Gradient Boosting Machines). Αυτή η τεχνική επιτρέπει την ταχεία και ακριβή επίλυση πολλών προβλημάτων που σχετίζονται με δεδομένα, κάνοντάς την ιδανική για χρήση σε μεγάλες κατανεμημένες υποδομές, όπως οι Hadoop, SGE, και MPI, και σε προβλήματα που περιλαμβάνουν δισεκατομμύρια παραδείγματα (Aydin & Ozturk, 2021).

Η διαδικασία του αλγορίθμου gradient boosting που ακολουθεί ο XGBoost περιλαμβάνει τα εξής βήματα:

- Δημιουργία ενός αρχικού μοντέλου, συνήθως ένα απλό δέντρο απόφασης, που εκτιμά τις αρχικές προβλέψεις.
- Υπολογισμός των σφαλμάτων ή των υπολειμμάτων (residuals) μεταξύ των πραγματικών τιμών και των προβλέψεων του αρχικού μοντέλου.
- Δημιουργία ενός νέου δέντρου απόφασης που εστιάζει στην πρόβλεψη των υπολειμμάτων.
- Προσθήκη του νέου δέντρου στο υπάρχον μοντέλο για βελτίωση των προβλέψεων, συνδυάζοντας τις προβλέψεις του νέου δέντρου με τις προβλέψεις του προηγούμενου μοντέλου.
- Επαναλαμβάνονται τα βήματα 2 έως 4 για πολλά δέντρα, βελτιώνοντας συνεχώς την ακρίβεια του μοντέλου.

Η ικανότητα του XGBoost να αντιμετωπίζει προβλήματα με εκατομμύρια παραδείγματα και η ευελιξία του σε κατανεμημένα περιβάλλοντα έχουν συμβάλλει στην υπεροχή του έναντι άλλων μοντέλων σε πολλές περιπτώσεις (Aydin & Ozturk, 2021).

3.1 Εφαρμογές και Πλεονεκτήματα του XGBoost

Ο XGBoost έχει αναδειχθεί ως ένας από τους πιο δημοφιλείς αλγορίθμους μηχανικής μάθησης, κερδίζοντας πολλές διαγωνισμούς δεδομένων στο Kaggle. Παρέχει ικανοποιητικά αποτελέσματα σε μια ποικιλία εφαρμογών, όπως:

- Πρόβλεψη ασθενειών
- Αναγνώριση τύπων καυσίμου ντίζελ
- Εκτίμηση της προόδου μηχανών διάτρησης τούνελ
- Πρόβλεψη ηλεκτρικής αντίστασης σκυροδέματος για παρακολούθηση της δομικής υγείας
- Ανάλυση συναισθημάτων σε κριτικές ξενοδοχείων
- Κατηγοριοποίηση άστρων/γαλαξιών
- Πρόβλεψη τραυματισμών επιβατών οχημάτων σε σηματοδοτούμενες διασταυρώσεις (Aydin & Ozturk, 2021).

Τα πλεονεκτήματα του αλγορίθμου XGBoost μπορούν να συνοψιστούν ως εξής:

- Μπορεί να διαχειριστεί ελλιπή δεδομένα χωρίς πρόσθετη επεξεργασία, γεγονός που το καθιστά ιδιαίτερα χρήσιμο σε πραγματικά σενάρια όπου η απουσία δεδομένων είναι συχνή.
- Δεν απαιτεί προκαταρκτική κλίμακα ή κανονικοποίηση των δεδομένων.
- Εφαρμόζει παράλληλη επεξεργασία για γρηγορότερη εκπαίδευση και προβλέψεις.
- Περιλαμβάνει μηχανισμούς κανονικοποίησης για την αποφυγή της υπερπροσαρμογής.
- Παρέχει εξαιρετικά ακριβείς προβλέψεις σε πολλές εφαρμογές.

Η ικανότητα του XGBoost να διαχειρίζεται ελλιπή δεδομένα και η γενική του ευελιξία συμβάλλουν στην ευρεία χρήση του, καθιστώντας τον ένα πολύτιμο εργαλείο στη μηχανική μάθηση (Aydin & Ozturk, 2021).

3. Μεθοδολογία και Πεδίο (Ποδόσφαιρο και Μέσα Κοινωνικής Δικτύωσης)

3.1 Μεθοδολογία

Ορίσαμε ένα κοινό χρονικό πλαίσιο που εξυπηρετεί και τις δύο βασικές αναλύσεις μας, το οποίο ορίσαμε ως το κύριο αντικείμενο μελέτης. Αυτό το χρονικό πλαίσιο επιλέχθηκε προσεκτικά ώστε να επιτρέπει τη συνδυασμένη εξέταση δύο κρίσιμων πτυχών της έρευνάς μας:

Δημιουργία και πειραματισμός με μοντέλα πρόβλεψης: Σε αυτό το στάδιο, κατασκευάσαμε τα μοντέλα μας και προχωρήσαμε σε πειραματισμούς με διάφορες παραμέτρους, προκειμένου να συγκρίνουμε την απόδοση των διαφορετικών προσεγγίσεων.

Ανάλυση συναισθήματος - διερεύνηση της συσχέτισης με την απόδοση των ομάδων: Η δεύτερη και κύρια ιδέα της διπλωματικής μας ήταν να αναλύσουμε το συναίσθημα των οπαδών όπως αυτό εκφράζεται μέσω των tweets και να εξετάσουμε πώς αυτό σχετίζεται με την αγωνιστική φόρμα των ομάδων. Αυτή η ανάλυση μας επιτρέπει να διερευνήσουμε τη δυναμική αλληλεπίδραση μεταξύ της απόδοσης της ομάδας και της αντίδρασης των οπαδών στο διαδίκτυο.

Η επιλογή του κοινού χρονικού πλαισίου για τις δύο αυτές αναλύσεις εξασφαλίζει τη συνέπεια και τη σύγκρισή των αποτελεσμάτων, επιτρέποντάς μας να εξάγουμε πιο αξιόπιστα συμπεράσματα σχετικά με τις σχέσεις μεταξύ της φόρμας των ομάδων και της συμπεριφοράς των οπαδών.

1. Δημιουργία και πειραματισμός με μοντέλα πρόβλεψης

Η πρώτη και θεμελιώδης ενέργεια ήταν η επιλογή ενός αξιόπιστου και κατάλληλου συνόλου δεδομένων που περιέχει τις απαραίτητες πληροφορίες για την ανάλυσή μας. Τα δεδομένα που χρησιμοποιήσαμε αφορούσαν τα αποτελέσματα ποδοσφαιρικών αγώνων, όπου κάθε αγώνας κατέληγε σε μία από τις τρεις πιθανές εκβάσεις: νίκη της γηπεδούχου ομάδας, νίκη της φιλοξενούμενης ομάδας ή ισοπαλία. Αυτές οι τρεις εκβάσεις απεικονίστηκαν αρχικά με τα σύμβολα 'H' (Home win), 'A' (Away win), και 'D' (Draw).

Για να διευκολύνουμε την εφαρμογή των μοντέλων μηχανικής μάθησης καθώς και το στόχο της ανάλυσης, πραγματοποιήσαμε μια διαδικασία μετασχηματισμού αυτών των τριών κατηγοριών σε δυαδικές τιμές ('1', '0'). Η τιμή '1' ανατέθηκε στη νίκη της γηπεδούχου ομάδας ('H'), ενώ οι άλλες δύο εκβάσεις ('A' και 'D') ανατέθηκαν στην τιμή '0'. Αυτός ο δυαδικός διαχωρισμός μας επιτρέπει να αναλύσουμε αν η γηπεδούχος ομάδα κερδίζει ή όχι, ενσωματώνοντας αυτές τις πληροφορίες σε ένα ενιαίο πλαίσιο.

Δεδομένου ότι τα μοντέλα μηχανικής μάθησης απαιτούν αριθμητικές εισόδους, προχωρήσαμε στη μετατροπή των ονομαστικών χαρακτηριστικών σε αριθμητικές τιμές. Συγκεκριμένα, κάθε ομάδα αντιπροσωπεύθηκε από έναν μοναδικό ακέραιο αριθμό, καθιστώντας έτσι τα δεδομένα μας κατάλληλα για την εισαγωγή τους στα μοντέλα.

Μετά την προεπεξεργασία των δεδομένων, τα διαιρέσαμε σε δύο υποσύνολα με βάση το χρονικό διάστημα των αγώνων. Το μεγαλύτερο χρονικό διάστημα αποτέλεσε το σύνολο εκπαίδευσης (training set), το οποίο περιλαμβάνει 709 αγώνες, ενώ το μικρότερο διάστημα αποτέλεσε το σύνολο επαλήθευσης (testing set), το οποίο περιλαμβάνει 51 αγώνες. Αυτός ο διαχωρισμός διασφαλίζει την ανεξαρτησία των δεδομένων που χρησιμοποιούνται για την εκπαίδευση από αυτά που χρησιμοποιούνται για την αξιολόγηση της απόδοσης των μοντέλων.

Στη συνέχεια, καθορίσαμε τη λίστα των μεταβλητών (predictors) που χρησιμοποιούνται ως είσοδοι στο μοντέλο για την πρόβλεψη του αποτελέσματος ενός μελλοντικού αγώνα.

Οι αρχικοί παράγοντες που επιλέξαμε περιλάμβαναν τους εξής:

- Δείκτης γηπεδούχου ομάδας : Ένας ακέραιος αριθμός που αναπαριστά την ταυτότητα της ομάδας που αγωνίζεται ως γηπεδούχος.
- Δείκτης φιλοξενούμενης ομάδας : Ένας ακέραιος αριθμός που αναπαριστά την ταυτότητα της ομάδας που αγωνίζεται ως φιλοξενούμενη.
- Δείκτης ημέρας : Ένας ακέραιος αριθμός που αναπαριστά την ημέρα διεξαγωγής του αγώνα.

Αρχικά, η ανάλυση μας περιορίστηκε σε αυτούς τους τρεις δείκτες. Ωστόσο, στη συνέχεια, αυξήσαμε τον αριθμό των μεταβλητών για να εξετάσουμε πώς η προσθήκη επιπλέον

πληροφοριών επηρεάζει την απόδοση των μοντέλων μας, επιδιώκοντας την επίτευξη καλύτερων και πιο ακριβών προβλέψεων.

Δημιουργήσαμε τα μοντέλα και στη συνέχεια προχωρήσαμε σε εκτεταμένο πειραματισμό, με στόχο την εύρεση του συνόλου των παραμέτρων που θα μπορούσαν να αποδώσουν τη μεγαλύτερη δυνατή ακρίβεια στο μοντέλο. Για την ενίσχυση της πρόβλεψης, αποφασίσαμε να επεκτείνουμε τη λίστα των predictors, προσθέτοντας επιπλέον πληροφορίες.

Συγκεκριμένα, αναπτύξαμε μια συνάρτηση η οποία αξιολογεί αν μια ομάδα βρίσκεται σε καλή αγωνιστική φόρμα ή όχι. Η συνάρτηση αυτή ενσωματώνει δεδομένα σχετικά με την πρόσφατη απόδοση της ομάδας και παράγει έναν δείκτη φόρμας, τον οποίο προσθέσαμε ως νέο predictor στο μοντέλο μας. Στόχος αυτής της προσθήκης ήταν να εξετάσουμε αν η συμπερίληψη της αγωνιστικής φόρμας της ομάδας μπορεί να βελτιώσει την ακρίβεια των προβλέψεων του μοντέλου μας.

Με αυτόν τον τρόπο, επιδιώξαμε να ενισχύσουμε την απόδοση των μοντέλων, διερευνώντας αν η πληροφορία σχετικά με τη φόρμα της ομάδας μπορεί να προσφέρει πρόσθετη αξία στην πρόβλεψη των αποτελεσμάτων, βελτιώνοντας την ακρίβεια και τη γενική αποδοτικότητα τω μοντέλων μας.

2. Ανάλυση συναισθήματος

Το δεύτερο και ιδιαίτερα σημαντικό σύνολο δεδομένων που απαιτήθηκε για την ανάλυσή μας ήταν τα tweets που συλλέχθηκαν από ένα συγκεκριμένο υποσύνολο του χρονικού πλαισίου που ορίσαμε προηγουμένως. Αφού συγκεντρώσαμε τα tweets, προχωρήσαμε σε λεπτομερή επεξεργασία κάθε ενός από αυτά με στόχο να βελτιώσουμε την ακρίβεια της ανάλυσης κειμένου.

Αρχικά, εντοπίσαμε και διαχειριστήκαμε τυχόν κενές τιμές στα δεδομένα μας, διαγράφοντας όσες από αυτές δεν μπορούσαν να προσφέρουν χρήσιμες πληροφορίες. Στη συνέχεια, δημιουργήσαμε μια λίστα από stopwords – λέξεις που δεν προσθέτουν ουσιαστικό νόημα στην ανάλυση του συναισθήματος, καθώς θεωρούνται συναισθηματικά ουδέτερες. Αφαιρέσαμε αυτές τις λέξεις από τα κείμενα των tweets, καθώς και όλους τους μη αλφαβητικούς χαρακτήρες, προκειμένου να εξασφαλίσουμε την καθαρότητα των δεδομένων μας.

Για την ανάλυση του συναισθήματος, επιλέξαμε το μοντέλο RoBERTa, το οποίο είναι εκπαιδευμένο σε εκατομμύρια tweets και είναι ιδανικό για την αναγνώριση συναισθηματικών αποχρώσεων σε κείμενα. Εφαρμόσαμε το μοντέλο RoBERTa σε κάθε tweet, με αποτέλεσμα να παραχθούν τρεις στήλες, καθεμία από τις οποίες αντιπροσωπεύει ένα σκορ μεταξύ 0 και 1. Οι στήλες αυτές εκφράζουν την πιθανότητα το tweet να είναι θετικό, αρνητικό ή ουδέτερο.

Για να ενισχύσουμε την ποιότητα των δεδομένων και την αξιοπιστία της ανάλυσής μας, περιορίσαμε το dataset κρατώντας μόνο τα tweets όπου η συναισθηματική δύναμη, είτε θετική είτε αρνητική, ήταν ίση ή μεγαλύτερη του 0,3. Με αυτόν τον τρόπο, εστιάσαμε σε tweets με σαφή και έντονη συναισθηματική φόρτιση.

Στη συνέχεια, δημιουργήσαμε μια νέα στήλη, την οποία ονομάσαμε Score, η οποία αναπαριστά τη συνολική συναισθηματική βαθμολογία του κάθε tweet. Συγκεκριμένα, η τιμή της στήλης αυτής είναι -1 όταν το αρνητικό συναίσθημα υπερισχύει του θετικού, και 1 όταν το θετικό συναίσθημα είναι ισχυρότερο. Αυτό μας επιτρέπει να αναγνωρίσουμε και να αναλύσουμε εύκολα τα tweets με βάση την κυρίαρχη συναισθηματική τους φόρτιση.

Επιπλέον, αποφασίσαμε να αξιοποιήσουμε και μια επιπλέον στήλη, το Retweets count, προσδίδοντας αντίστοιχα βαθμολογίες 1 και -1 όπως προαναφέρθηκε, ανάλογα με την κυρίαρχη συναισθηματική κατεύθυνση του tweet. Με αυτή τη διαδικασία, κάθε tweet αντιστοιχίζεται με έναν αριθμητικό δείκτη, ο οποίος εκφράζει πόσο θετικά ή αρνητικά αντιδρούν οι χρήστες του Twitter για την ομάδα τους σε μια συγκεκριμένη ημερομηνία.

Το τελικό αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία μιας συνολικής βαθμολογίας Score για κάθε συνδυασμό ημερομηνιών και ομάδων. Στη συνέχεια, περιορίσαμε το dataset ώστε να περιλαμβάνει μόνο τις τιμές από την ημέρα του αγώνα και τις δύο προηγούμενες ημέρες, εστιάζοντας στα δεδομένα που είναι άμεσα σχετιζόμενα με την ημέρα του αγώνα. Τέλος, δημιουργήσαμε δύο τελικά datasets: το ένα αντιστοιχεί στην ομάδα που αγωνίζεται ως φιλοξενούμενη και το άλλο στην ομάδα που αγωνίζεται ως γηπεδούχος, εξασφαλίζοντας ότι η ανάλυση θα πραγματοποιηθεί με σαφή και διακριτό τρόπο για κάθε περίπτωση.

3.2 Ποδόσφαιρο και Μέσα Κοινωνικής Δικτύωσης

1. Σχέση Ποδοσφαίρου με Μέσα Κοινωνικής δικτύωσης

Η σχέση μεταξύ ποδοσφαίρου και μέσων κοινωνικής δικτύωσης είναι ιδιαίτερα δυναμική και συνεχώς εξελίσσεται, αναδεικνύοντας μια σειρά από μοναδικές ιδιαιτερότητες. Στον χώρο των αθλητικών μέσων κοινωνικής δικτύωσης, οι πλατφόρμες αυτές λειτουργούν όχι μόνο ως πεδίο ενημέρωσης, αλλά και ως χώρος συλλογικής εμπειρίας και αλληλεπίδρασης. Οι οπαδοί, μέσα από αυτές τις πλατφόρμες, βρίσκουν ένα κοινό έδαφος για να μοιραστούν τη χαρά των επιτυχιών της ομάδας τους, να εκφράσουν τα συναισθήματά τους, και να συζητήσουν για τα αγαπημένα τους αθλήματα.

Ένας από τους πιο χαρακτηριστικούς τρόπους έκφρασης στα μέσα κοινωνικής δικτύωσης είναι τα hashtags. Αυτά τα σύμβολα όχι μόνο συνδέουν τους οπαδούς γύρω από συγκεκριμένα αθλητικά γεγονότα, αλλά και δημιουργούν μια αίσθηση συμμετοχής σε μια ζωντανή, κοινή εμπειρία. Μέσα από τα hashtags, οι οπαδοί μπορούν να συμμετέχουν σε συζητήσεις σε πραγματικό χρόνο, να μοιράζονται απόψεις, να σχολιάζουν αγώνες, και να εκφράζουν τη στήριξή τους στις ομάδες τους.

Η γλώσσα που χρησιμοποιείται στις αθλητικές σελίδες κοινωνικής δικτύωσης είναι επίσης ιδιαίτερη. Συχνά περιλαμβάνει έντονα συναισθηματικά φορτισμένες εκφράσεις, ενθουσιασμό, αλλά και απογοήτευση, ενώ το ύφος μπορεί να είναι απόλυτα ειλικρινές και αυθόρμητο. Οι οπαδοί χρησιμοποιούν συχνά φράσεις όπως "πάμε για τη νίκη!", "δεν μας σταματάει τίποτα!", ή "ήρωες!", οι οποίες ενισχύουν το αίσθημα της κοινότητας και της συλλογικής ταυτότητας.

Επιπλέον, οι αθλητικές ομάδες αξιοποιούν τα μέσα κοινωνικής δικτύωσης ως πολύτιμο εργαλείο μάρκετινγκ και επικοινωνίας με τους οπαδούς τους. Συχνά, ανακοινώνουν ειδήσεις, αποτελέσματα αγώνων, και άλλες σημαντικές πληροφορίες μέσω πλατφορμών όπως το Twitter. Με αυτόν τον τρόπο, εξασφαλίζεται ότι οι φίλαθλοι έχουν πρόσβαση σε πληροφορίες που ενδεχομένως δεν θα ήταν διαθέσιμες μέσω άλλων μέσων, εμπλουτίζοντας έτσι την αθλητική τους εμπειρία.

Τα μέσα κοινωνικής δικτύωσης προσφέρουν επίσης στις μικρότερες, λιγότερο προβεβλημένες επαγγελματικές αθλητικές ομάδες και πρωταθλήματα την ευκαιρία να προωθηθούν και να κερδίσουν την προσοχή που ίσως να μην έβρισκαν μέσω των παραδοσιακών μέσων ενημέρωσης. Αυτές οι πλατφόρμες γίνονται ένα εργαλείο

επικοινωνίας που διευκολύνει την αλληλεπίδραση μεταξύ των φιλάθλων, ενισχύοντας έτσι τη σύνδεσή τους με την ομάδα.

Επίσης τα μέσα κοινωνικής δικτύωσης επιτρέπουν την καταγραφή και ανάλυση δεδομένων συνομιλίας, τα οποία μπορούν να περιλαμβάνουν πληροφορίες όπως το αναγνωριστικό παιχνιδιού, το αναγνωριστικό χρήστη, το κείμενο του μηνύματος, την αγαπημένη ομάδα του χρήστη και τη χρονική στιγμή της δημοσίευσης. Αυτά τα δεδομένα αποτελούν τη βάση για περαιτέρω αναλύσεις που μπορεί να αποκαλύψουν σημαντικές πληροφορίες σχετικά με τις τάσεις, τις αντιδράσεις των οπαδών και τις δυναμικές της κοινωνικής αλληλεπίδρασης γύρω από το ποδόσφαιρο.

Αναμφίβολα, τα μέσα κοινωνικής δικτύωσης παίζουν συμπληρωματικό ρόλο με τα παραδοσιακά μέσα ενημέρωσης και βοηθούν την ομάδα να διαδώσει τις πληροφορίες σε μεγαλύτερο βαθμό (Himelboim I., & Espina C. 2017) .

Η χρήση των μέσων κοινωνικής δικτύωσης έχει επιτρέψει στους ποδοσφαιρικούς συλλόγους να αναπτύξουν μια αμφίδρομη σχέση με τους οπαδούς τους. Σύμφωνα με τον (Pacak, 2020) πλατφόρμες όπως το Facebook και το Twitter προσφέρουν στους οπαδούς έναν άμεσο σύνδεσμο με τον σύλλογο, ακόμα και μεταξύ των αγώνων. Αυτές οι πλατφόρμες δίνουν στους συλλόγους τη δυνατότητα να επιβραβεύουν την αφοσίωση των οπαδών με αποκλειστικές προσφορές, διαγωνισμούς, καθώς και άμεσες ενημερώσεις από το γήπεδο και το προπονητικό κέντρο. Επισημαίνεται ότι οι αφοσιωμένοι οπαδοί είναι πιο πιθανό να μοιραστούν πληροφορίες και προσφορές, να επισκεφθούν την επίσημη ιστοσελίδα του συλλόγου πιο συχνά, και να συμβάλουν στην αύξηση της συμμετοχής και της αλληλεπίδρασης των οπαδών (Pacak, 2020).

Όταν αξιοποιούνται σωστά, τα μέσα κοινωνικής δικτύωσης μπορούν να βοηθήσουν τις ποδοσφαιρικές οργανώσεις να ενισχύσουν τη συμμετοχή των οπαδών, να αυξήσουν την κίνηση προς την επίσημη ιστοσελίδα τους, και ακόμη να αναπτύξουν προγράμματα χορηγίας, αυξάνοντας έτσι τα έσοδα (Cheong and Cheong 2011).

2. Εκφράσεις - Ορολογία

Το πεδίο του ποδοσφαίρου, ειδικά όπως εμφανίζεται στα μέσα κοινωνικής δικτύωσης, χαρακτηρίζεται από ένα ιδιαίτερο λεξιλόγιο και ύφος που αντικατοπτρίζουν την ένταση, το πάθος και τη συλλογική εμπειρία των οπαδών. Το περιεχόμενο που αφορά το ποδόσφαιρο

στα μέσα αυτά περιλαμβάνει όχι μόνο πληροφορίες για αγώνες, αποτελέσματα και μεταγραφές, αλλά και εκφράσεις ενθουσιασμού, απογοήτευσης, ή και στήριξης προς τις ομάδες.

Το λεξιλόγιο του ποδοσφαίρου περιλαμβάνει ειδική ορολογία που είναι γνώριμη στους οπαδούς, όπως "γκολ", "πέναλτι", "κόρνερ", "αποβολή", "ασίστ", και "ενδεκάδα". Αυτές οι λέξεις χρησιμοποιούνται συχνά σε συνομιλίες που αφορούν αγωνιστικά γεγονότα και εξελίξεις. Παράλληλα, οι εκφράσεις που σχετίζονται με την απόδοση των παικτών και την τακτική των ομάδων, όπως "άμυνα", "επιθετική διάταξη", "κατοχή μπάλας" και "πρέσινγκ", κυριαρχούν στο περιεχόμενο των συζητήσεων.

Στα μέσα κοινωνικής δικτύωσης, οι οπαδοί χρησιμοποιούν εκφράσεις που συχνά αντικατοπτρίζουν τα έντονα συναισθήματά τους, όπως "πάμε γερά!", "όλοι μαζί για τη νίκη!", "δεν μας σταματάει κανείς!", και "έχουμε την καλύτερη ομάδα!". Το ύφος των συνομιλιών είναι συχνά αυθόρμητο, γεμάτο ενέργεια και πάθος, ενώ μπορεί να περιέχει και χιούμορ, ειρωνεία ή ακόμα και σαρκασμό, ανάλογα με την πορεία του αγώνα ή την απόδοση της ομάδας.

Οι ιδιαιτερότητες του ποδοσφαιρικού πεδίου στα μέσα κοινωνικής δικτύωσης περιλαμβάνουν επίσης τη χρήση συμβόλων και hashtags που συνδέουν τους οπαδούς σε πραγματικό χρόνο. Τα hashtags όπως #Goal, #Victory, ή #FootballFever ενώνουν τους χρήστες, επιτρέποντάς τους να συμμετέχουν σε παγκόσμιες συζητήσεις και να εκφράζουν την υποστήριξή τους. Επίσης, οι οπαδοί συχνά δημιουργούν συνθήματα και memes που εκφράζουν την αγάπη τους για την ομάδα ή σχολιάζουν με χιούμορ τα δρώμενα, κάνοντας το περιεχόμενο πιο πλούσιο και διαδραστικό.

Συνολικά, το περιεχόμενο και το λεξιλόγιο του ποδοσφαίρου στα μέσα κοινωνικής δικτύωσης είναι βαθιά συνυφασμένα με την κουλτούρα του αθλήματος και τη συλλογική εμπειρία των οπαδών, προσφέροντας ένα ζωντανό, διαδραστικό πεδίο συζήτησης και συμμετοχής.

3. Τόνος και Ύφος Οπαδών

Τα σχόλια των οπαδών σε πλατφόρμες όπως το Twitter είναι συχνά γεμάτα πάθος και έντονα συναισθήματα, αντανακλώντας την αφοσίωσή τους στην ομάδα τους. Ο τόνος των tweets μπορεί να ποικίλλει ευρέως, ανάλογα με το αποτέλεσμα του αγώνα, την απόδοση

της ομάδας ή ακόμα και συγκεκριμένων παικτών. Παρακάτω είναι μερικά χαρακτηριστικά του τόνου και του ύφους αυτών των σχολίων:

- Ενθουσιασμός και Υποστήριξη: Όταν η ομάδα κερδίζει ή σημειώνει ένα γκολ, ο τόνος είναι συχνά γεμάτος ενθουσιασμό και περηφάνια. Οι οπαδοί χρησιμοποιούν εκφράσεις όπως "Τεράστια νίκη!", "Μπράβο παλικάρια!", "Δεν μας σταματάει κανείς!". Το ύφος είναι θετικό, γεμάτο ενέργεια, και πολλές φορές περιλαμβάνει emojis όπως 🔥, 💪, ή 🙌 για να ενισχύσει το συναίσθημα.
- Απογοήτευση και Κριτική: Σε περιπτώσεις ήττας ή κακής απόδοσης, ο τόνος γίνεται πιο αρνητικός και κριτικός. Οι οπαδοί μπορεί να εκφράζουν απογοήτευση με φράσεις όπως "Τι κάνετε εκεί;", "Απαράδεκτοι!", "Πάλι τα ίδια!". Το ύφος είναι πολλές φορές επιθετικό, με σκληρή γλώσσα, και μπορεί να εμπεριέχει και ειρωνεία ή σαρκασμό.
- Ειρωνεία και Σαρκασμός: Όταν οι οπαδοί είναι απογοητευμένοι ή θέλουν να εκφράσουν δυσαρέσκεια, συχνά χρησιμοποιούν ειρωνεία ή σαρκασμό. Tweets όπως "Σπουδαία δουλειά σήμερα... 👎", "Καλύτερα να μην είχατε βγει στο γήπεδο!", χρησιμοποιούν την αντίθεση για να εκφράσουν την αντίθετη πραγματικότητα από αυτήν που περιγράφουν, ενισχύοντας το αρνητικό συναίσθημα.
- Χιούμορ και Memes: Συχνά, οι οπαδοί χρησιμοποιούν χιούμορ για να διαχειριστούν την απογοήτευσή τους ή απλά για να ψυχαγωγηθούν μεταξύ τους. Τα σχόλια μπορεί να περιλαμβάνουν αστείες φράσεις ή memes που διακωμωδούν μια κατάσταση, έναν αγώνα ή ακόμα και τους ίδιους τους παίκτες. Το ύφος εδώ είναι ελαφρύ, διασκεδαστικό, και συχνά περιλαμβάνει viral περιεχόμενο.
- Ενότητα και Κάλεσμα σε Δράση: Ορισμένα tweets ενθαρρύνουν τη συσπείρωση των οπαδών και την υποστήριξη της ομάδας ανεξαρτήτως αποτελέσματος. Εκφράσεις όπως "Μαζί στα εύκολα και στα δύσκολα!", "Πάμε γερά, η επόμενη νίκη είναι δική μας!" ενισχύουν το αίσθημα της κοινότητας και της πίστης στην ομάδα. Το ύφος εδώ είναι ενωτικό και θετικό.

Σε γενικές γραμμές, ο τόνος και το ύφος των σχολίων των οπαδών στο Twitter είναι έντονα συναισθηματικός και καθοδηγείται από την τρέχουσα κατάσταση και την απόδοση της ομάδας τους. Η αυθόρμητη και στιγμιαία φύση των tweets επιτρέπει στους οπαδούς να

εκφράσουν άμεσα τα συναισθήματά τους, κάτι που κάνει το περιεχόμενο εξαιρετικά δυναμικό και ποικίλο.

4. Σύγκριση Μοντέλων Μηχανικής Μάθησης

4.1 Διαδικασία Σύγκρισης Μοντέλων Μηχανικής Μάθησης

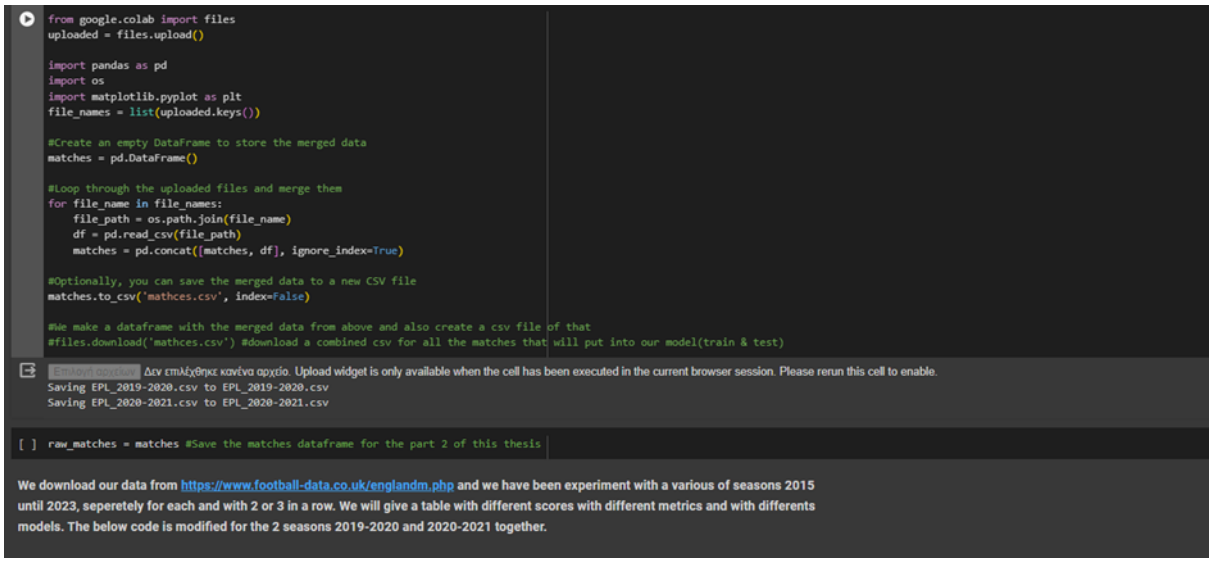
Το πρώτο στάδιο της παρούσας ανάλυσης και επεξεργασίας είναι να συγκρίνουμε διαφορετικά μοντέλα μηχανικής μάθησης, χρησιμοποιώντας διάφορες μετρικές και συναρτήσεις, βασισμένοι σε ιστορικά δεδομένα ποδοσφαίρου από το Αγγλικό Πρωτάθλημα (English Premier League). Αρχικά, θα δημιουργήσουμε και θα αξιολογήσουμε τα μοντέλα μας για την ακρίβεια και την αποδοτικότητά τους, χρησιμοποιώντας ποικίλες μετρικές και τεχνικές. Στη συνέχεια, θα εργαστούμε πάνω στη βελτίωση της απόδοσής τους, χρησιμοποιώντας ένα σύνολο διαφόρων συναρτήσεων τεχνικών και παραμέτρων ως είσοδο. Η επιλογή του συνόλου αυτού πραγματοποιήθηκε με κριτήριο τη μέγιστη δυνατή βελτίωση των προβλέψεων, επιδιώκοντας υψηλή ακρίβεια και αξιοπιστία στα αποτελέσματα (Baboota & Kaur, 2019).

Η διαδικασία της σύγκρισης των μοντέλων μηχανικής μάθησης μπορεί να συνοψιστεί σε τέσσερα γενικά βήματα, συγκεκριμένα, την επιλογή (1) και την επισκόπηση (2) του συνόλου δεδομένων που θα επεξεργαστεί (Βήμα 1 και Βήμα 2 αντίστοιχα), την αξιολόγηση των εμφανιζόμενων πληροφοριών του συνόλου δεδομένων (Βήμα 3) και την χρήση και ενεργοποίηση μοντέλων Μηχανικής Μάθησης (Βήμα 4). για την δημιουργία μοντέλων μηχανικής μάθησης για την επεξεργασία των δεδομένων του Αγγλικού Πρωταθλήματος (English Premier League) εξετάζονται οι αλγόριθμοι Random Forest, MLPClassifier και XGBoost.

1. Επιλογή Συνόλου Δεδομένων

Το πρώτο βήμα στην ανάλυση και επεξεργασία που παρουσιάζεται εδώ είναι η επιλογή συνόλου δεδομένων που περιέχει τις πληροφορίες στις οποίες στοχεύει η παρούσα προσέγγιση. Το παρακάτω site https://www.football-data.co.uk/englandm.php είναι ευρέως γνωστό για την πληθώρα πληροφοριών που παρέχει σχετικά με το ποδόσφαιρο και, πιο συγκεκριμένα, τους αγώνες του Αγγλικού Πρωταθλήματος (English Premier League). Πρόκειται για τα δεδομένα από τις χρονιές 2019-2020 και 2020-2021 (Εικόνα 4.1).

Εικόνα 4.1 Επιλογή Ποδοσφαιρικών Σεζόν



2. Επισκόπηση Συνόλου Δεδομένων

Στη συνέχεια, πραγματοποιούμε μια επισκόπηση του συνόλου των δεδομένων για να λάβουμε αρχικές πληροφορίες για τα χαρακτηριστικά και την δομή τους και για τον τρόπο με τον οποίο μπορούν να αξιοποιηθούν στην ανάλυση και επεξεργασία τους. Επίσης, στην επισκόπηση των δεδομένων αυτών διαπιστώνουμε ότι ο αριθμός των αγώνων που πραγματοποιήθηκαν από το 2019 έως το 2021 είναι 760.

3. Αξιολόγηση Εμφανιζόμενων Πληροφοριών του Συνόλου Δεδομένων

Στο σύνολο δεδομένων προς ανάλυση και επεξεργασία είναι δυνατή η εμφάνιση διαφορετικών ειδών πληροφοριών προς αξιολόγηση. Οι πληροφορίες αυτές κυμαίνονται από πληροφορίες γενικής φύσεως (Εικόνα 4.2) μέχρι ειδικές πληροφορίες όπως, εμφανίζονται στα παραδείγματα που ακολουθούν (Εικόνες 4.3, 4.4, 4.5).

Στην Εικόνα 4.2 διακρίνονται πληροφορίες γενικού τύπου FTHG : goal της γηπεδούχας ομάδας, FTAG : goal της φιλενούμενης ομάδας , FTR : ποια ομάδα κέρδισε τον αγώνα.

Εικόνα 4.2 Αποτελέσματα Αγώνων & Εύστοχα Γκολ

Index	Home Team	Away Team	FTHG	FTAG	FTR
0	Liverpool	Norwich	4	1	H
1	West Ham	Man City	0	5	A
2	Bournemouth	Sheffield United	1	1	D
3	Burnley	Southampton	3	0	H
4	Crystal Palace	Everton	0	0	D
5	Watford	Brighton	0	3	A
6	Tottenham	Aston Villa	3	1	H
7	Leicester	Wolves	0	0	D
8	Newcastle	Arsenal	0	1	A
9	Man United	Chelsea	4	0	H
10	Arsenal	Burnley	2	1	H
11	Aston Villa	Bournemouth	1	2	A
12	Brighton	West Ham	1	1	D
13	Everton	Watford	1	0	H
14	Norwich	Newcastle	3	1	H
15	Southampton	Liverpool	1	2	A
16	Man City	Tottenham	2	2	D
17	Sheffield United	Crystal Palace	1	0	H
18	Chelsea	Leicester	1	1	D
19	Wolves	Man United	1	1	D
20	Aston Villa	Everton	2	0	H
21	Norwich	Chelsea	2	3	A
22	Brighton	Southampton	0	2	A
23	Man United	Crystal Palace	1	2	A
24	Sheffield United	Leicester	1	2	A

Show 25 per page 1 2 10 30 31

Στην Εικόνα 4.3 παρουσιάζονται εξειδικευμένες πληροφορίες, όπως το HC, που αναφέρεται στο συνολικό αριθμό των κόρνερ για την ομάδα που αγωνίζεται ως γηπεδούχος, και το AC, που αντιστοιχεί στον αντίστοιχο αριθμό για την ομάδα που αγωνίζεται ως φιλοξενούμενη. Επιπλέον, το HF αναφέρεται στο συνολικό αριθμό των φάουλ που πραγματοποίησε η γηπεδούχος ομάδα, ενώ το AF αφορά στον αντίστοιχο αριθμό για την ομάδα που αγωνίζεται ως φιλοξενούμενη.

Εικόνα 4.3 Σύνολο Κόρνερ & Φάουλ ανά Ομάδα

Index	Home Team	Away Team	HC	AC	HF	AF
0	Liverpool	Norwich	11	2	9	9
1	West Ham	Man City	1	1	6	13
2	Bournemouth	Sheffield United	3	4	10	19
3	Burnley	Southampton	2	7	6	12
4	Crystal Palace	Everton	6	2	16	14
5	Watford	Brighton	5	2	15	11
6	Tottenham	Aston Villa	14	0	13	9
7	Leicester	Wolves	12	3	3	13
8	Newcastle	Arsenal	5	3	12	7
9	Man United	Chelsea	3	5	15	13
10	Arsenal	Burnley	10	7	13	11
11	Aston Villa	Bournemouth	10	5	10	13
12	Brighton	West Ham	8	6	11	10
13	Everton	Watford	4	7	11	11
14	Norwich	Newcastle	7	5	9	11
15	Southampton	Liverpool	5	9	10	6
16	Man City	Tottenham	13	2	14	4
17	Sheffield United	Crystal Palace	8	4	16	11
18	Chelsea	Leicester	4	5	9	14
19	Wolves	Man United	4	6	17	8
20	Aston Villa	Everton	0	6	10	18
21	Norwich	Chelsea	1	8	9	9
22	Brighton	Southampton	8	5	9	10
23	Man United	Crystal Palace	8	1	8	18
24	Sheffield United	Leicester	7	4	11	6

Show 25 per page 1 2 10 30 31

Στην Εικόνα 4.4 παρουσιάζονται εξειδικευμένες πληροφορίες για τον αγώνα ειδικότερα. Συγκεκριμένα, η στήλη B365>2.5 δηλώνει την απόδοση μια συγκεκριμένης στοιχηματικής εταιρείας για τον αγώνα αν θα σημειωθούν περισσότερα από 2,5 γκολ στο παιχνίδι, ενώ η

στήλη B365<2.5 την απόδοση μια συγκεκριμένης στοιχηματικής εταιρείας για τον αγώνα αν θα σημειωθούν λιγότερα από 2,5 γκολ.

Εικόνα 4.4 Απόδοση Στοιχηματικής Εταιρείας για Κάποιο Γεγονός

Index	HomeTeam	AwayTeam	B365>2.5	B365<2.5
0	Liverpool	Norwich	1.4	3.0
1	West Ham	Man City	1.44	2.75
2	Bournemouth	Sheffield United	1.9	1.9
3	Burnley	Southampton	2.1	1.72
4	Crystal Palace	Everton	2.2	1.66
5	Watford	Brighton	2.1	1.72
6	Tottenham	Aston Villa	1.66	2.2
7	Leicester	Wolves	2.2	1.66
8	Newcastle	Arsenal	1.8	2.0
9	Man United	Chelsea	2.0	1.8
10	Arsenal	Burnley	1.57	2.37
11	Aston Villa	Bournemouth	1.8	2.0
12	Brighton	West Ham	1.9	1.9
13	Everton	Watford	1.72	2.1
14	Norwich	Newcastle	1.9	1.9
15	Southampton	Liverpool	1.66	2.2
16	Man City	Tottenham	1.53	2.5
17	Sheffield United	Crystal Palace	2.3	1.61
18	Chelsea	Leicester	1.9	1.9
19	Wolves	Man United	2.1	1.72
20	Aston Villa	Everton	1.72	2.1
21	Norwich	Chelsea	1.53	2.5
22	Brighton	Southampton	2.1	1.72
23	Man United	Crystal Palace	1.72	2.1
24	Sheffield United	Leicester	2.3	1.61

Show 25 per page 1 2 10 30 31

Στην Εικόνα 4.5 παρουσιάζονται οι αποδόσεις από ένα σύνολο στοιχηματικών εταιρειών στην αγορά: MaxH, η μέγιστη απόδοση αγοράς για νίκη γηπεδούχου, MaxD, η μέγιστη απόδοση αγοράς για ισοπαλία, MaxA, η μέγιστη απόδοση αγοράς για νίκη φιλοξενούμενου, AvgH, η μέση απόδοση αγοράς για νίκη γηπεδούχου, AvgD, η μέση απόδοση αγοράς για ισοπαλία, και AvgA, η μέση απόδοση αγοράς για νίκη φιλοξενούμενου.

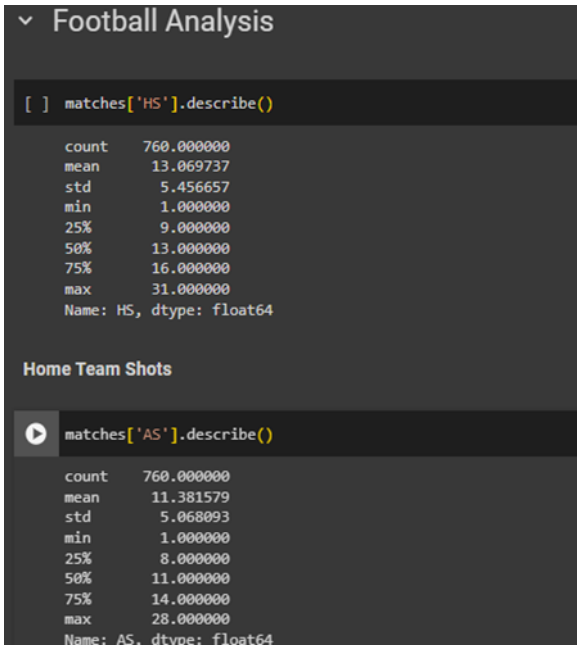
Εικόνα 4.5 Απόδοση Στοιχηματικής Εταιρείας για Κάποιο Γεγονός

Index	HomeTeam	AwayTeam	MaxH	MaxD	MaxA	AvgH	AvgD	AvgA
0	Liverpool	Norwich	1.16	10.0	23.0	1.14	8.75	19.83
1	West Ham	Man City	13.0	6.75	1.29	11.84	6.29	1.25
2	Bournemouth	Sheffield United	2.06	3.65	4.0	2.01	3.53	3.83
3	Burnley	Southampton	2.8	3.33	2.88	2.68	3.22	2.76
4	Crystal Palace	Everton	3.21	3.4	2.52	3.13	3.27	2.4
5	Watford	Brighton	2.0	3.5	4.0	1.94	3.41	4.26
6	Tottenham	Aston Villa	1.33	5.95	12.0	1.3	5.53	10.51
7	Leicester	Wolves	2.29	3.38	3.09	2.23	3.28	3.48
8	Newcastle	Arsenal	4.7	4.0	1.83	4.48	3.82	1.79
9	Man United	Chelsea	2.38	3.43	3.63	2.19	3.32	3.49
10	Arsenal	Burnley	1.36	5.75	11.0	1.33	5.49	9.49
11	Aston Villa	Bournemouth	2.42	3.64	3.2	2.33	3.49	3.08
12	Brighton	West Ham	2.65	3.42	3.0	2.52	3.33	2.9
13	Everton	Watford	1.8	4.05	5.2	1.73	3.9	4.86
14	Norwich	Newcastle	2.3	3.5	3.5	2.23	3.38	3.38
15	Southampton	Liverpool	7.5	4.8	1.52	6.77	4.67	1.47
16	Man City	Tottenham	1.4	5.5	9.0	1.36	5.22	8.46
17	Sheffield United	Crystal Palace	2.65	3.35	3.1	2.57	3.19	2.95
18	Chelsea	Leicester	1.77	4.0	5.25	1.73	3.77	5.07
19	Wolves	Man United	3.41	3.4	2.35	3.32	3.28	2.3
20	Aston Villa	Everton	3.26	3.72	2.35	3.16	3.5	2.27
21	Norwich	Chelsea	4.4	4.07	1.87	4.18	3.92	1.83
22	Brighton	Southampton	2.55	3.4	3.2	2.43	3.25	3.08
23	Man United	Crystal Palace	1.39	5.3	10.75	1.36	5.08	9.16
24	Sheffield United	Leicester	3.6	3.34	2.32	3.45	3.23	2.26

Show 25 per page 1 2 10 30 31

Στις Εικόνες 4.6 και 4.7 απεικονίζονται μια ανάλυση των σουτ από την γηπεδούχο και τη φιλοξενούμενη ομάδα αντίστοιχα, καθώς και μια ανάλυση των φάουλ που έχουν σφυριχτεί.

Εικόνα 4.6 Σουτ ανά Γηπεδούχα Ομάδα



Εικόνα 4.7 Σουτ ανά Φιλοξενούμενη Ομάδα

Home Team Fouls Committed

```
[ ] matches['HF'].describe()

count    760.000000
mean      10.853947
std        3.528314
min        0.000000
25%        9.000000
50%       11.000000
75%       13.000000
max       23.000000
Name: HF, dtype: float64
```

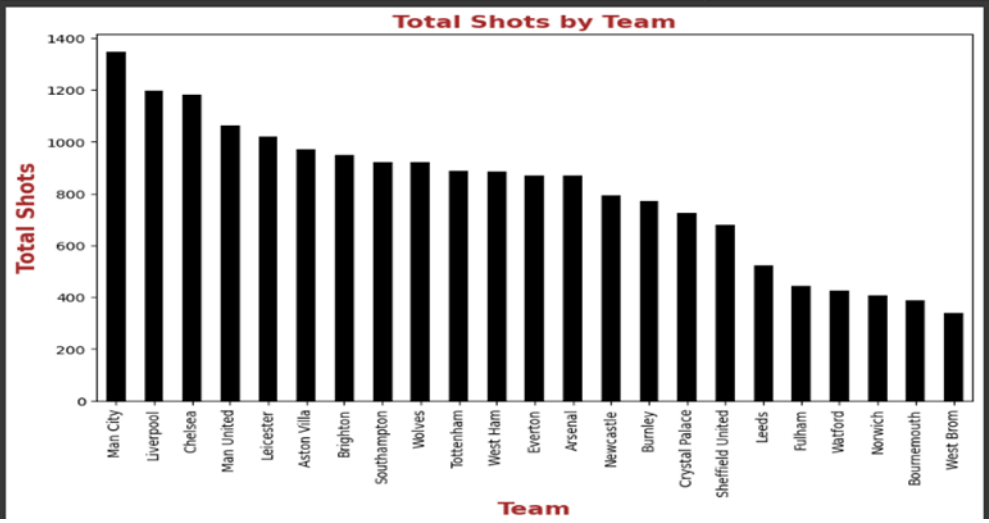
Away Team Fouls Committed

```
[ ] matches['AF'].describe()

count    760.000000
mean      10.765789
std        3.600306
min        1.000000
25%        8.000000
50%       11.000000
75%       13.000000
max       24.000000
Name: AF, dtype: float64
```

Στην Εικόνα 4.8 περιγράφονται το σύνολο των σουτ ανά ομάδα σε φθίνουσα σειρά.

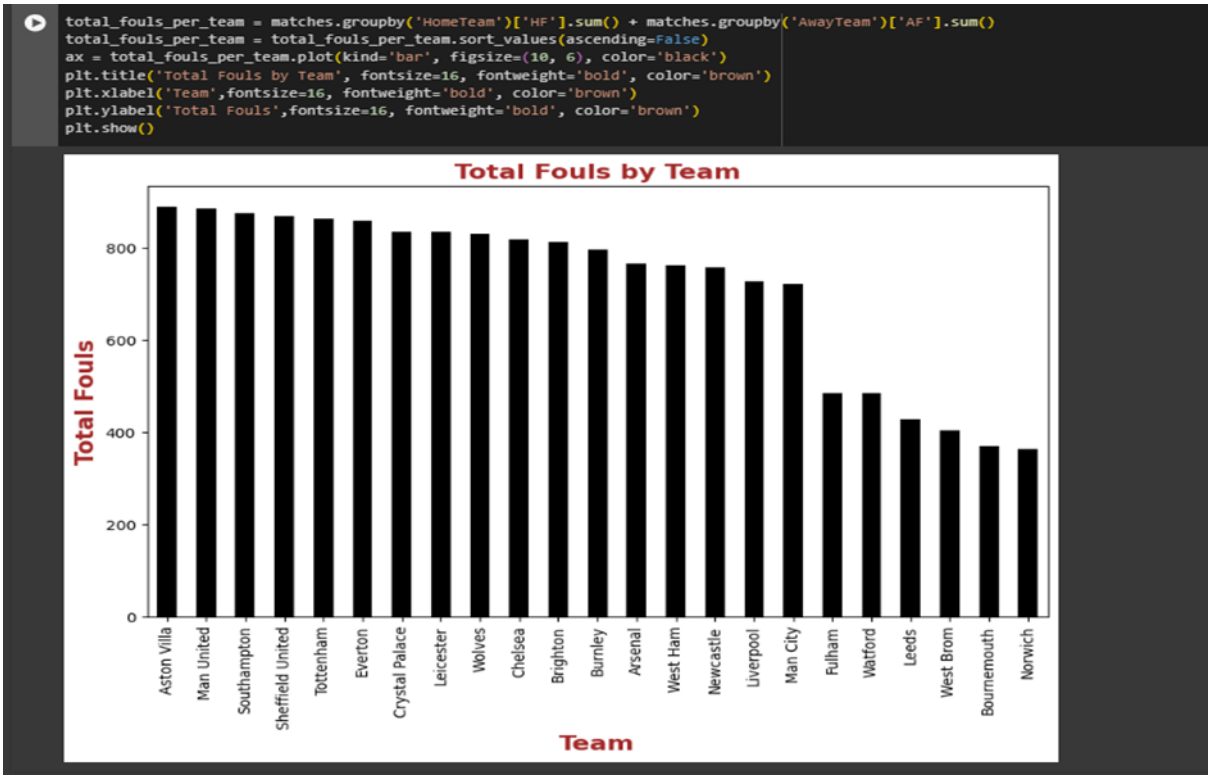
Εικόνα 4.8 Σουτ ανά Ομάδα

```
total_shots_per_team = matches.groupby('HomeTeam')['HS'].sum() + matches.groupby('AwayTeam')['AS'].sum()
total_shots_per_team = total_shots_per_team.sort_values(ascending=False)
ax = total_shots_per_team.plot(kind='bar', figsize=(10, 6), color='black')
plt.title('Total Shots by Team', fontsize=16, fontweight='bold', color='brown')
plt.xlabel('Team',fontsize=16, fontweight='bold', color='brown')
plt.ylabel('Total Shots',fontsize=16, fontweight='bold', color='brown')
plt.show()
```



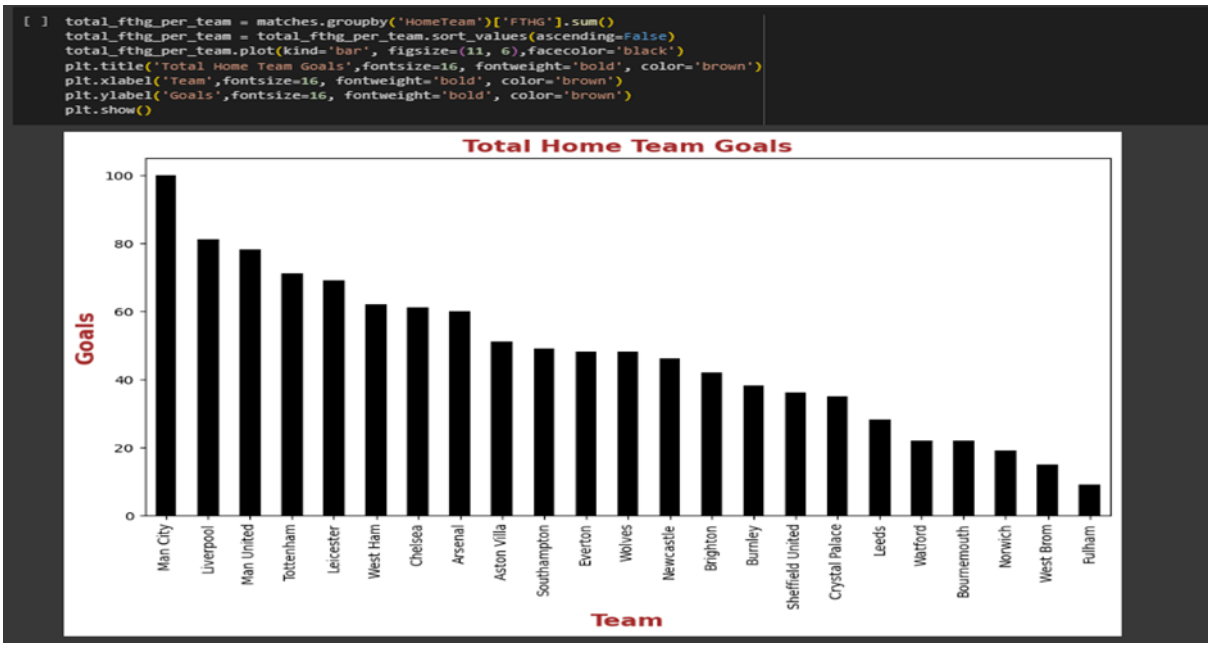
Στην Εικόνα 4.9 περιγράφονται το σύνολο των φάουλ ανά ομάδα.

Εικόνα 4.9 Σύνολο Φάουλ ανά Ομάδα



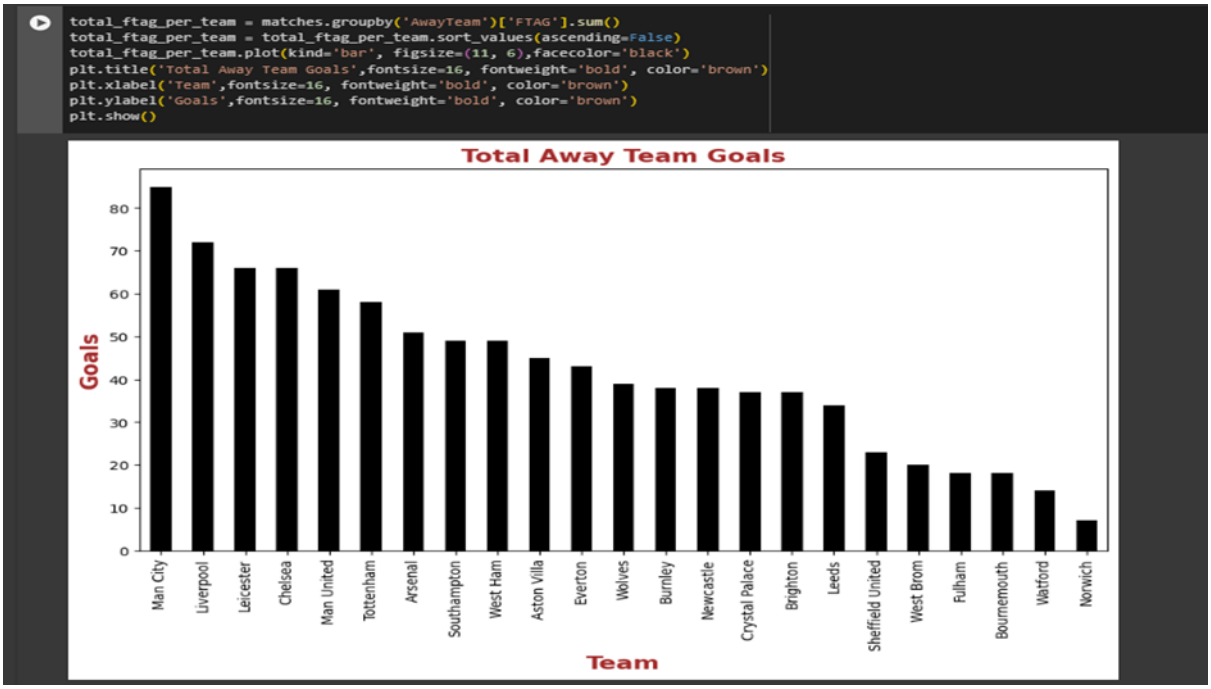
Επιπλέον, εμφανίζονται πληροφορίες για το σύνολο των γκολ ανά γηπεδούχα ομάδα (Εικόνα 4.10).

Εικόνα 4.10 Σύνολο Γκολ ανά Γηπεδούχα Ομάδα



Στην Εικόνα 4.11 περιγράφονται τα συνολικά γκολ που σημειώθηκαν από τις ομάδες ως φιλοξενούμενες.

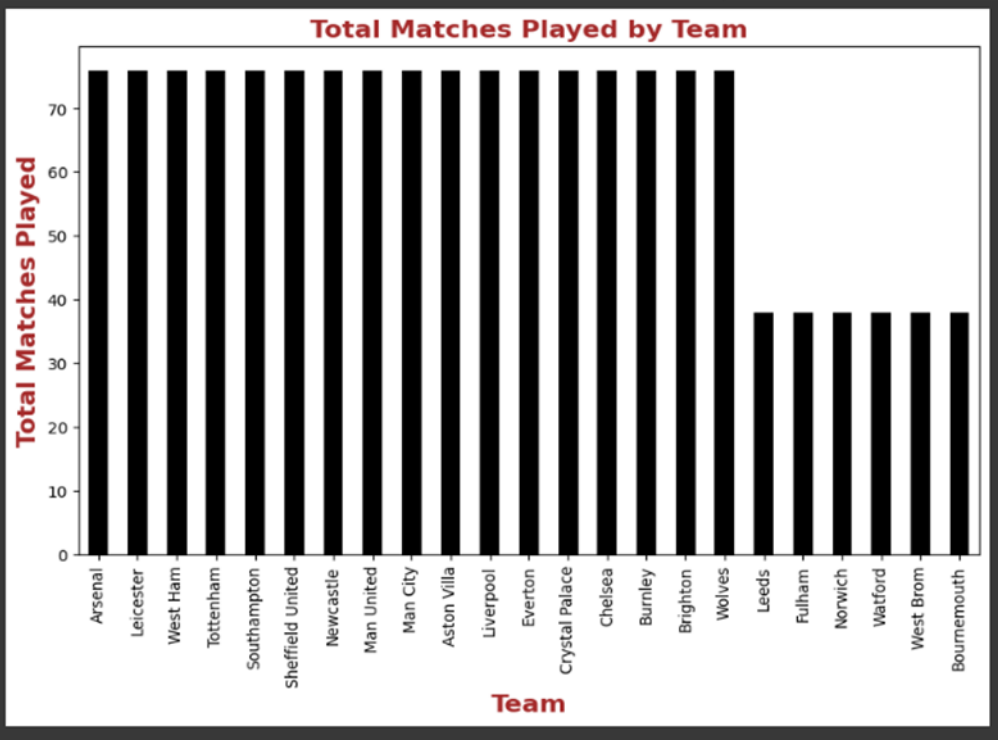
Εικόνα 4.11 Σύνολο Γκολ ανά Φιλοξενούμενη Ομάδα



Τέλος, εμφανίζονται πληροφορίες για το σύνολο των αγώνων που έχουν πραγματοποιηθεί ανά ομάδα. (Εικόνα 4.12)

Εικόνα 4.12 Σύνολο Αγώνων ανά Ομάδα

```
[ ] home_team_counts = matches["HomeTeam"].value_counts()
    away_team_counts = matches["AwayTeam"].value_counts()
    total_matches_counts_per_team = home_team_counts.add(away_team_counts, fill_value=0)
    total_matches_counts_per_team = total_matches_counts_per_team.sort_values(ascending=False)
    ax = total_matches_counts_per_team.plot(kind='bar', figsize=(10, 6),color='black')
    plt.title('Total Matches Played by Team',fontsize=16, fontweight='bold', color='brown')
    plt.xlabel('Team',fontsize=16, fontweight='bold', color='brown')
    plt.ylabel('Total Matches Played',fontsize=16, fontweight='bold', color='brown')
    plt.show()
```



4.2 Χρήση και Ενεργοποίηση Μοντέλων Μηχανικής Μάθησης

4.2.1 Εισαγωγή

Στους στόχους της παρούσας προσέγγισης περιλαμβάνεται, μεταξύ άλλων, να ερευνήσουμε εάν μπορούν τα μοντέλα μηχανικής μάθησης που θα δημιουργήσουμε να μπορούν να

προβλέψουν το τελικό αποτέλεσμα (με υψηλή ακρίβεια) χρησιμοποιώντας ιστορικά δεδομένα για τη σεζόν 2019-2021. Με άλλα λόγια, στόχο αποτελεί ο βαθμός της δυνατότητας πρόβλεψης των τελικών αποτελεσμάτων ποδοσφαιρικών αγώνων με μοντέλα μηχανικής μάθησης σε ανάλυση συναισθήματος – γνώμης από τα μέσα κοινωνικής δικτύωσης (social media).

Η χρήση και ενεργοποίηση μοντέλων μηχανικής μάθησης αφορά τον ορισμό πλαισίου αξιολόγησης των αποτελεσμάτων των αγώνων, την προετοιμασία των δεδομένων που θα επεξεργαστούν από τις διαδικασίες μηχανικής μάθησης, την προετοιμασία συνόλων δεδομένων για εκπαίδευση και για αξιολόγηση, τον ορισμό δεικτών (το input που θα βάλουμε στο μοντέλα) πρόβλεψης και, τέλος, την δημιουργία των μοντέλων μηχανικής μάθησης (Baboota & Kaur, 2019).

4.2.2 Πλαίσιο Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων των αγώνων θα δημιουργήσουμε ένα πλαίσιο με δύο διακριτές τιμές για το αποτέλεσμα κάθε αγώνα, το οποίο μπορούμε να εκφράσουμε ως δυαδικό σύστημα. Οι τιμές που μπορεί να πάρει κάθε αγώνας ως τελικό αποτέλεσμα είναι «Νίκη» (αν κερδίσει η γηπεδούχος ομάδα) ή «Ήττα» (αν χάσει ή φέρει ισοπαλία η γηπεδούχος ομάδα).

Βάσει του προαναφέρομενου πλαισίου αξιολόγησης, θέλουμε να δημιουργήσουμε προγνωστικά και δείκτες που θα μπορούμε να εισάγουμε σε κάθε αλγόριθμο μηχανικής μάθησης, ώστε να είναι δυνατή η εκτίμηση του τρόπου με τον οποίον ανταποκρίνονται οι αλγόριθμοι και του είδους των αποτελεσμάτων που παρουσιάζουν. Εδώ αξίζει να σημειωθεί ότι κάθε αλγόριθμος μηχανικής μάθησης χρειάζεται δεδομένα σε συγκεκριμένη μορφή για να μπορεί να εκπαιδευτεί, και αυτή η μορφή είναι αριθμητική (Baboota & Kaur, 2019).

4.2.3 Προετοιμασία Δεδομένων

Κατά συνέπεια, για την προετοιμασία των δεδομένων που θα επεξεργαστούν από τις διαδικασίες μηχανικής μάθησης, προβαίνουμε στις απαραίτητες ενέργειες για να μετατρέψουμε τα δεδομένα μας στην επιθυμητή μορφή, χωρίς να αλλοιώσουμε τις αρχικές πληροφορίες. Η προετοιμασία των δεδομένων μπορεί να διαχωριστεί σε τρείς επί μέρους

διαδικασίες. Συγκεκριμένα, αρχικά ομαδοποιούμε τις τιμές που μπορεί να πάρει ένα αποτέλεσμα αγώνα, όπως αναφέραμε προηγουμένως, και μετατρέπουμε το όνομα κάθε ομάδας στο dataset μας σε έναν αριθμό (Προετοιμασία Δεδομένων-1: Αρχική ομαδοποίηση τιμών αποτελεσμάτων αγώνα). Ακολουθούμε την ίδια διαδικασία για τις ομάδες που παίζουν εντός έδρας και για τις ομάδες που παίζουν εκτός έδρας (Προετοιμασία Δεδομένων-2: Ομαδοποίηση τιμών αποτελεσμάτων για ομάδες εντός / εκτός έδρας). Στη συνέχεια, πραγματοποιούμε διασταυρώσεις για να είμαστε σίγουροι ότι δεν έχουμε χάσει ή αλλοιώσει κάποια πληροφορία (Προετοιμασία Δεδομένων-3: Έλεγχος – διασταύρωση πληροφοριών).

Για την περιγραφή της προετοιμασίας των δεδομένων σε συνάρτηση με το πλαίσιο αξιολόγησης, παρουσιάζεται χαρακτηριστικό παράδειγμα της Αγγλικής ομάδας Liverpool.

Ειδικότερα, στα δεδομένα της ομάδας Liverpool (Εικόνα 4.13, 4.14) παρατηρούμε τον δείκτη αποτελέσματος στη στήλη "FTR_indicator", όπου το νούμερο αντιπροσωπεύει την ομάδα Liverpool ως "HomeTeam_indicator" και την αντίπαλη ομάδα ως "AwayTeam_indicator". Η τελευταία στήλη, "Code_Match", δείχνει αν η μετάβαση από κείμενο σε αριθμητική μορφή ήταν επιτυχής. Η ένδειξη "false" σημαίνει ότι όλες οι μετατροπές έγιναν σωστά, χωρίς ανεπιθύμητα αποτελέσματα.

Εικόνα 4.13 Προεπισκόπηση Δεδομένων

```
value_to_category = {'H': 1, 'A': 0, 'D': 0} #We will make a new column that will has this format
matches["FTR_indicator"] = matches["FTR"].map(value_to_category)
matches["HomeTeam_indicator"] = matches["HomeTeam"].astype("category").cat.codes

#Assuming you already have the 'HomeTeam' column mapped to codes
matches["HomeTeam_indicator"] = matches["HomeTeam"].astype("category").cat.codes

#Create a mapping dictionary based on the codes in the 'HomeTeam' column
team_mapping = dict(zip(matches["HomeTeam"], matches["HomeTeam_indicator"]))

#Apply the mapping to the second column ('AwayTeam' in this example)
matches["AwayTeam_indicator"] = matches["AwayTeam"].map(team_mapping)

#Let's make sure tha the mapping was performed correctly

matches["Code_Match"] = matches["HomeTeam_indicator"] == matches["AwayTeam_indicator"]
#Display rows where the codes do not match
print(matches[matches["Code_Match"] == False])
mismatch_count = matches["Code_Match"].sum()
#Display the count of mismatches
print("Number of mismatches:", mismatch_count)
```

```
    Div        Date   Time          HomeTeam          AwayTeam  FTHG  FTAG  \
0    E0  09/08/2019  20:00         Liverpool           Norwich     4     1
1    E0  10/08/2019  12:30          West Ham          Man City     0     5
2    E0  10/08/2019  15:00       Bournemouth  Sheffield United     1     1
3    E0  10/08/2019  15:00           Burnley       Southampton     3     0
4    E0  10/08/2019  15:00    Crystal Palace           Everton     0     0
..   ..         ...    ...               ...               ...   ...   ...
755  E0  23/05/2021  16:00         Liverpool    Crystal Palace     2     0
756  E0  23/05/2021  16:00          Man City           Everton     5     0
757  E0  23/05/2021  16:00  Sheffield United           Burnley     1     0
758  E0  23/05/2021  16:00          West Ham       Southampton     3     0
759  E0  23/05/2021  16:00            Wolves         Man United     1     2

    FTR  HTHG  HTAG  ... PCAHH PCAHA  MaxCAHH  MaxCAHA  AvgCAHH  AvgCAHA  \
0     H     4     0  ...  1.94  1.98     1.99     2.07     1.90     1.99
1     A     0     1  ...  1.96  1.97     2.07     1.98     1.97     1.92
2     D     0     0  ...  1.98  1.95     2.00     1.96     1.96     1.92
3     H     0     0  ...  1.89  2.03     1.90     2.07     1.86     2.02
4     D     0     0  ...  1.97  1.96     2.03     2.08     1.96     1.93
..   ..   ...   ...  ...   ...   ...      ...      ...      ...      ...
755   H     1     0  ...  1.88  2.03     1.98     2.14     1.88     2.00
756   H     2     0  ...  1.99  1.89     2.20     2.00     2.03     1.85
757   H     1     0  ...  2.05  1.86     2.17     1.90     2.03     1.84
758   H     2     0  ...  2.02  1.91     2.06     2.01     1.99     1.89
759   A     1     2  ...  2.10  1.84     2.10     1.94     2.00     1.88

     FTR_indicator  HomeTeam_indicator  AwayTeam_indicator  Code_Match
0                1                  11                  15       False
1                0                  21                  12       False
2                0                   2                  16       False
3                1                   4                  17       False
4                0                   6                   7       False
..             ...                 ...                 ...         ...
755              1                  11                   6       False
756              1                  12                   7       False
757              1                  16                   4       False
758              1                  21                  17       False
759              0                  22                  13       False

[760 rows x 110 columns]
Number of mismatches: 0
```

Εικόνα 4.14 Παράδειγμα Ομάδας

```
matches[matches["HomeTeam"] == 'Liverpool']
```

	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	...	PCAHH	PCAHA	MaxCAHH	MaxCAHA	AvgCAHH	AvgCAHA	FTR_indicator	HomeTeam_indicator	AwayTeam_indicator	Code_Match
0	E0	09/08/2019	20:00	Liverpool	Norwich	4	1	H	4	0	...	1.94	1.98	1.99	2.07	1.90	1.99	1	11	15	False
26	E0	24/08/2019	17:30	Liverpool	Arsenal	3	1	H	1	0	...	1.97	1.97	2.08	2.04	1.96	1.93	1	11	0	False
40	E0	14/09/2019	12:30	Liverpool	Newcastle	3	1	H	2	1	...	1.92	2.02	1.92	2.17	1.82	2.07	1	11	14	False
72	E0	05/10/2019	15:00	Liverpool	Leicester	2	1	H	1	0	...	2.03	1.89	2.05	1.95	1.98	1.90	1	11	10	False
98	E0	27/10/2019	16:30	Liverpool	Tottenham	2	1	H	0	1	...	2.04	1.89	2.06	1.92	2.01	1.88	1	11	18	False
119	E0	10/11/2019	16:30	Liverpool	Man City	3	1	H	2	0	...	2.01	1.93	2.04	2.08	1.98	1.90	1	11	12	False
133	E0	30/11/2019	15:00	Liverpool	Brighton	2	1	H	2	0	...	2.09	1.83	2.13	1.96	2.05	1.83	1	11	3	False
147	E0	04/12/2019	20:15	Liverpool	Everton	5	2	H	4	2	...	1.84	2.10	1.85	2.27	1.80	2.10	1	11	7	False
160	E0	14/12/2019	12:30	Liverpool	Watford	2	0	H	1	0	...	2.15	1.80	2.18	1.89	2.10	1.79	1	11	19	False
197	E0	29/12/2019	16:30	Liverpool	Wolves	1	0	H	1	0	...	1.88	2.07	1.88	2.14	1.83	2.05	1	11	22	False
208	E0	02/01/2020	20:00	Liverpool	Sheffield United	2	0	H	1	0	...	1.98	1.95	2.01	2.05	1.90	1.97	1	11	16	False
228	E0	19/01/2020	16:30	Liverpool	Man United	2	0	H	1	0	...	2.09	1.84	2.13	1.94	2.04	1.84	1	11	13	False
243	E0	01/02/2020	15:00	Liverpool	Southampton	4	0	H	0	0	...	2.03	1.89	2.05	1.99	1.96	1.91	1	11	17	False
269	E0	24/02/2020	20:00	Liverpool	West Ham	3	2	H	1	1	...	2.06	1.86	2.07	1.98	2.02	1.85	1	11	21	False
278	E0	07/03/2020	12:30	Liverpool	Bournemouth	2	1	H	2	1	...	2.13	1.79	2.15	1.93	2.08	1.80	1	11	2	False
306	E0	24/06/2020	20:15	Liverpool	Crystal Palace	4	0	H	2	0	...	1.95	1.97	1.96	2.17	1.88	2.00	1	11	8	False
327	E0	05/07/2020	16:30	Liverpool	Aston Villa	2	0	H	0	0	...	1.98	1.94	2.20	1.95	1.99	1.89	1	11	1	False
342	E0	11/07/2020	15:00	Liverpool	Burnley	1	1	D	1	0	...	1.92	2.00	1.96	2.08	1.89	1.98	0	11	4	False
369	E0	22/07/2020	20:15	Liverpool	Chelsea	5	3	H	3	1	...	1.93	2.01	1.95	2.04	1.91	1.96	1	11	5	False
382	E0	12/09/2020	17:30	Liverpool	Leeds	4	3	H	3	2	...	1.85	2.08	1.90	2.16	1.84	2.04	1	11	9	False
407	E0	28/09/2020	20:00	Liverpool	Arsenal	3	1	H	2	1	...	2.06	1.87	2.13	1.89	2.05	1.84	1	11	0	False
432	E0	24/10/2020	20:00	Liverpool	Sheffield United	2	1	H	1	1	...	2.00	1.91	2.02	2.03	1.96	1.92	1	11	16	False
441	E0	31/10/2020	17:30	Liverpool	West Ham	2	1	H	1	1	...	1.93	2.00	1.94	2.04	1.90	1.99	1	11	21	False
465	E0	22/11/2020	19:15	Liverpool	Leicester	3	0	H	2	0	...	2.05	1.88	2.08	1.93	2.01	1.88	1	11	10	False
485	E0	06/12/2020	19:15	Liverpool	Wolves	4	0	H	1	0	...	2.00	1.92	2.02	2.03	1.96	1.90	1	11	22	False
503	E0	16/12/2020	20:00	Liverpool	Tottenham	2	1	H	1	1	...	2.05	1.88	2.09	1.88	2.05	1.83	1	11	18	False
525	E0	27/12/2020	16:30	Liverpool	West Brom	1	1	D	1	0	...	2.03	1.88	2.07	1.95	1.99	1.87	0	11	20	False
556	E0	17/01/2021	16:30	Liverpool	Man United	0	0	D	0	0	...	1.94	1.98	2.01	1.98	1.94	1.93	0	11	13	False
563	E0	21/01/2021	20:00	Liverpool	Burnley	0	1	A	0	0	...	1.86	2.07	1.86	2.27	1.78	2.13	0	11	4	False
593	E0	03/02/2021	20:15	Liverpool	Brighton	0	1	A	0	0	...	1.87	2.06	1.90	2.20	1.82	2.06	0	11	3	False
602	E0	07/02/2021	16:30	Liverpool	Man City	1	4	A	0	0	...	1.93	2.00	2.02	2.03	1.92	1.97	0	11	12	False

Example of how many matches has Liverpool for the given dataset

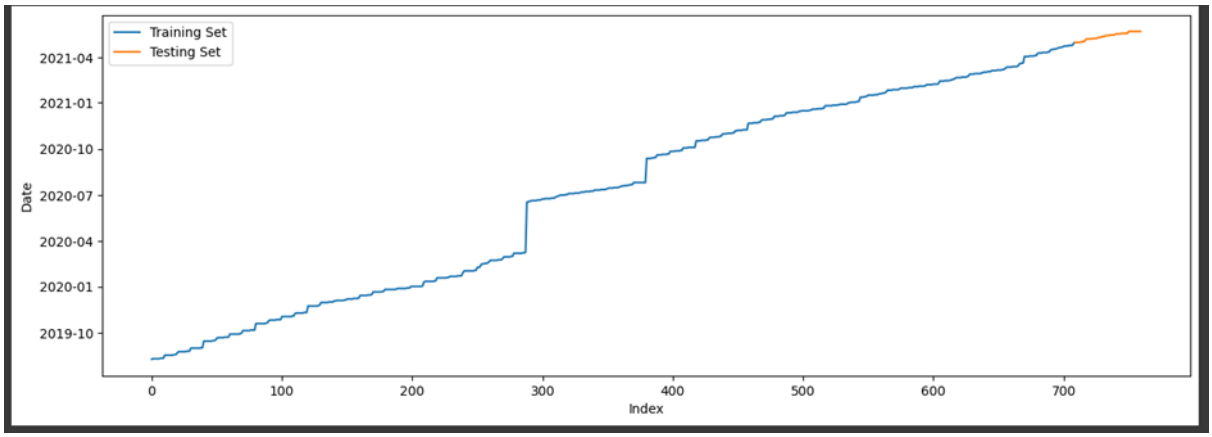
4.2.4 Προετοιμασία Συνόλων Δεδομένων για Εκπαίδευση και για Αξιολόγηση

Μετά την ολοκλήρωση της προετοιμασίας των δεδομένων για την Μηχανική Μάθηση στην συνέχεια, ακολουθεί το στάδιο, στο οποίο προετοιμάζουμε το σύνολο δεδομένων - dataset για να το χωρίσουμε σε δύο μέρη: το εκπαιδευτικό σύνολο και το σύνολο αξιολόγησης (Εικόνα 4.15α, 4.15β). Συγκεκριμένα, το εκπαιδευτικό σετ θα περιλαμβάνει τους αγώνες που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου μας, ενώ το σετ αξιολόγησης θα αποτελείται από δεδομένα που θα χρησιμοποιηθούν για την αξιολόγηση της απόδοσης του αλγορίθμου που φτιάξαμε (Mahesh, 2020).

Εικόνα 4.15 Διαίρεση Συνόλου μεταξύ Εκπαίδευσης & Επαυλήθευσης

```
Prepearing the dataset

[ ]  proper_date = '2021-05-01'
     training_set = matches[matches["Date"] < proper_date] #SPLIT THE SET
     testing_set = matches[matches["Date"] >= proper_date]
```

Εικόνα 4.15β Διαίρεση Συνόλου ως προς το Σύνολο των Αγώνων



Το σύνολο των δεδομένων, δηλαδή οι διαφορετικοί αγώνες που θα περιλαμβάνονται στο εκπαιδευτικό σύνολο, ανέρχεται σε 709. Οι υπόλοιποι 51 αγώνες θα αποτελέσουν το σύνολο αξιολόγησης, το οποίο θα χρησιμοποιηθεί για να εκτιμήσουμε την απόδοση των μοντέλων μας.

4.2.5 Ορισμός Δεικτών (εκπαιδευτικά δεδομένα εκίνησης) Πρόβλεψης

Για την αξιολόγηση των αποτελεσμάτων απαραίτητη προϋπόθεση είναι ο ορισμός δεικτών πρόβλεψης- predictors δηλαδή το input που θα χρησιμοποιήσουμε για να feedαρουμε τα μοντέλα μας ώστε να μπορούν να εξάγουν αποτελέσματα. Η λίστα των predictors (δεικτών πρόβλεψης) αποτελεί τα δεδομένα που το μοντέλο χρησιμοποιεί ως δεδομένα εισόδου - input για να προβλέψει το αποτέλεσμα ενός μελλοντικού αγώνα Mahesh (2020).

Οι δείκτες πρόβλεψης-predictors που ορίστηκαν για την αξιολόγηση του συνόλου δεδομένων από το Αγγλικό Πρωτάθλημα (English Premier League) είναι οι εξής:

- home_team_indicator : ένας ακέραιος αριθμός που δηλώνει την ομάδα που παίζει ως γηπεδούχος.
- away_team_indicator : ένας ακέραιος αριθμός που δηλώνει την ομάδα που παίζει ως φιλοξενούμενη.
- day_indicator : ένας ακέραιος αριθμός που δηλώνει την ημέρα που διεξήχθη ο αγώνας.

Με βάση την προαναφερόμενη λίστα των predictors, κάθε μοντέλο που θα πειραματιστούμε παρακάτω θα μας δώσει ένα accuracy.

Το ποσοστό-δείκτης ακρίβειας (accuracy) αποτελεί έναν κοινό μετρικό δείκτη για την αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης. Αυτός ο δείκτης υπολογίζει το ποσοστό των προβλέψεων του μοντέλου που είναι σωστές σε σχέση με τις πραγματικές τιμές των δεδομένων ελέγχου (Juba and Le, 2019).

Συγκεκριμένα, η ακρίβεια ορίζεται ως ο λόγος του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων. Για παράδειγμα, εάν ένα μοντέλο καταφέρνει να κάνει σωστές προβλέψεις για 80 από τα 100 δεδομένα ελέγχου, τότε η ακρίβειά του θα είναι 80%.

Ο δείκτης ακρίβειας είναι χρήσιμος για να αξιολογήσουμε πόσο καλά λειτουργεί ένα μοντέλο σε σχέση με τον στόχο του προβλήματος. Ωστόσο, πρέπει να λαμβάνουμε υπόψη το πιθανό αποτέλεσμα του τυχαίου για ορισμένα προβλήματα, καθώς η ακρίβεια μπορεί να παραπλανηθεί αν το μοντέλο δεν είναι καλά ισορροπημένο για όλες τις κλάσεις του προβλήματος (Juba and Le, 2019) .

Συνολικά το accuracy υποδεικνύει πόσο καλά μπορεί το μοντέλο να προβλέψει το σωστό αποτέλεσμα ενός αγώνα, δηλαδή την νίκη («Νίκη») ή την ήττα («Ήττα») της γηπεδούχου ομάδας.

4.3. Δημιουργία Μοντέλων

4.3.1 Εισαγωγή

Όπως προαναφέρθηκε, για την δημιουργία μοντέλων μηχανικής μάθησης για την επεξεργασία των δεδομένων του Πρωταθλήματος της Engish Premier League εξετάζονται οι αλγόριθμοι Random Forest, MLPClassifier και XGBoost.

4.3.2 Αλγόριθμος Random Forest

Ο αλγόριθμος random forest είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης που ανήκει στην κατηγορία των μεθόδων συνόλου. Η βασική ιδέα του Random Forest είναι να δημιουργήσει ένα "δάσος" από πολλά δέντρα απόφασης, όπου κάθε δέντρο εκπαιδεύεται σε διαφορετικά τμήματα των δεδομένων και μερικές φορές σε διαφορετικά χαρακτηριστικά (Rigatti, 2017).

Χαρακτηριστικό στοιχείο του αλγορίθμου random forest είναι ότι είναι ανθεκτικός στο υπεπροσδιορισμό, θεωρείται ότι αποτελεί αποδοτική λύση για μεγάλα σύνολα δεδομένων, και αντιμετωπίζει αποτελεσματικά δεδομένα υψηλής διάστασης. Παρέχει κατάταξη σημαντικότητας χαρακτηριστικών και μπορεί να χειριστεί τόσο προβλήματα παλινδρόμησης όσο και ταξινόμησης. Επιπλέον, ανιχνεύει μη γραμμικές σχέσεις στα δεδομένα με αποτελεσματικό τρόπο (Rigatti, 2017).

Για τον αλγόριθμο random forest θα πειραματιστούμε με δύο μετρικές. Αμφότερες οι μετρικές είναι σημαντικές για την ολοκληρωμένη αξιολόγηση της απόδοσης του μοντέλου. Η ακρίβεια μας δίνει μια γενική εικόνα της συνολικής ακρίβειας των προβλέψεων, ενώ η precision μας δείχνει πόσο αξιόπιστες είναι οι θετικές προβλέψεις του μοντέλου (Juba & Le, 2019). Επιλέγοντας κατάλληλα τις μετρικές για το πρόβλημα που αντιμετωπίζουμε, μπορούμε να καταλήξουμε σε καλύτερη κατανόηση και βελτίωση της απόδοσης του αλγορίθμου Random Forest.

Οι δύο μετρικές με τις οποίες θα πειραματιστούμε για το συγκεκριμένο αλγόριθμο είναι το accuracy και το precision, οι οποίες ορίζονται και περιγράφονται ως εξής:

Accuracy: Μετρά τη συνολική ορθότητα των προβλέψεων διαιρώντας τον αριθμό των σωστών προβλέψεων με τον συνολικό αριθμό προβλέψεων (Juba & Le, 2019).

Precision: Επικεντρώνεται στην ακρίβεια των θετικών προβλέψεων, υπολογίζοντας τα πραγματικά θετικά διαιρώντας με το άθροισμα των πραγματικών θετικών και των ψευδών θετικών. Είναι ιδιαίτερα χρήσιμη όταν η ελαχιστοποίηση των ψευδών θετικών είναι κρίσιμη (Juba & Le, 2019).

Επιπλέον, για τον αλγόριθμο random forest θα εξετάσουμε δύο (2) προσεγγίσεις.

Η πρώτη προσέγγιση (Εικόνα 4.16) θα εστιάσει στην ακρίβεια των προβλέψεων (accuracy), αναζητώντας τη βέλτιστη συνολική απόδοση του μοντέλου σε όλες τις κλάσεις. Αυτή η προσέγγιση είναι κατάλληλη για γενικές εφαρμογές όπου δεν υπάρχει έμφαση σε συγκεκριμένη κλάση.

Η δεύτερη προσέγγιση θα αποτελείται από την ανάπτυξη μιας συνάρτησης που θα αξιολογεί την απόδοση της ομάδας μέσω της φόρμας της. Θα δημιουργήσουμε μια μετρική που θα εκτιμά την τρέχουσα φόρμα της ομάδας, εξετάζοντας την τάση των αποτελεσμάτων της σε μια προηγούμενη χρονική περίοδο. Η συνάρτηση αυτή θα λαμβάνει υπόψη την ακρίβεια (accuracy) των προβλέψεων για την ομάδα, αξιολογώντας εάν η ομάδα βρίσκεται σε καλή φόρμα ή όχι. Με αυτόν τον τρόπο, θα εξετάσουμε εάν η γενική απόδοση του μοντέλου σε όλες τις κλάσεις βελτιώνεται με βάση την εκτίμηση της ομάδας σε καλή φόρμα.

1η προσέγγιση του Random Forest

Εύρεση βέλτιστης ακρίβειας με βάση την απόδοση του μοντέλου

Εικόνα 4.16 Random Forest classifier

Στη συνέχεια, για τον αλγόριθμο random forest, θα προσπαθήσουμε να βρούμε το σύνολο των παραμέτρων που μπορεί να αποδώσει την καλύτερη ακρίβεια για το μοντέλο.

Σε κάθε αλγόριθμο μηχανικής μάθησης μπορούμε να πειραματιστούμε με παραμέτρους για να πετύχουμε το καλύτερο δυνατό αποτέλεσμα. Για τον συγκεκριμένο αλγόριθμο, παρατηρούμε ότι η ακρίβεια μπορεί να αυξηθεί σημαντικά με ορισμένες συγκεκριμένες παραμέτρους.

Εικόνα 4.17 Πειραματισμός Παραμέτρων του Random Forest classifier

```
n_estimators_values = [50, 100, 150]
min_samples_split_values = [5, 10, 15]

best_accuracy = 0
best_params = {}

# Iterate through the combinations
for n_estimators in n_estimators_values:
    for min_samples_split in min_samples_split_values:
        # Create and train the model
        random_forest_1 = RandomForestClassifier(n_estimators=n_estimators, min_samples_split=min_samples_split, random_state=1)
        random_forest_1.fit(training_set[predictors], training_set["Target"])

        # Make predictions on the testing set
        predictions = random_forest_1.predict(testing_set[predictors])

        # Calculate accuracy
        accuracy = accuracy_score(testing_set["Target"], predictions)

        # Check if this model has the best accuracy
        if accuracy > best_accuracy:
            best_accuracy = accuracy
            best_params = {'n_estimators': n_estimators, 'min_samples_split': min_samples_split}

# Print the best parameters and accuracy
print(f"Best parameters: {best_params}")
print(f"Best accuracy: {best_accuracy}")
```

```
Best parameters: {'n_estimators': 50, 'min_samples_split': 15}
Best accuracy: 0.6274509803921569
```

2η προσέγγιση του Random Forest

Στη συνέχεια, θα αναπτύξουμε μια προσέγγιση που (ίσως) θα βοηθήσει τον αλγόριθμο να μάθει από προηγούμενες εμπειρίες, επιτρέποντάς μας να αναγνωρίσουμε ένα μοτίβο που μπορεί να λειτουργήσει ως εκτιμητής για την απόδοση μιας ομάδας. Θα δημιουργήσουμε μια συνάρτηση που θα επεξεργάζεται τα δεδομένα κάθε ομάδας για μια συγκεκριμένη χρονική περίοδο, καταγράφοντας την τρέχουσα απόδοσή της. Αυτή η συνάρτηση θα εξετάζει αν η ομάδα βρίσκεται σε φόρμα ή όχι και θα χρησιμοποιεί τους εκτιμητές μας για να καταγράφει πληροφορίες βάσει της φόρμας της ομάδας.

Για να το επιτύχουμε αυτό, θα συλλέξουμε δεδομένα για κάθε ομάδα και θα τα οργανώσουμε χρονολογικά, ώστε να διατηρούμε πληροφορίες για την απόδοση της ομάδας

με την πάροδο του χρόνου. Με αυτόν τον τρόπο, θα μπορούμε να αναλύουμε την τρέχουσα φόρμα της ομάδας και να κατανοούμε πώς αυτή επηρεάζει την απόδοσή της.

Οι νέες επιπλέον στήλες που θα προσθέσουμε στον αλγόριθμο για να ενσωματώσουμε την επιπλέον πληροφορία, βάσει της συνάρτησης που δημιουργήσαμε για την αξιολόγηση της απόδοσης κάθε ομάδας, είναι οι εξής:

- FTHG (Full Time Home Goals): Αριθμός γκολ που σημείωσε η γηπεδούχος ομάδα στο τέλος του αγώνα.
- FTAG (Full Time Away Goals): Αριθμός γκολ που σημείωσε η φιλοξενούμενη ομάδα στο τέλος του αγώνα.
- HS (Home Team Shots): Αριθμός συνολικών σουτ της γηπεδούχου ομάδας κατά τη διάρκεια του αγώνα.
- AS (Away Team Shots): Αριθμός συνολικών σουτ της φιλοξενούμενης ομάδας κατά τη διάρκεια του αγώνα.
- HST (Home Team Shots on Target): Αριθμός σουτ της γηπεδούχου ομάδας που ήταν εντός στόχου.
- AST (Away Team Shots on Target): Αριθμός σουτ της φιλοξενούμενης ομάδας που ήταν εντός στόχου.
- HR (Home Team Red Cards): Αριθμός κόκκινων καρτών που δόθηκαν στη γηπεδούχο ομάδα.
- AR (Away Team Red Cards): Αριθμός κόκκινων καρτών που δόθηκαν στη φιλοξενούμενη ομάδα.

Εδώ βλέπουμε ένα παράδειγμα μιας ομάδας και όλες τις καινούργιες στήλες που θα λάβουμε υπόψη ως εκτιμητές:

Εικόνα 4.18 Ενδεικτικό Παράδειγμα Ομάδας

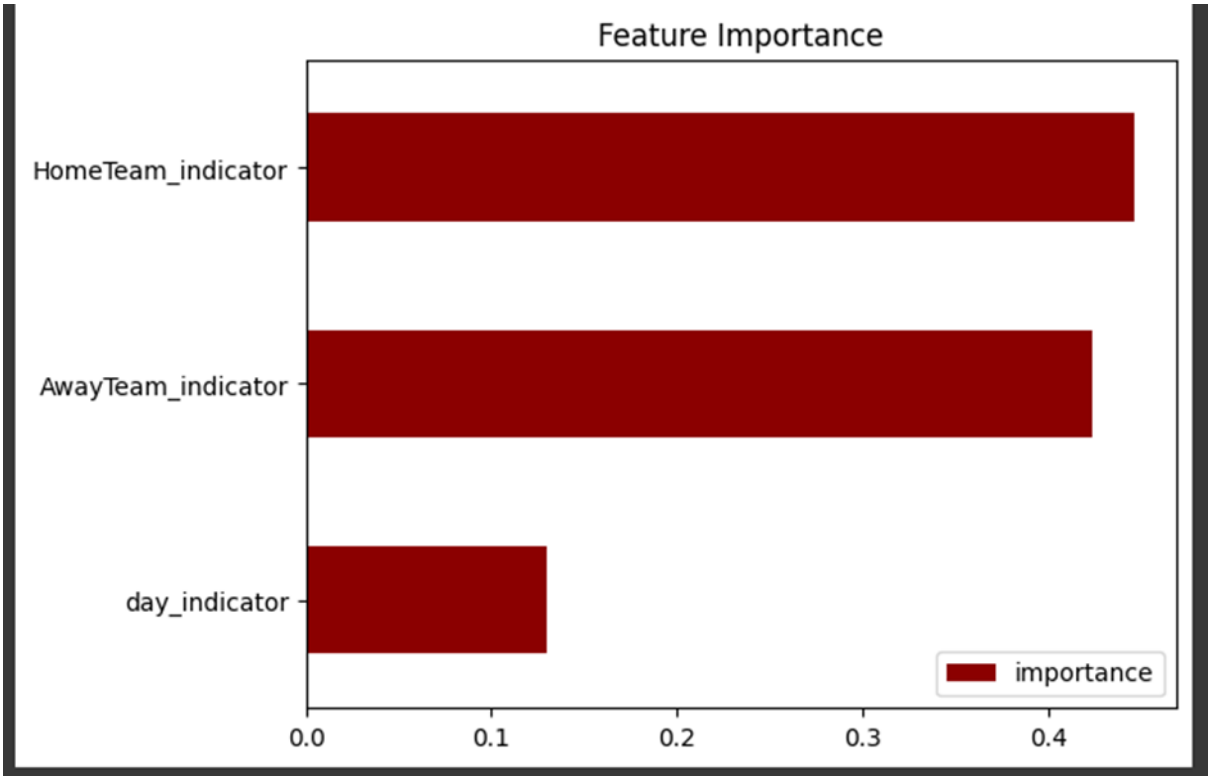
	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	...	Code_Match	day_indicator	FTHG_rolling	FTAG_rolling	HS_rolling	AS_rolling	HST_rolling	AST_rolling	HR_rolling	AR_rolling
76	E0	2019-10-06	14:00	Arsenal	Bournemouth	1	0	H	1	0	...	False	6	2.333333	1.666667	21.000000	15.000000	7.666667	7.666667	0.333333	0.000000
97	E0	2019-10-27	16:30	Arsenal	Crystal Palace	2	2	D	2	1	...	False	6	2.000000	1.333333	19.666667	12.333333	5.333333	6.666667	0.333333	0.000000
101	E0	2019-11-02	15:00	Arsenal	Wolves	1	1	D	1	0	...	False	5	2.000000	1.333333	16.000000	11.333333	4.666667	5.000000	0.333333	0.000000
121	E0	2019-11-23	15:00	Arsenal	Southampton	2	2	D	1	1	...	False	5	1.333333	1.000000	12.333333	15.000000	4.000000	4.666667	0.000000	0.000000
149	E0	2019-12-05	20:15	Arsenal	Brighton	1	2	A	0	1	...	False	3	1.666667	1.666667	12.333333	18.666667	5.000000	6.000000	0.000000	0.000000
168	E0	2019-12-15	16:30	Arsenal	Man City	0	3	A	0	3	...	False	6	1.333333	1.666667	11.333333	22.000000	4.666667	7.666667	0.000000	0.000000
196	E0	2019-12-29	14:00	Arsenal	Chelsea	1	2	A	1	0	...	False	6	1.000000	2.333333	10.000000	18.333333	3.666667	7.333333	0.000000	0.000000
207	E0	2020-01-01	20:00	Arsenal	Man United	2	0	H	2	0	...	False	2	0.666667	2.333333	8.333333	15.666667	2.666667	6.666667	0.000000	0.000000
220	E0	2020-01-18	15:00	Arsenal	Sheffield United	1	1	D	1	0	...	False	5	1.000000	1.666667	7.666667	12.333333	2.333333	5.000000	0.000000	0.000000
257	E0	2020-02-16	16:30	Arsenal	Newcastle	4	0	H	0	0	...	False	6	1.333333	1.000000	9.333333	11.666667	3.333333	4.000000	0.000000	0.000000
268	E0	2020-02-23	16:30	Arsenal	Everton	3	2	H	2	2	...	False	6	2.333333	0.333333	12.000000	10.666667	5.000000	3.333333	0.000000	0.000000

Εδώ αξίζει να τονιστεί ότι δοκιμάζουμε ξανά τον αλγόριθμο Random Forest με τους νέους εκτιμητές, αλλά δεν καταλήγουμε σε καλύτερο αποτέλεσμα.

Επίσης, αξίζει να αναφερθεί η σημασία κάθε εκτιμητή καθότι αν και η προσθήκη αυτών των εκτιμητών δεν οδήγησε σε βελτίωση του αποτελέσματος με τον αλγόριθμο Random Forest, η ανάλυση των δεδομένων που παρέχουν μπορεί να προσφέρει πολύτιμες γνώσεις και να βοηθήσει στη βελτιστοποίηση άλλων αλγορίθμων ή την ανάπτυξη στρατηγικών βελτιστοποίησης της ομάδας.

Παρατηρούμε ένα διάγραμμα που απεικονίζει τη σημασία κάθε εκτιμητή, υποδεικνύοντας τον καθοριστικό ρόλο του καθενός στο μοντέλο μας.

Εικόνα 4.19 Διάγραμμα με Ποσοστό Σημαντικότητας ανά Δείκτη



4.3.3 Αλγόριθμος Multi Layer Perceptron Classifier

Ο MLPClassifier (Multi-layer Perceptron Classifier) είναι ένας τύπος ταξινομητή νευρωνικών δικτύων στη βιβλιοθήκη scikit-learn. Βασίζεται στην αρχιτεκτονική των τεχνητών νευρωνικών δικτύων, και συγκεκριμένα σε ένα νευρωνικό δίκτυο με προώθηση προς τα εμπρός (feedforward neural network).

Ο MLPClassifier αποτελείται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου. Τα επίπεδα είναι πλήρως συνδεδεμένα, που σημαίνει ότι κάθε νευρώνας σε ένα επίπεδο συνδέεται με κάθε νευρώνα στο επόμενο επίπεδο (What Is a Multilayer Perceptron (MLP) Neural Network?, 2024).

Όπως πραγματοποιήθηκε για τον αλγόριθμο random forest , έτσι και για τον αλγόριθμο MLPClassifier θα εξετάσουμε δύο (2) προσεγγίσεις. Η πρώτη προσέγγιση (Εικόνα 4.20) για τον αλγόριθμο MLPClassifier εστιάζει επίσης στην ακρίβεια των προβλέψεων. Ο στόχος είναι να βελτιστοποιηθεί η συνολική απόδοση του μοντέλου σε όλες τις κλάσεις,

χωρίς να δίνεται έμφαση σε κάποια συγκεκριμένη κλάση. Αυτή η προσέγγιση είναι κατάλληλη για περιπτώσεις όπου η ακρίβεια της πρόβλεψης είναι ο βασικός στόχος και επιδιώκεται η γενική βελτίωση της απόδοσης του μοντέλου..Ενώ η δεύτερη προσέγγιση επικεντρώνεται στην αξιολόγηση της απόδοσης της ομάδας μέσω της φόρμας της. Αναπτύσσουμε μια συνάρτηση που εκτιμά την τρέχουσα φόρμα της ομάδας, χρησιμοποιώντας δεδομένα από προηγούμενες χρονικές περιόδους για να αναγνωρίσουμε τάσεις και μοτίβα. Η συνάρτηση αυτή αξιολογεί την ακρίβεια των προβλέψεων για την ομάδα, λαμβάνοντας υπόψη αν η ομάδα βρίσκεται σε καλή φόρμα ή όχι. Με αυτή την προσέγγιση, εξετάζουμε αν η γενική απόδοση του μοντέλου βελτιώνεται όταν λαμβάνονται υπόψη οι εκτιμήσεις για τη φόρμα της ομάδας.

1η προσέγγιση του MLP Classifier

Εύρεση βέλτιστης ακρίβειας με βάση την απόδοση του μοντέλου

Εικόνα 4.20 Multi Layer Perceptron Classifier

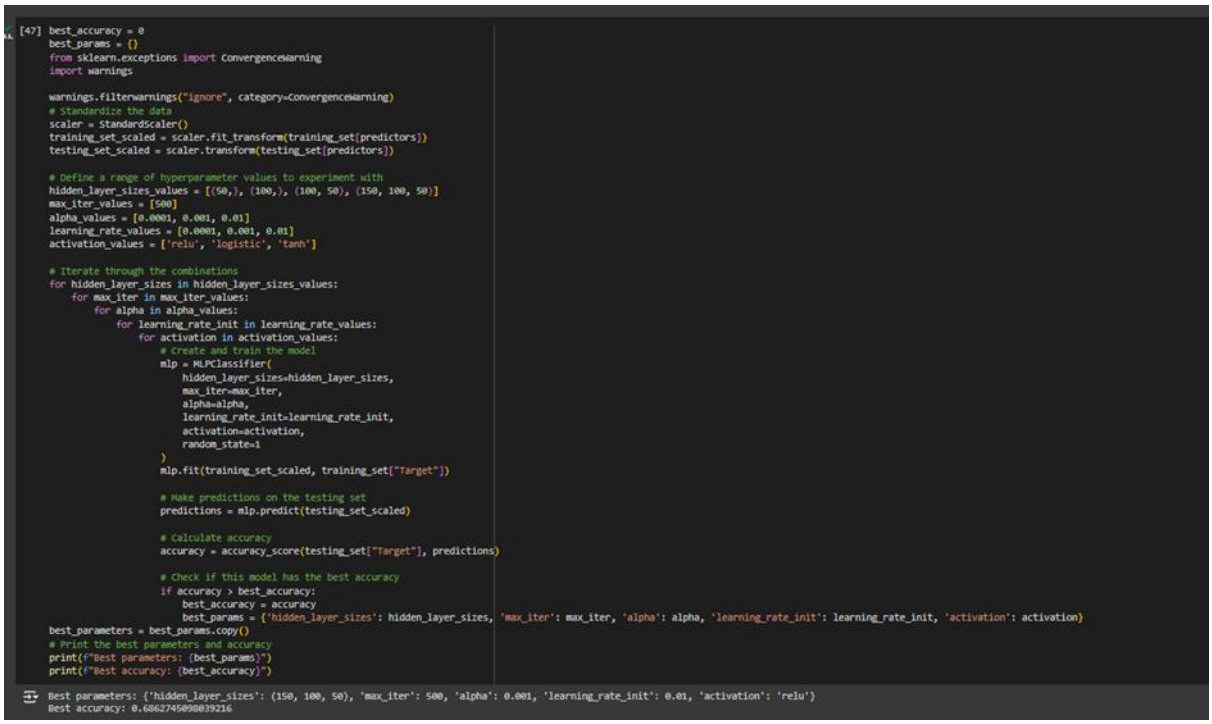
```
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

#Standardize the data
scaler = StandardScaler()
training_set_scaled = scaler.fit_transform(training_set[predictors])
testing_set_scaled = scaler.transform(testing_set[predictors])

#Define and create the MLP model
mlp = MLPClassifier(hidden_layer_sizes=(100,), max_iter=500, activation='relu', random_state=1)

#Train the model
mlp.fit(training_set_scaled, training_set["Target"])

#Make predictions on the testing set
predictions = mlp.predict(testing_set_scaled)

#Calculate and print th
accuracy = accuracy_score(testing_set["Target"], predictions)
print(f"MLP Accuracy: {accuracy}")
```

```
MLP Accuracy: 0.5294117647058824
```

Στη συνέχεια, για τον αλγόριθμο MLP Classifier, θα εξερευνήσουμε τις παραμέτρους που μπορούν να βελτιστοποιήσουν την απόδοσή του μοντέλου.

Κάθε αλγόριθμος μηχανικής μάθησης προσφέρει ένα σύνολο παραμέτρων που μπορούμε να προσαρμόσουμε για να επιτύχουμε βέλτιστα αποτελέσματα. Αναζητούμε τις συγκεκριμένες παραμέτρους που μπορούν να μεγιστοποιήσουν την ακρίβεια του μοντέλου, παρατηρώντας ποιες παράμετροι έχουν σημαντική επίδραση σε αυτό το είδος αλγορίθμου.

Εικόνα 4.21 Πειραματισμός Παραμέτρων Multi Layer Perceptron Classifier



```
[47] best_accuracy = 0
     best_params = {}
     from sklearn.exceptions import ConvergenceWarning
     import warnings

     warnings.filterwarnings("ignore", category=ConvergenceWarning)
     # Standardize the data
     scaler = StandardScaler()
     training_set_scaled = scaler.fit_transform(training_set[predictors])
     testing_set_scaled = scaler.transform(testing_set[predictors])

     # Define a range of hyperparameter values to experiment with
     hidden_layer_sizes_values = [(50,), (100,), (100, 50), (150, 100, 50)]
     max_iter_values = [500]
     alpha_values = [0.0001, 0.001, 0.01]
     learning_rate_values = [0.0001, 0.001, 0.01]
     activation_values = ['relu', 'logistic', 'tanh']

     # Iterate through the combinations
     for hidden_layer_sizes in hidden_layer_sizes_values:
         for max_iter in max_iter_values:
             for alpha in alpha_values:
                 for learning_rate_init in learning_rate_values:
                     for activation in activation_values:
                         # Create and train the model
                         mlp = MLPClassifier(
                             hidden_layer_sizes=hidden_layer_sizes,
                             max_iter=max_iter,
                             alpha=alpha,
                             learning_rate_init=learning_rate_init,
                             activation=activation,
                             random_state=1
                         )
                         mlp.fit(training_set_scaled, training_set["Target"])

                         # Make predictions on the testing set
                         predictions = mlp.predict(testing_set_scaled)

                         # Calculate accuracy
                         accuracy = accuracy_score(testing_set["Target"], predictions)

                         # Check if this model has the best accuracy
                         if accuracy > best_accuracy:
                             best_accuracy = accuracy
                             best_params = {'hidden_layer_sizes': hidden_layer_sizes, 'max_iter': max_iter, 'alpha': alpha, 'learning_rate_init': learning_rate_init, 'activation': activation}
     best_parameters = best_params.copy()
     # Print the best parameters and accuracy
     print(f"Best parameters: {best_params}")
     print(f"Best accuracy: {best_accuracy}")

Best parameters: {'hidden_layer_sizes': (150, 100, 50), 'max_iter': 500, 'alpha': 0.001, 'learning_rate_init': 0.01, 'activation': 'relu'}
Best accuracy: 0.6862745098039216
```

Εδώ σημειώνεται ότι αυτό το σύνολο των παραμέτρων φέρνει μια εξαιρετικά και μεγάλη διαφορά στην ακρίβεια περισσότερο από το προηγούμενο μοντέλο που - πειραματιστήκαμε random forest.

2η προσέγγιση του MLP Classifier

Με την προσέγγιση μια συνάρτησης που θα επεξεργάζεται τα δεδομένα κάθε ομάδας για μια συγκεκριμένη χρονική περίοδο, καταγράφοντας την τρέχουσα απόδοσή της. Αυτή η συνάρτηση θα εξετάζει αν η ομάδα βρίσκεται σε φόρμα ή όχι και θα χρησιμοποιεί τους

εκτιμητές μας για να καταγράφει πληροφορίες βάσει της φόρμας της ομάδας. Με βάση αυτή την λογική καταλήψαμε στο εξής αποτέλεσμα:

Best accuracy: 0.6666

Συχνά, παρέχουμε περισσότερες πληροφορίες στο μοντέλο μας με την ελπίδα να βελτιώσουμε την ακρίβειά του. Ωστόσο, παρά τις προσπάθειές μας, αυτή η προσέγγιση δεν πάντα οδηγεί σε καλύτερα αποτελέσματα.

4.3.4 Αλγόριθμος XGBoost

Το XGBoost είναι μια βελτιστοποιημένη βιβλιοθήκη ανάδειξης κλίσης που σχεδιάστηκε για να είναι υψηλά αποδοτική, ευέλικτη και φορητή (Εικόνα 4.22). Υλοποιεί αλγορίθμους μηχανικής μάθησης μέσω του πλαισίου Gradient Boosting (Masui, 2024). Το XGBoost παρέχει ένα παράλληλο δέντρο ενίσχυσης (επίσης γνωστό ως GBDT, GBM) που λύνει πολλά προβλήματα επιστήμης δεδομένων με γρήγορο και ακριβές τρόπο. Ο ίδιος κώδικας εκτελείται σε κύρια κατανεμημένα περιβάλλοντα (Hadoop, SGE, MPI) και μπορεί να λύσει προβλήματα πέρα από τα δισεκατομμύρια παραδείγματα (Chen & Guestrin, 2016).

Εικόνα 4.22 XGBoost Classifier

```
import xgboost as xgb
from sklearn.metrics import mean_squared_error #metric
from sklearn.metrics import r2_score
from sklearn.model_selection import GridSearchCV

training_set = df[df["Date"] < proper_date]
testing_set = df[df["Date"] >= proper_date]

X_train = training_set[['HomeTeam_indicator', 'AwayTeam_indicator', 'day_indicator']]
y_train = training_set["Target"]

X_test = testing_set[['HomeTeam_indicator', 'AwayTeam_indicator', 'day_indicator']]
y_test = testing_set["Target"]

train = xgb.DMatrix(X_train, label =y_train) #
test = xgb.DMatrix(X_test, label =y_test) #
#Convert our Data into DMatrix format because the model expects that
```

We can define our hyper parameters values, which it is the most crucial part and can give us many differt numbers of accuracy for each set of hyperparameters.

Basically our main job our target job is to find the best hypereparameters for each testing model on a given dataset. Some dataset with the features can give us a good result in ony model with a number of parameters and if we change a little bit the our main dataset this can change into a another model with another set of hyperparameters.

```
[27] param_grid = {
        'max_depth': [3, 4, 5],
        'learning_rate': [0.1, 0.3, 0.5],
        'n_estimators': [50, 100, 200],
    }
```

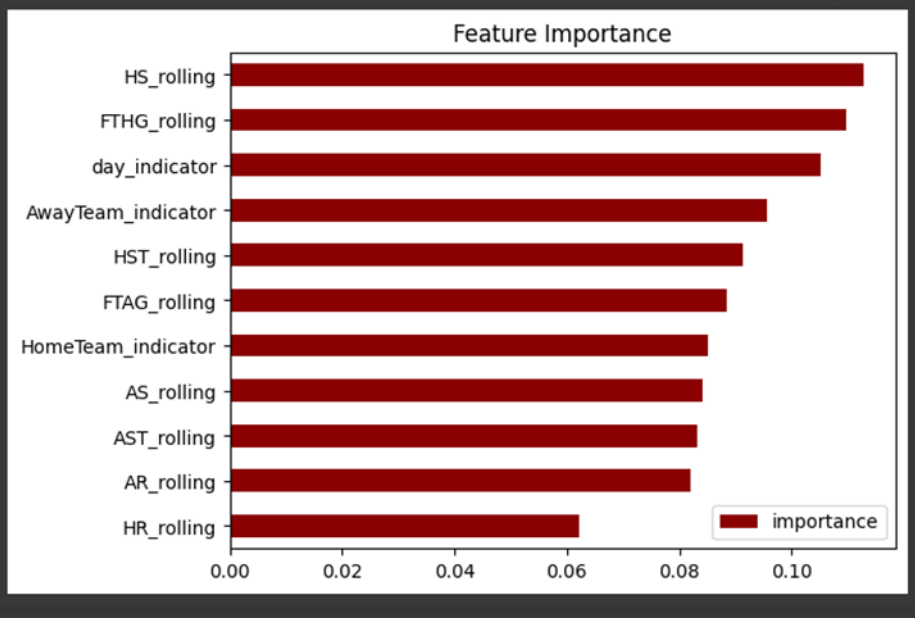
```
[29] from sklearn.metrics import accuracy_score
     model = xgb.XGBClassifier(objective='binary:logistic', eval_metric='logloss')
     #objective='binary:logistic' specifies that the model is for a binary classification
     grid_search = GridSearchCV(model, param_grid, scoring='accuracy', cv=3)
     #ridSearchCV will perform a search over the specified hyperparameter grid, using 3-fold cross-validation and optimizing for accuracy.
     #Fit the model to the training data
     grid_search.fit(X_train, y_train)
     #Get the best parameters from the grid search
     best_params = grid_search.best_params_
     print("Best Hyperparameters:", best_params)
     #Predict on the test set using the best model
     best_model = grid_search.best_estimator_
     predictions = best_model.predict(X_test)
```

Η καλύτερη ακρίβεια με το μοντέλο XGBoost είναι 0.6277 με παραμέτρους {learning_rate: 0.1, max_depth: 3, n_estimators: 50}

Παρατηρούμε ένα διάγραμμα που απεικονίζει τη σημασία κάθε εκτιμητή, υποδεικνύοντας τον καθοριστικό ρόλο του καθενός εκτιμητή στο μοντέλο του XGBooster.

Εικόνα 4.23 Διάγραμμα με Ποσοστό Σημαντικότητας ανά Δείκτη

```
model2 = xgb.XGBClassifier(objective='binary:logistic', eval_metric='logloss')
model2.fit(X_train, y_train)
#Create a DataFrame for feature importances
df1 = pd.DataFrame(data=model2.feature_importances_, index=X_train.columns, columns=['importance'])
#Sort by importance
df1 = df1.sort_values('importance')
bar_color = 'darkred'
title_color = 'black'
ax = df1.plot(kind='barh', title='Feature Importance', color=bar_color)
ax.title.set_color(title_color)
plt.show()
```



4.4 Σύνοψη και Συμπεράσματα

Εδώ παρουσιάζονται συνοπτικά και αξιολογούνται τα βασικά στοιχεία και χαρακτηριστικά των εφαρμογών Random Forest, MLP Classifier και XGBoost που δοκιμάστηκαν και χρησιμοποιήθηκαν στην παρούσα προσέγγιση.

Τα βασικά συμπεράσματα από την εφαρμογή του αλγορίθμου Random Forest μπορούν να συνοψιστούν ως εξής:

- Το Random Forest είναι σταθερό και ανθεκτικό στον υπερπροσδιορισμό, κάτι που μπορεί να είναι πλεονέκτημα σε σύνολα δεδομένων με ποικιλία και θόρυβο.
- Το ποσοστό ακρίβειας 62.74% δείχνει ότι το μοντέλο μπορεί να κάνει χρήσιμες προβλέψεις, αλλά υπάρχει χώρος για βελτίωση.

- Ενδέχεται να επωφεληθεί από βελτιστοποίηση παραμέτρων ή από ενσωμάτωση επιπλέον χαρακτηριστικών δεδομένων.

Τα βασικά συμπεράσματα από την εφαρμογή του μοντέλου MLP Classifier μπορούν να συνοψιστούν ως εξής:

- Το MLP Classifier αποδεικνύεται το πιο αποτελεσματικό από τα τρία μοντέλα, υποδεικνύοντας ότι η δομή του νευρωνικού δικτύου μπορεί να κατανοήσει καλύτερα τα πρότυπα στα δεδομένα.
- Το ποσοστό ακρίβειας 68.62% είναι αρκετά υψηλό, δείχνοντας ότι το μοντέλο έχει καλή ικανότητα πρόβλεψης.

Η απόδοσή του μπορεί να βελτιωθεί περαιτέρω με προσεκτική ρύθμιση των παραμέτρων (π.χ. αριθμός επιπέδων, νευρώνες ανά επίπεδο) και ίσως με την αύξηση του μεγέθους του εκπαιδευτικού συνόλου δεδομένων.

Τα βασικά συμπεράσματα από την εφαρμογή του μοντέλου XGBoost μπορούν να συνοψιστούν ως εξής:

- Το XGBoost είναι γνωστό για την ταχύτητα και την απόδοσή του, καθώς και για την ικανότητά του να χειρίζεται μεγάλες ποσότητες δεδομένων και πολύπλοκα χαρακτηριστικά.
- Η ακρίβεια 62.77% υποδεικνύει ότι το μοντέλο έχει καλές προοπτικές, αν και δεν υπερέχει έναντι του MLP Classifier.
- Βελτιώσεις στην απόδοση του XGBoost μπορεί να επιτευχθούν με fine-tuning παραμέτρων όπως το learning rate, το μέγεθος των δέντρων και το number of boosting rounds.

Με βάση τα αποτελέσματα ακρίβειας από την αξιολόγηση διαφορετικών μοντέλων σε δεδομένα αγώνων ποδοσφαίρου της αγγλικής Premier League, ο ταξινομητής MLP αναδείχθηκε ο κορυφαίο μοντέλο με ακρίβεια 0,6862. Αυτό ξεπερνά την ακρίβεια τόσο των μοντέλων Random Forest όσο και XGBoost . Δεδομένων αυτών των αποτελεσμάτων, συνιστούμε τη χρήση του ταξινομητή MLP για εκπαίδευση και τη δημιουργία

προβλεπόμενων αποτελεσμάτων για μελλοντικούς αγώνες, καθώς έχει επιδείξει ανώτερη απόδοση στο σύνολο δεδομένων μας.

5. Διερεύνηση Σχέσης Αποτελέσματος – Οπαδών

5.1 Διαδικασία Διερεύνησης Σχέσης Αποτελέσματος – Οπαδών

Το δεύτερο στάδιο της παρούσας ανάλυσης και επεξεργασίας σχετικά με τους αγώνες ποδοσφαίρου του Αγγλικού Πρωταθλήματος - English Premier League είναι να ερευνηθεί αν υπάρχει σχέση μεταξύ του αποτελέσματος μιας ομάδας και των οπαδών της. Πιο συγκεκριμένα, θα επικεντρωθούμε στους οπαδούς από την πλατφόρμα μέσων κοινωνικής δικτύωσης (social media) του Twitter. Στην ουσία, θα πραγματοποιήσουμε ανάλυση συναισθήματος από δεδομένα του Twitter για να διαπιστώσουμε πως οι οπαδοί – fans (φανς) μιλάνε και εκφράζουν την υποστήριξή τους στην ομάδα. Στη συνέχεια, θα συγκρίνουμε αυτά τα συναισθηματικά δεδομένα με τα πραγματικά αποτελέσματα που έχουν φέρει οι ομάδες. Με αυτόν τον τρόπο, σαν αποτέλεσμα θα μπορέσουμε να ερευνήσουμε αν τα δεδομένα της πλατφόρμας του twitter έχουν μια σημαντική ή ασήμαντη συσχέτιση με τα αποτελέσματα που φέρνει η εκάστοτε ομάδα του Αγγλικού Πρωταθλήματος - English Premier League (EPL).

Για την διαδικασία διερεύνησης της σχέσης αποτελέσματος – οπαδών χρησιμοποιείται σύνολο δεδομένων από την πλατφόρμα μέσων κοινωνικής δικτύωσης (social media) του Twitter. Συγκεκριμένα, έχουμε εντοπίσει ένα σύνολο δεδομένων από το Twitter, το οποίο αφορά την Premier League (EPL) και είναι διαθέσιμο στην πλατφόρμα του Kaggle στον παρακάτω σύνδεσμο: https://www.kaggle.com/datasets/wjia26/epl-teams-twitter-sentiment-dataset/data (Kaggle: Your Machine Learning and Data Science Community, n.d.).

Η διαδικασία της διερεύνησης της σχέσης αποτελέσματος – οπαδών μπορεί να συνοψιστεί σε τέσσερα γενικά βήματα, συγκεκριμένα, την Επισκόπηση και Ανάλυση Δεδομένων (1), την Επεξεργασία Δεδομένων (2), την Επιλογή Μοντέλου για Ανάλυση Συναισθήματος (3) και την Ανάλυση Συσχέτισης Μεταξύ Υποστήριξης Οπαδών και Νικηφόρων Αγώνων (4).

5.1.1 Επισκόπηση και Ανάλυση Δεδομένων

Ως πρώτο βήμα, θα ανεβάσουμε τα δεδομένα για να εξετάσουμε τις διαθέσιμες πληροφορίες και να καθορίσουμε πώς μπορούμε να τις αξιοποιήσουμε καλύτερα. Τα tweets που περιλαμβάνονται στο dataset καλύπτουν την περίοδο από 9 Ιουλίου 2020 έως 13 Οκτωβρίου 2020 (Εικόνα 5.1).

Μία απεικόνιση των δεδομένων παρουσιάζεται παρακάτω.

Εικόνα 5.1 Απεικόνιση Δεδομένων

Date_Created	Team_Name	Followers	group_name	Retweets_Count	Text
2020-07-09	Liverpool	697325	Liverpool FC	0.0	This is a strange claim #LFC https://t.co/U1...
2020-07-09	Liverpool	2348	Liverpool FC	65.0	RT @TheKopiteOFF: 🔥| #LFC have won 30 of their...
2020-07-09	Liverpool	465	Liverpool FC	NaN	#liverpoolfc OR #YNWA OR #LFC
2020-07-09	Liverpool	334	Liverpool FC	0.0	Outrageous... Poor auld Martin Tyler has to ju...
2020-07-09	Liverpool	760	Liverpool FC	NaN	#liverpoolfc OR #YNWA OR #LFC
...	...	...	...	...	...
2020-09-30	Everton	10962	Everton FC	0.0	BTTS, Each Team Over 3 Corners & Each Team...
2020-09-30	Everton	3439	Everton FC	0.0	Matchday ! 💙 #UpTheToffees #EFC #COYB
2020-09-30	Everton	0	Everton FC	0.0	First post!! Come ride along and enjoy the bea...
2020-09-30	Everton	31	Everton FC	1.0	New home on the way..... #EFC #Everton #Spirit...
2020-09-30	Everton	2829	Everton FC	0.0	Echo: Ancelotti names area of Everton's passin...

Τα δεδομένα που παρουσιάζουμε περιέχουν τις εξής πληροφορίες:

Date_Created: Η ημερομηνία δημιουργίας του tweet.

Team_Name: Η ομάδα στην οποία απευθύνεται το tweet.

Followers: Ο αριθμός των followers του προφίλ που δημοσίευσε το tweet.

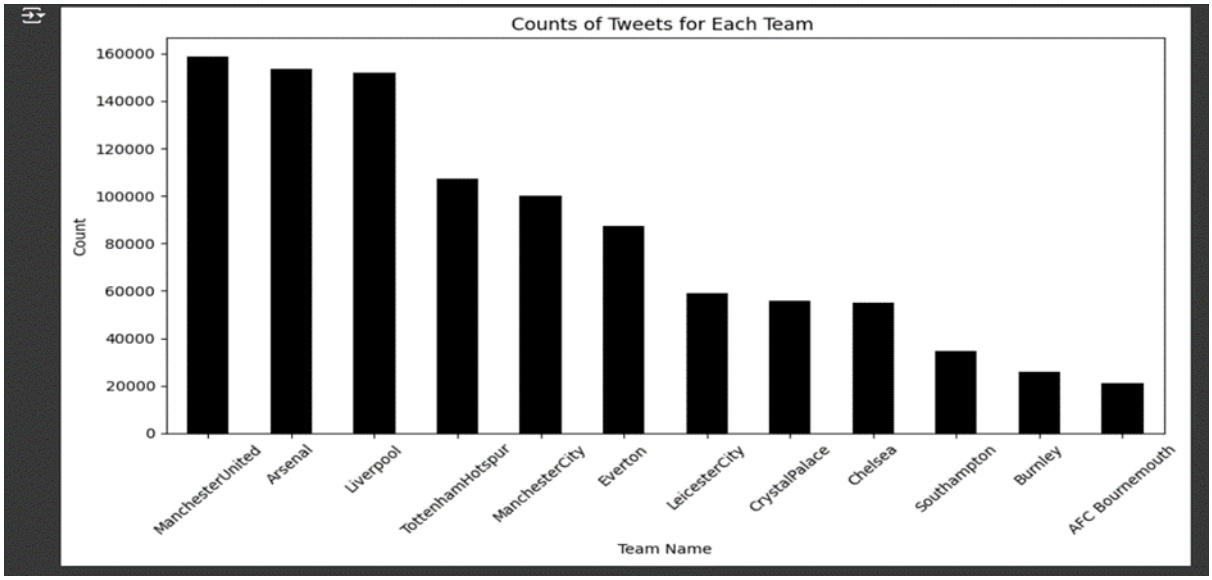
Group_Name: Η ομάδα στην οποία απευθύνεται το tweet (επαναλαμβανόμενο πεδίο).

Retweets_Count: Ο αριθμός των retweets που έχει λάβει το συγκεκριμένο tweet (η αναδημοσίευση του tweet).

Text: Το κείμενο του tweet.

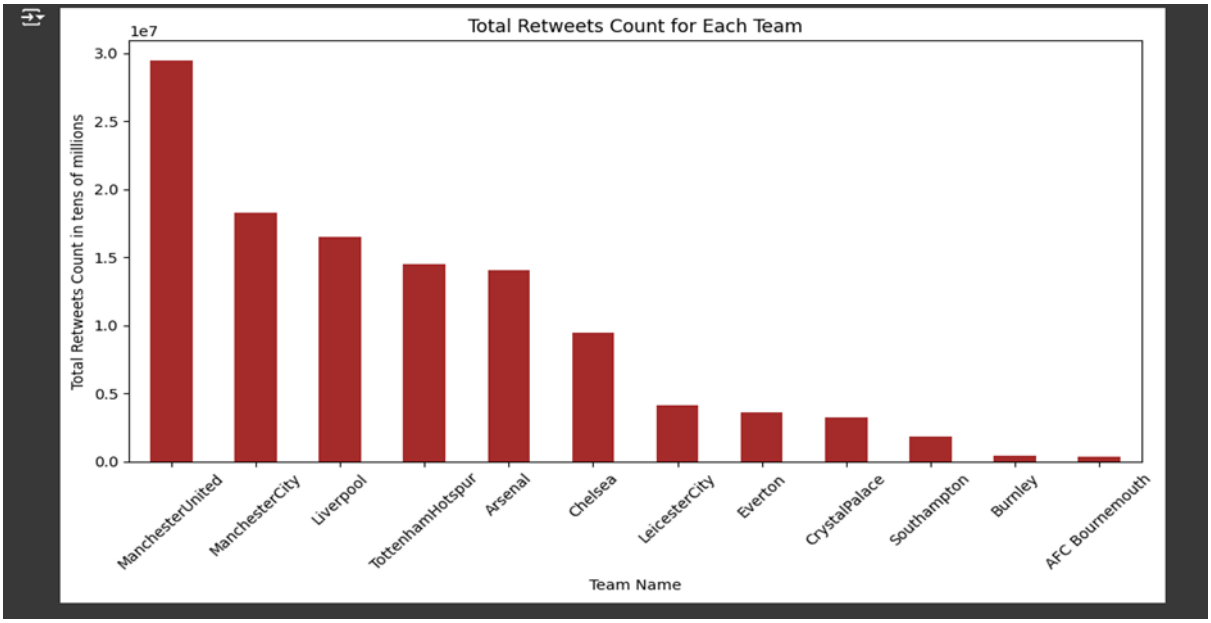
Στην παρακάτω απεικόνιση (Εικόνα 5.2) παρουσιάζεται ο συνολικός αριθμός των tweets που έχουμε για κάθε ομάδα, για την περίοδο από 9 Ιουλίου 2020 έως 13 Οκτωβρίου 2020. Συνολικά, έχουμε δεδομένα από 96 ημέρες για τις παρακάτω ομάδες.

Εικόνα 5.2 Διάγραμμα με Καταμέτρηση των Tweet ανά Ομάδα



Στην παρακάτω απεικόνιση (Εικόνα 5.3) παρουσιάζεται ο συνολικός αριθμός των retweets που έχουμε για κάθε ομάδα, για την περίοδο από 9 Ιουλίου 2020 έως 13 Οκτωβρίου 2020.

Εικόνα 5.3 Διάγραμμα με Καταμέτρηση των Retweet ανά Ομάδα



5.1.2 Επεξεργασία Δεδομένων

Όπως προαναφέρθηκε, σαν δεύτερο βήμα στην διαδικασία διερεύνησης της σχέσης αποτελέσματος – οπαδών είναι η επεξεργασία των δεδομένων. Ειδικότερα, σαν επόμενο βήμα θα περάσουμε το Text πεδίο (δηλαδή, το tweet) από κάποιες συναρτήσεις ώστε να γίνει όσο πιο απλό χωρίς να αλλοιώσουμε το περιεχόμενο για να το περάσουμε στο μοντέλο της ανάλυσης συναισθήματος που έχουμε επιλέξει (Εικόνα 5.4) . Αυτή η διαδικασία περιγράφεται ως εξής:

1. Κατεβάζουμε τις stopwords από το nltk: Αυτό το βήμα κατεβάζει μια λίστα από κοινές λέξεις (όπως "and", "the", "is") που συχνά δεν προσθέτουν νόημα στην ανάλυση και αφαιρούνται για να βελτιωθεί η ακρίβεια της επεξεργασίας κειμένου.

2. Ελέγχουμε για κενές τιμές στο dataset: Αυτή η εντολή ελέγχει αν υπάρχουν κενές τιμές στα δεδομένα μας και πόσες υπάρχουν σε κάθε στήλη. Είναι σημαντικό να γνωρίζουμε αν υπάρχουν ελλείποντα δεδομένα για να τα διαχειριστούμε κατάλληλα.

3. Ορίζουμε έναν PorterStemmer: Ο PorterStemmer είναι ένα εργαλείο που μετατρέπει τις λέξεις στις ρίζες τους. Για παράδειγμα, οι λέξεις "running", "runs", και "ran" θα μετατραπούν όλες στη ρίζα "run". Αυτό βοηθά στην ενοποίηση των διαφορετικών μορφών μιας λέξης για καλύτερη ανάλυση (IBM Developer, n.d.).

4. Δημιουργούμε μια συνάρτηση για stemming και καθαρισμό του κειμένου.

Εικόνα 5.4 Δημιουργία Συνάρτησης Stem για Καθαρισμό Κειμένου

```
def stem(content):
  stem_content = re.sub('[^a-zA-Z]',' ', content)
  stem_content = stem_content.lower()
  stem_content = stem_content.split()
  stem_content = [port_stem.stem(word) for word in stem_content if not word in stopwords.words('english')]
  stem_content = ' '.join(stem_content)

  return stem_content
```

- Αφαιρεί όλα τα μη αλφαβητικά χαρακτήρες.
- Μετατρέπει το κείμενο σε πεζά και το χωρίζει σε λέξεις.
- Εφαρμόζει stemming και αφαιρεί τις stopwords.
- Ενώνει τις λέξεις πίσω σε κείμενο.
- Εφαρμόζουμε τη συνάρτηση στα tweets.

Διαγράφουμε το αρχικό πεδίο "Text", το οποίο περιέχει τα ακατέργαστα tweets, και μετονομάζουμε το πεδίο "stemmed_content" σε "Text", ώστε να περιέχει πλέον το επεξεργασμένο κείμενο.

5.1.3 Επιλογή Μοντέλου για Ανάλυση Συναισθήματος

Κρίσιμης σημασίας στοιχείο στην διαδικασία της διερεύνησης της σχέσης αποτελέσματος – είναι η επιλογή μοντέλου για ανάλυση συναισθήματος. Ειδικότερα, το μοντέλο που έχουμε επιλέξει για να κάνουμε την ανάλυση συναισθήματος είναι το Roberta. Το μοντέλο twitter-roberta-base-sentiment από το Cardiff NLP στο Hugging Face είναι ένα RoBERTa-base μοντέλο εκπαιδευμένο σε περίπου 58 εκατομμύρια tweets και βελτιστοποιημένο για ανάλυση συναισθήματος χρησιμοποιώντας το TweetEval benchmark (Siebert/Sentiment-roberta-large-english · Hugging Face, n.d.). Είναι κατάλληλο για αγγλικά και ταξινομεί τα tweets σε τρεις κατηγορίες: αρνητικό, ουδέτερο, και θετικό. Το μοντέλο μπορεί να χρησιμοποιηθεί μέσω της βιβλιοθήκης Transformers για διάφορες εφαρμογές ανάλυσης συναισθήματος (Deniz et al., 2022). Περισσότερα μπορούμε να βρούμε στο site του hugging face : https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment .

Στη συνέχεια, αφού επεξεργαστούμε όλα τα δεδομένα μας με το μοντέλο RoBERTa, θα δημιουργηθούν τρεις νέες στήλες: roberta_positive, roberta_negative, και roberta_neutral (Εικόνα 5.5) . Κάθε στήλη θα περιέχει μια τιμή από 0 έως 1, που αντιπροσωπεύει την ισχύ της αντίστοιχης συναισθηματικής κατηγορίας για κάθε tweet. Αυτές οι τιμές δείχνουν πόσο πιθανό είναι το tweet να είναι θετικό, αρνητικό ή ουδέτερο. Η παρακάτω εικόνα παρουσιάζει το αποτέλεσμα.

Εικόνα 5.5 Απεικόνιση Αποτελεσμάτων από το Μοντέλο Roberta

	roberta_negative	roberta_neutral	roberta_positive	Date_Created	Team_Name	Retweets_Count	Text
0	0.114547	0.841231	0.044222	2020-07-09	Liverpool	0.0	strang claim lfc http co u cj w xj
1	0.005433	0.499352	0.495215	2020-07-09	Liverpool	65.0	rt thekopiteoff lfc premier leagu game season ...
2	0.092521	0.813406	0.094074	2020-07-09	Liverpool	NaN	liverpoolfc ynwa lfc
3	0.299847	0.667607	0.032545	2020-07-09	Liverpool	0.0	outrag poor auld martin tyler ignor word ynwa ...
4	0.092521	0.813406	0.094074	2020-07-09	Liverpool	NaN	liverpoolfc ynwa lfc
...	...	...	...	...	...	...	...
1010577	0.050368	0.903607	0.046026	2020-09-30	Everton	0.0	btt team corner amp team card massiv william h...
1010578	0.054720	0.851688	0.093592	2020-09-30	Everton	0.0	matchday upthetoffe efc coyb
1010579	0.002694	0.159032	0.838274	2020-09-30	Everton	0.0	first post come ride along enjoy beauti game f...
1010580	0.019193	0.842431	0.138376	2020-09-30	Everton	1.0	new home way efc everton spiritoftheblu premie...
1010581	0.088594	0.875386	0.036020	2020-09-30	Everton	0.0	echo ancelotti name area everton pass insist m...

1010582 rows × 7 columns

Το χρονικό διάστημα που θα επικεντρωθούμε είναι ένα υποσύνολο από των αγώνων που χρησιμοποιήσαμε στην παράγραφο 4 σχετικά με την σύγκριση μοντέλων μηχανικής μάθησης. Αυτό το διάστημα είναι από 9 Ιουλίου 2020 έως 13 Οκτωβρίου 2020, κατά το οποίο διαθέτουμε δεδομένα tweets (Εικόνα 5.6). Αυτά τα δεδομένα θα χρησιμοποιηθούν για την ανάλυσή μας και περιλαμβάνουν τους αγώνες που πραγματοποιήθηκαν κατά τη διάρκεια αυτής της περιόδου, καθώς και τα αποτελέσματά τους. Βλέπουμε ότι το σύνολο των αγώνων που πραγματοποιήθηκαν είναι 81.

Εικόνα 5.6 Περιορισμός Συνόλου με Βάση Αγώνων που Πραγματοποιήθηκαν σε ένα Εύρος

```
[ ] matches['Date'] = pd.to_datetime(matches['Date']) #Convert the date
    #Filter the Dataframe
    final_matches = matches[(matches['Date'] >= '2020-07-09') & (matches['Date'] <= '2020-10-13')] #Slice of the whole dataframe
```

```
final_matches
```

	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	...	PCAHA	MaxCAHH	MaxCAHA	AvgCAHH	AvgCAHA	Target	HomeTeam_indicator	AwayTeam_indicator	Code_Match	day_indicator
337	E0	2020-07-09	18:00	Bournemouth	Tottenham	0	0	D	0	0	...	2.02	2.00	2.08	1.92	1.97	0	2	18	False	3
338	E0	2020-07-09	18:00	Everton	Southampton	1	1	D	1	1	...	2.13	1.86	2.16	1.80	2.10	0	7	17	False	3
339	E0	2020-07-09	20:15	Aston Villa	Man United	0	3	A	0	2	...	2.00	1.99	2.03	1.92	1.96	0	1	13	False	3
340	E0	2020-07-11	12:30	Norwich	West Ham	0	4	A	0	2	...	2.11	1.89	2.11	1.83	2.05	0	15	21	False	5
341	E0	2020-07-11	12:30	Watford	Newcastle	2	1	H	0	1	...	2.01	1.97	2.01	1.92	1.96	1	19	14	False	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
413	E0	2020-10-04	12:00	Southampton	West Brom	2	0	H	1	0	...	2.06	1.89	2.09	1.85	2.03	1	17	20	False	6
414	E0	2020-10-04	14:00	Arsenal	Sheffield United	2	1	H	0	0	...	2.09	1.89	2.21	1.83	2.06	1	0	16	False	6
415	E0	2020-10-04	14:00	Wolves	Fulham	1	0	H	0	0	...	1.85	2.12	1.86	2.05	1.83	1	22	8	False	6
416	E0	2020-10-04	16:30	Man United	Tottenham	1	6	A	1	4	...	2.11	1.85	2.14	1.80	2.08	0	13	18	False	6
417	E0	2020-10-04	19:15	Aston Villa	Liverpool	7	2	H	4	1	...	2.00	2.03	2.01	1.93	1.95	1	1	11	False	6

81 rows × 111 columns

tweets : dataset που περιέχει δεδομένα tweets από την πλατφόρμα του twitter

final_matches : σύνολο αγώνων με διάφορες πληροφορίες του αγώνα που αντιστοιχεί στο εύρος της ημερομηνίας που έχουμε δεδομένα tweets (9 Ιουλίου 2020 μέχρι και 13 Οκτωμβρίου 2020).

Στη συνέχεια, επιθυμούμε να καθαρίσουμε το dataset των tweets, ώστε να περιέχει μόνο ουσιώδη δεδομένα, δηλαδή δεδομένα που θα χρησιμοποιήσουμε για την ανάλυση μας. Θα διατηρήσουμε τα tweets που έχουν score μεγαλύτερο του 0.3, είτε είναι αρνητικό είτε θετικό (Εικόνα 5.7). Με αυτόν τον τρόπο, θα διατηρήσουμε τα αρχεία που εκφράζουν έντονα συναισθήματα βάσει του περιεχομένου τους, απορρίπτοντας τα αρχεία που δεν εκφράζουν κάτι σημαντικό. Στόχος μας είναι να εστιάσουμε σε tweets με σαφή και έντονη συναισθηματική φόρτιση, ενισχύοντας την ποιότητα των δεδομένων μας και την αξιοπιστία των αναλύσεων που θα ακολουθήσουν.

Εικόνα 5.7 Περιορισμός Συνόλου Δεδομένων με Βάση Αποτελεσμάτων Roberta

```
tweets = tweets[(tweets['roberta_positive'] > 0.3) | (tweets['roberta_negative'] > 0.3)] #Hold the texts that has some negative or positive value so i can compare it
tweets = tweets.drop_duplicates(subset=['Text']) #Drop duplicates
tweets
```

	roberta_negative	roberta_neutral	roberta_positive	Date_Created	Team_Name	Retweets_Count	Text
1	0.005433	0.499352	0.495215	2020-07-09	Liverpool	65	rt thekopiteoff lfc premier leagu game season ...
8	0.313789	0.491151	0.195060	2020-07-09	Liverpool	0	boy
11	0.012546	0.639650	0.347804	2020-07-09	Liverpool	147	rt stanchart use standr tweet support lfc take...
12	0.007590	0.601031	0.391379	2020-07-09	Liverpool	176	rt theredmentv leagu fastest team win game yea...
38	0.002296	0.104447	0.893257	2020-07-09	Liverpool	0	power platform provid great opportun businessl...
...	...	...	...	...	...	...	...
1010550	0.009877	0.168255	0.821868	2020-09-30	Everton	5	nice one england biggest club swagger step foo...
1010557	0.333362	0.573009	0.093629	2020-09-30	Everton	0	prouder club fan bunch neg bastard want excel ...
1010559	0.586481	0.400193	0.013326	2020-09-30	Everton	0	sourc sex session report lie protest efc forwa...
1010562	0.006975	0.154482	0.838544	2020-09-30	Everton	0	love everton fan get old song team top itun ch...
1010567	0.005776	0.529912	0.464311	2020-09-30	Everton	0	everton vs westham win matchday

Θα δημιουργήσουμε μια νέα στήλη στο dataset των tweets μας με την ονομασία "Score", η οποία θα αποτελεί δείκτη της ισχυρότερης συναισθηματικής βαθμολογίας. Συγκεκριμένα, η τιμή της στήλης θα είναι -1 εάν η αρνητική τιμή που υπολογίζει το μοντέλο RoBERTa είναι μεγαλύτερη από τη θετική, και 1 στην αντίθετη περίπτωση. Με αυτόν τον τρόπο, θα μπορούμε να αναγνωρίζουμε και να αναλύουμε εύκολα τα tweets με βάση την κυρίαρχη συναισθηματική τους φόρτιση. Επιπλέον, διορθώνουμε τη στήλη "Retweets_Count" και μετατρέπουμε τις τιμές σε ακέραιο αριθμό. Για τα records όπου δεν υπάρχουν τιμές, όπως

null ή NaN (Not a Number), εισάγουμε την τιμή 0. Με αυτή τη διαδικασία, εξασφαλίζουμε ότι τα δεδομένα μας είναι ομοιόμορφα και έτοιμα για περαιτέρω ανάλυση, αποφεύγοντας προβλήματα που μπορεί να προκύψουν από ελλιπή ή μη αριθμητικά δεδομένα.

Στη συνέχεια, επιδιώκουμε να αξιοποιήσουμε τις πληροφορίες του "Retweets_Count" προς όφελός μας. Σχεδιάζουμε να πραγματοποιήσουμε μια απλή πράξη πολλαπλασιασμού του αριθμού των "Retweets_Count" με την αντίστοιχη βαθμολογία, η οποία παίρνει τιμές 1 και -1 ανάλογα με το εάν το συναίσθημα είναι θετικό ή αρνητικό, όπως είχαμε προαναφέρει (Εικόνα 5.8). Με αυτόν τον τρόπο, κάθε tweet που δημοσιεύεται στην πλατφόρμα του Twitter θα αντιστοιχίζεται με έναν ακέραιο αριθμό, ο οποίος δηλώνει πόσοι διαφορετικοί χρήστες εκφράζουν θετική ή αρνητική γνώμη για την ομάδα τους, εντός μιας συγκεκριμένης ημερομηνίας.

Εικόνα 5.8 Δημιουργία Score του Tweet

	roberta_negative	roberta_neutral	roberta_positive	Date_Created	Team_Name	Retweets_Count	Text	Score	Score_Multiplication
1	0.005433	0.499352	0.495215	2020-07-09	Liverpool	65	rt thekopiteoff lfc premier leagu game season ...	1	65
8	0.313789	0.491151	0.195060	2020-07-09	Liverpool	0	boy	-1	-1
11	0.012546	0.639650	0.347804	2020-07-09	Liverpool	147	rt stanchart use standr tweet support lfc take...	1	147
12	0.007590	0.601031	0.391379	2020-07-09	Liverpool	176	rt theredmentv leagu fastest team win game yea...	1	176
38	0.002296	0.104447	0.893257	2020-07-09	Liverpool	0	power platform provid great opportun businessl...	1	1
...	...	...	...	...	...	...	...	...	...
1010550	0.009877	0.168255	0.821868	2020-09-30	Everton	5	nice one england biggest club swagger step foo...	1	5
1010557	0.333362	0.573009	0.093629	2020-09-30	Everton	0	prouder club fan bunch neg bastard want excel ...	-1	-1
1010559	0.586481	0.400193	0.013326	2020-09-30	Everton	0	sourc sex session report lie protest efc forwa...	-1	-1
1010562	0.006975	0.154482	0.838544	2020-09-30	Everton	0	love everton fan get old song team top itun ch...	1	1
1010567	0.005776	0.529912	0.464311	2020-09-30	Everton	0	everton vs westham win matchday	1	1

133028 rows × 9 columns

Στη συνέχεια, θα ομαδοποιήσουμε τα αποτελέσματα που προέκυψαν ανά ημερομηνία και ομάδα (Εικόνα 5.9). Αυτό σημαίνει ότι θα υπολογίσουμε ένα σκορ για κάθε ομάδα, το οποίο αντιστοιχεί στο σύνολο των διαφορετικών οπαδών που εξέφρασαν θετική άποψη για την ομάδα τους. Με αυτόν τον τρόπο, δημιουργούμε μια συνολική εικόνα της θετικής αντίληψης που έχουν οι οπαδοί για την ομάδα τους, λαμβάνοντας υπόψη την ημερομηνία και την ομάδα στην οποία αναφέρονται τα tweets.

Εικόνα 5.9 Score Tweet & Retweet ανά Ημέρα & Ομάδα

```
sum_scores_tweets = tweets.groupby(['Team_Name', 'Date_Created']).agg({'Score_Multiplication': 'sum'}).reset_index()
final_tweets = sum_scores_tweets
final_tweets
```

	Team_Name	Date_Created	Score_Multiplication
0	AFC Bournemouth	2020-07-08	108
1	AFC Bournemouth	2020-07-09	73
2	AFC Bournemouth	2020-07-10	15
3	AFC Bournemouth	2020-07-11	4
4	AFC Bournemouth	2020-07-12	10
...	...	...	...
1155	TottenhamHotspur	2020-10-08	777
1156	TottenhamHotspur	2020-10-09	123
1157	TottenhamHotspur	2020-10-10	494
1158	TottenhamHotspur	2020-10-11	224
1159	TottenhamHotspur	2020-10-12	699

1160 rows × 3 columns

Σε αυτό το σημείο, λοιπόν, διαθέτουμε δύο dataframes: το "final_tweets", που περιλαμβάνει τις διάφορες ομάδες με τη βαθμολογία ανάλυσης των tweets για κάθε ημερομηνία, και το "final_matches", που περιέχει τους αγώνες που διεξήχθησαν κατά τη διάρκεια της περιόδου που εξετάζουμε και έχουν το τελικό αποτέλεσμα του αγώνα. Εάν η γηπεδούχος ομάδα (HomeTeam) κερδίσει, σημειώνεται με "A", ενώ αν η φιλοξενούμενη ομάδα (AwayTeam) κερδίσει, σημειώνεται με "B", και το "D" δηλώνει ισοπαλία.

Αυτή τη στιγμή, λοιπόν, θα δημιουργήσουμε κουβάδες για τους αγώνες.

Έχουμε τη βαθμολογία για όλες τις διαφορετικές ημερομηνίες και τις διαφορετικές ομάδες (Εικόνα 5.10).

Θα κρατήσουμε μόνο τις σχετικές τιμές για κάθε αγώνα. Συγκεκριμένα, θα κρατήσουμε τις τιμές των tweets score την ημέρα του αγώνα και τις δύο προηγούμενες ημέρες συνολικά. Κατά αυτόν τον τρόπο, τα δεδομένα που θα έχουμε θα αφορούν την ημέρα του αγώνα και θα συνυπολογίσουμε επίσης δεδομένα και τιμές που προηγούνται μέχρι και δύο ημέρες πριν. Μπορούμε επίσης να δοκιμάσουμε πειράματα για δεδομένα που αναφέρονται σε περισσότερες ή λιγότερες ημέρες πριν από τον αγώνα, αλλά για το παρόν επιλέγουμε δύο ημέρες νωρίτερα.

Εικόνα 5.10 Περιορισμός Συνόλου για Δεδομένα με Τουλάχιστον 1 Τιμή

```
#Initialize variables to track the previous team and its total value
prev_date = None
prev_team1 = None
prev_team2 = None
prev_total_value1 = 0
prev_total_value2 = 0

#Iterate through each row in final_matches_sorted
for _, row in final_matches_sorted.iterrows():
    match_date = row['Date']
    team1 = row['HomeTeam']
    team2 = row['AwayTeam']

    #Convert match_date to datetime if needed
    #if not isinstance(match_date, pd.Timestamp):
    #    match_date = pd.to_datetime(match_date)

    #If date changes, reset total value to zero
    if match_date != prev_date:
        prev_date = match_date
        prev_team = None
        prev_total_value = 0

     #If team1 changes, reset total value to zero
    if team1 != prev_team1:
        prev_team1 = team1
        prev_total_value1 = 0

    #If team2 changes, reset total value to zero
    if team2 != prev_team2:
        prev_team2 = team2
        prev_total_value2 = 0

      #Calculate the date range
    date_range = [match_date - pd.DateOffset(days=i) for i in range(3)] #To filter the tweets DataFrame to include only the rows with dates that are equal to match_date or 1 or 2 days earlier (

     #Subset of rows in final_tweets for the specified date range and HomeTeam
    subset_tweets1 = final_tweets[(final_tweets['Date_Created'].isin(date_range)) & (final_tweets['Team_Name'] == team1)]
    total_value_team1 = subset_tweets1['Score_Multiplication'].sum()

    #Subset of rows in final_tweets for the specified date range and AwayTeam
    subset_tweets2 = final_tweets[(final_tweets['Date_Created'].isin(date_range)) & (final_tweets['Team_Name'] == team2)]
    total_value_team2 = subset_tweets2['Score_Multiplication'].sum()

    #Update total value for the current HomeTeam and AwayTeam
    prev_total_value1 += total_value_team1
    prev_total_value2 += total_value_team2
```

```
    #Append the result to the result DataFrame
    # result_df_Home_Team = result_df_Home_Team.append({'Date': match_date, 'HomeTeam': team1, 'Total_Value': total_value_team1}, ignore_index=True)
    # result_df_Away_Team = result_df_Away_Team.append({'Date': match_date, 'AwayTeam': team2, 'Total_Value': total_value_team2}, ignore_index=True)
    result_df_Home_Team = pd.concat([result_df_Home_Team, pd.DataFrame([{'Date': match_date, 'HomeTeam': team1, 'Total_Value': total_value_team1}])], ignore_index=True)
    result_df_Away_Team = pd.concat([result_df_Away_Team, pd.DataFrame([{'Date': match_date, 'AwayTeam': team2, 'Total_Value': total_value_team2}])], ignore_index=True)


# df = pd.concat([df, pd.DataFrame([new_row])], ignore_index=True) #new
# df = pd.DataFrame(df).append(new_row, ignore_index=True)         #old
#Display the result DataFrame
print(result_df_Home_Team)
# print(result_df_Away_Team)
```

```
         Date      HomeTeam Total_Value
0  2020-07-09  Bournemouth         181
1  2020-07-09       Everton          93
2  2020-07-09  Aston Villa           0
3  2020-07-11       Norwich          0
4  2020-07-11       Watford          0
..        ...           ...         ...
76 2020-10-04     Leicester         -44
77 2020-10-04  Southampton          45
78 2020-10-04       Arsenal         973
79 2020-10-04        Wolves          0
80 2020-10-04  Aston Villa           0

[81 rows x 3 columns]
```

Ως αποτέλεσμα, έχουμε δύο διαφορετικά dataframes (datasets). Ένα από αυτά αντιστοιχεί στην ομάδα που είναι φιλοξενούμενη, ενώ το άλλο αντιστοιχεί στην ομάδα που είναι γηπεδούχος. Δηλαδή η τιμή που αντιστοιχεί σε κάθε συνδυασμό ομάδας και ημερομηνίας συμβολίζει το overall αποτέλεσμα που έχουνε οι fans την ίδια μέρα του αγώνα που διαδραματίζεται καθώς και 2 ημέρες πριν.

Τα 2 dataset που βλέπουμε αποτελούνται από τρεις στήλες και αναφέρονται στις ομάδες που αγωνίζονται ως γηπεδούχες και φιλοξενούμενες αντιστοίχως κατά τη διάρκεια της περιόδου που πραγματοποιούμε το πείραμα, δηλαδή από τις 9 Ιουλίου 2020 έως τις 13 Οκτωβρίου 2020 (Εικόνα 5.11).

Οι στήλες του dataset είναι οι εξής:

- Date: Η ημερομηνία που διεξήχθη ο αγώνας.
- HomeTeam: Η ομάδα που αγωνίζεται ως γηπεδούχος στον αγώνα.
- Total_Value: Αυτός ο δείκτης εκφράζει τη συναισθηματική στάση των οπαδών της ομάδας στο Twitter, όπως προκύπτει από την ανάλυση συναισθήματος που έχουμε πραγματοποιήσει στα επεξεργασμένα κείμενα (tweets). Η ερμηνεία των τιμών της στήλης έχει ως εξής:
    1. Τιμές μεγαλύτερες από το 0: Οι οπαδοί της ομάδας εκφράζονται θετικά για την ομάδα τους στο Twitter. Αυτό υποδηλώνει ότι, σύμφωνα με την ανάλυση συναισθήματος που έχουμε διεξάγει, η ομάδα και οι οπαδοί της έχουν ένα θετικό κλίμα ή "καλό vibe".
    2. Τιμές μικρότερες από το 0: Οι οπαδοί της ομάδας εκφράζονται αρνητικά για την ομάδα τους στο Twitter. Αυτό δείχνει ότι, βάσει της ανάλυσης συναισθήματος που έχουμε πραγματοποιήσει, η ομάδα και οι οπαδοί της βιώνουν ένα αρνητικό κλίμα ή "κακό vibe".

Με αυτό τον τρόπο, το dataset προσφέρει μια εικόνα για το πώς οι οπαδοί των γηπεδούχων ομάδων (και αντιστοίχα των φιλεξενούμενων ομάδων) αντιδρούν συναισθηματικά στα παιχνίδια της συγκεκριμένης περιόδου, όπως αυτή η αντίδραση αποτυπώνεται στην πλατφόρμα του Twitter.

Εικόνα 5.11 Αποτελέσματα Ανάλυσης με Βάση την Ημερομηνία του Αγώνα

```
result_df_Home_Team
```

	Date	HomeTeam	Total_Value
0	2020-07-09	Bournemouth	181
1	2020-07-09	Everton	93
2	2020-07-09	Aston Villa	0
3	2020-07-11	Norwich	0
4	2020-07-11	Watford	0
...	...	...	...
76	2020-10-04	Leicester	-44
77	2020-10-04	Southampton	45
78	2020-10-04	Arsenal	973
79	2020-10-04	Wolves	0
80	2020-10-04	Aston Villa	0

81 rows × 3 columns

```
result_df_Away_Team
```

	Date	AwayTeam	Total_Value
0	2020-07-09	Tottenham	-19
1	2020-07-09	Southampton	210
2	2020-07-09	Man United	12487
3	2020-07-11	West Ham	0
4	2020-07-11	Newcastle	0
...	...	...	...
76	2020-10-04	West Ham	0
77	2020-10-04	West Brom	0
78	2020-10-04	Sheffield United	0
79	2020-10-04	Fulham	0
80	2020-10-04	Liverpool	2052

81 rows × 3 columns

Το πρόβλημα που αντιμετωπίζουμε σε αυτή την περίπτωση είναι η έλλειψη πλήρους διαθεσιμότητας δεδομένων για όλες τις ομάδες και για όλες τις ημερομηνίες της περιόδου που καλύπτει το πείραμα. Αυτό συνεπάγεται έναν σημαντικό περιορισμό στην ανάλυσή μας: δεν μπορούμε να εξετάσουμε κάθε αγώνα με την ίδια λεπτομέρεια και πληρότητα.

Συγκεκριμένα, ο περιορισμός που προκύπτει είναι ότι η ανάλυση μας θα πρέπει να εστιάσει αποκλειστικά στους αγώνες για τους οποίους υπάρχει τουλάχιστον ένα tweet που έχει αναρτηθεί στην πλατφόρμα Twitter κατά τη διάρκεια της περιόδου του πειράματος. Αυτή η προσέγγιση διασφαλίζει ότι η συναισθηματική ανάλυση που πραγματοποιούμε βασίζεται σε πραγματικά δεδομένα και όχι σε υποθέσεις ή ελλιπή στοιχεία.

Με άλλα λόγια, για να θεωρήσουμε έναν αγώνα έγκυρο για την ανάλυση μας, πρέπει να υπάρχει τουλάχιστον μία δημοσίευση στο Twitter από οπαδό της γηπεδούχου ομάδας κατά τη διάρκεια από 9 Ιουλίου 2020 έως 13 Οκτωβρίου 2020. Αυτή η δημοσίευση θα μας επιτρέψει να εξάγουμε συμπεράσματα για τη συναισθηματική κατάσταση των οπαδών και, κατ' επέκταση, για το "vibe" της ομάδας κατά την εν λόγω χρονική περίοδο.

Αντιλαμβανόμαστε ότι αυτός ο περιορισμός μειώνει τον αριθμό των αγώνων που μπορούμε να αναλύσουμε, αλλά είναι απαραίτητος για να διατηρήσουμε την αξιοπιστία και την ακρίβεια των αποτελεσμάτων μας. Με την εξασφάλιση ότι κάθε αναλυόμενος αγώνας έχει επαρκή δεδομένα από τις δημοσιεύσεις στο Twitter, μπορούμε να παρέχουμε πιο εμπεριστατωμένες και αξιόπιστες εκτιμήσεις για τη συναισθηματική ατμόσφαιρα γύρω από τις γηπεδούχες ομάδες.

Σαν αποτελέσμα έχουμε τους 2 παρακάτω πίνακες :

Πίνακας που τα αποτελέσματα των Count_Total_Tweets δηλώνουν τον δείκτη για την ομάδα του HomeTeam

Πίνακας 5.1 Αποτελέσματα Ανάλυσης (Home)

Date	Home_Team	Count_Total_Tweets	Away_Team	FTR
9/7/2020	Bournemouth	181	Tottenham	D
9/7/2020	Everton	93	Southampton	D
11/7/2020	Liverpool	21689	Burnley	D
12/7/2020	Bournemouth	29	Leicester	H
12/7/2020	Tottenham	56	Arsenal	H

13/7/2020	Man United	14375	Southampton	D
14/7/2020	Chelsea	-4	Norwich	H
15/7/2020	Burnley	-46	Wolves	D
15/7/2020	Man City	-3327	Bournemouth	H
15/7/2020	Arsenal	354	Liverpool	H
16/7/2020	Southampton	223	Brighton	D
16/7/2020	Crystal Palace	5782	Man United	A
16/7/2020	Everton	-6	Aston Villa	D
16/7/2020	Leicester	161	Sheffield United	H
19/7/2020	Bournemouth	45	Southampton	A
19/7/2020	Tottenham	1949	Leicester	H
22/7/2020	Man United	3447	West Ham	D
22/7/2020	Liverpool	23981	Chelsea	H
26/7/2020	Man City	1488	Norwich	H
26/7/2020	Southampton	62	Sheffield United	H
26/7/2020	Leicester	1779	Man United	A
26/7/2020	Crystal Palace	571	Tottenham	D
26/7/2020	Arsenal	-1566	Watford	H
26/7/2020	Burnley	127	Brighton	A
26/7/2020	Everton	457	Bournemouth	A
26/7/2020	Chelsea	32	Wolves	H
12/9/2020	Crystal Palace	-262	Southampton	H
12/9/2020	Liverpool	1129	Leeds	H
13/9/2020	Tottenham	246	Everton	A
19/9/2020	Everton	95	West Brom	H
19/9/2020	Man United	-361	Crystal Palace	A
19/9/2020	Arsenal	653	West Ham	H
20/9/2020	Leicester	211	Burnley	H
20/9/2020	Southampton	25	Tottenham	A
20/9/2020	Chelsea	34	Liverpool	A
26/9/2020	Crystal Palace	17	Everton	A
26/9/2020	Burnley	-15	Southampton	A
27/9/2020	Tottenham	329	Newcastle	D
27/9/2020	Man City	608	Leicester	A
28/9/2020	Liverpool	3701	Arsenal	H
3/10/2020	Chelsea	282	Crystal Palace	H
3/10/2020	Everton	1375	Brighton	H
4/10/2020	Man United	-1154	Tottenham	A
4/10/2020	Leicester	-44	West Ham	A
4/10/2020	Southampton	45	West Brom	H
4/10/2020	Arsenal	973	Sheffield United	H

Πίνακας 5.2 Αποτελέσματα Ανάλυσης (Away)

Date	Away_Team	Count_Total_Tweets	Home_Team	FTR
9/7/2020	Tottenham	-19	Bournemouth	D
9/7/2020	Southampton	210	Everton	D
9/7/2020	Man United	12487	Aston Villa	A
11/7/2020	Burnley	87	Liverpool	D
11/7/2020	Chelsea	164	Sheffield United	H
11/7/2020	Man City	564	Brighton	A
12/7/2020	Leicester	-108	Bournemouth	H
12/7/2020	Arsenal	-7406	Tottenham	H
12/7/2020	Crystal Palace	-4747	Aston Villa	H
12/7/2020	Everton	746	Wolves	H
13/7/2020	Southampton	118	Man United	D
15/7/2020	Tottenham	1946	Newcastle	A
15/7/2020	Liverpool	21014	Arsenal	H
16/7/2020	Man United	6703	Crystal Palace	A
18/7/2020	Burnley	418	Norwich	A
19/7/2020	Southampton	-16	Bournemouth	A
19/7/2020	Leicester	252	Tottenham	H
20/7/2020	Everton	2114	Sheffield United	A
20/7/2020	Crystal Palace	-179	Wolves	H
21/7/2020	Man City	3959	Watford	A
21/7/2020	Arsenal	10267	Aston Villa	H
22/7/2020	Chelsea	167	Liverpool	H
26/7/2020	Man United	1564	Leicester	A
26/7/2020	Liverpool	17628	Newcastle	A
26/7/2020	Tottenham	3132	Crystal Palace	D
26/7/2020	Bournemouth	82	Everton	A
12/9/2020	Arsenal	5341	Fulham	A
12/9/2020	Southampton	16	Crystal Palace	H
13/9/2020	Everton	540	Tottenham	A
13/9/2020	Leicester	382	West Brom	A
14/9/2020	Chelsea	402	Brighton	A
19/9/2020	Crystal Palace	127	Man United	A
20/9/2020	Burnley	43	Leicester	H
20/9/2020	Tottenham	1230	Southampton	A
20/9/2020	Liverpool	975	Chelsea	A
21/9/2020	Man City	2987	Wolves	A
26/9/2020	Man United	-82	Brighton	A
26/9/2020	Everton	777	Crystal Palace	A
26/9/2020	Chelsea	231	West Brom	D

26/9/2020	Southampton	67	Burnley	A
27/9/2020	Leicester	670	Man City	A
28/9/2020	Arsenal	928	Liverpool	H
3/10/2020	Crystal Palace	52	Chelsea	H
3/10/2020	Man City	934	Leeds	D
3/10/2020	Burnley	61	Newcastle	H
4/10/2020	Tottenham	1055	Man United	A
4/10/2020	Liverpool	2052	Aston Villa	H

5.1.4 Ανάλυση Συσχέτισης Μεταξύ Υποστήριξης Οπαδών & Νικηφόρων Αγώνων

Το τελευταίο βήμα στην διαδικασία της διερεύνησης της σχέσης αποτελέσματος – οπαδών είναι η ανάλυση συσχέτισης μεταξύ υποστήριξης οπαδών και νικηφόρων αγώνων. Συγκεκριμένα, η συσχέτιση που εξετάζουμε είναι ο λόγος των αγώνων όπου η ομάδα κέρδισε και οι οπαδοί την υποστήριξαν θετικά προς το σύνολο των αγώνων όπου οι οπαδοί εξέφρασαν θετική υποστήριξη. Η ανάλυση αυτή μας επιτρέπει να κατανοήσουμε σε ποιο βαθμό η θετική υποστήριξη των οπαδών σχετίζεται με την επιτυχία της ομάδας στον αγωνιστικό χώρο.

Για να υπολογίσουμε αυτή τη συσχέτιση, ακολουθήσαμε τα εξής βήματα:

1. Καταγραφή Αγώνων με Θετική Υποστήριξη και Νίκη: Προσδιορίσαμε το πλήθος των αγώνων στους οποίους η ομάδα κέρδισε και ταυτόχρονα οι οπαδοί εξέφρασαν θετική υποστήριξη στα κοινωνικά δίκτυα.
2. Καταγραφή Αγώνων με Θετική Υποστήριξη: Προσδιορίσαμε το συνολικό πλήθος των αγώνων όπου οι οπαδοί εξέφρασαν θετική υποστήριξη, ανεξαρτήτως του αποτελέσματος του αγώνα.

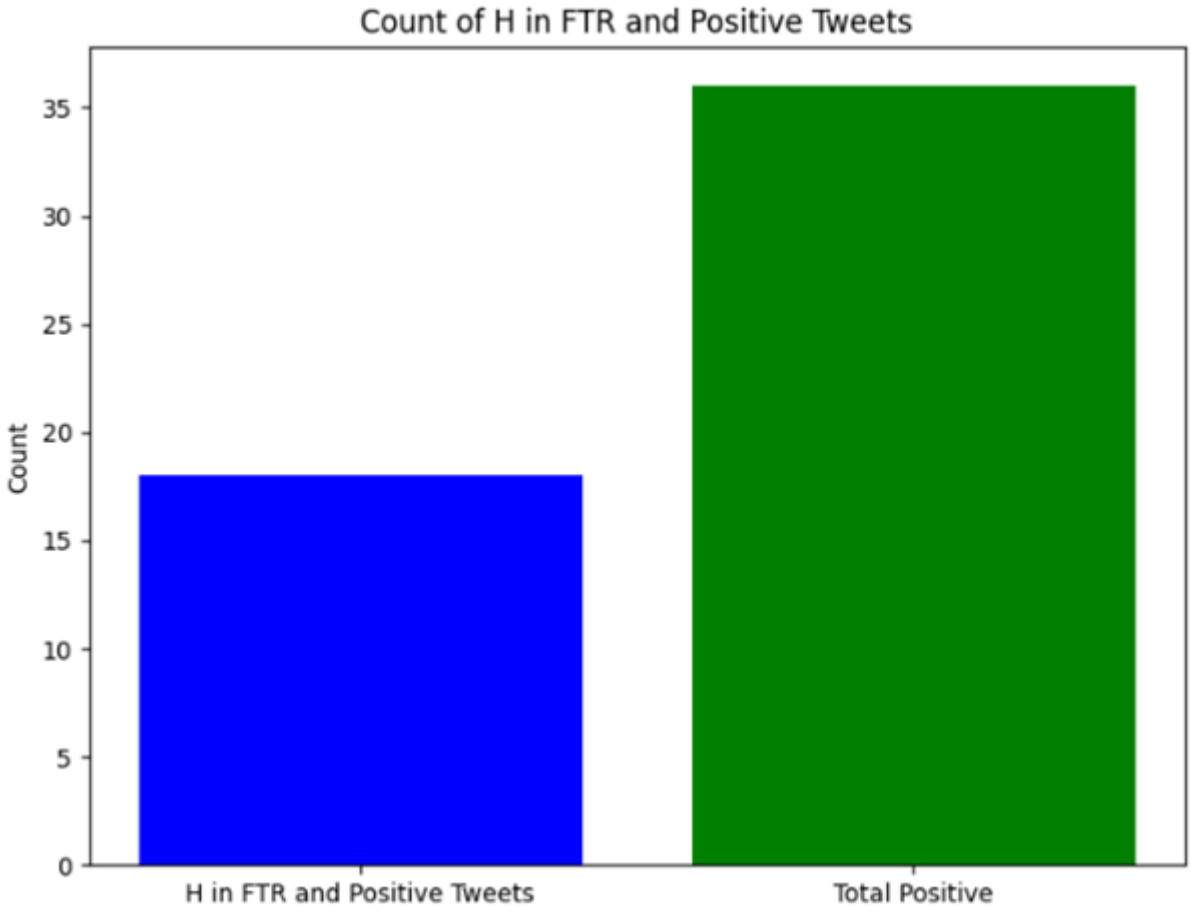
Από την ανάλυση των δεδομένων μας, προέκυψαν τα εξής:

- Συνολικός αριθμός αγώνων όπου οι οπαδοί εξέφρασαν θετική υποστήριξη: 36 αγώνες.
- Αριθμός αυτών των αγώνων στους οποίους η ομάδα κέρδισε: 18 αγώνες.

Η συσχέτιση υπολογίζεται ως ο λόγος των αγώνων με νίκη και θετική υποστήριξη προς το σύνολο των αγώνων με θετική υποστήριξη.

1. **Σχετικά με το dataset HomeTeam.**

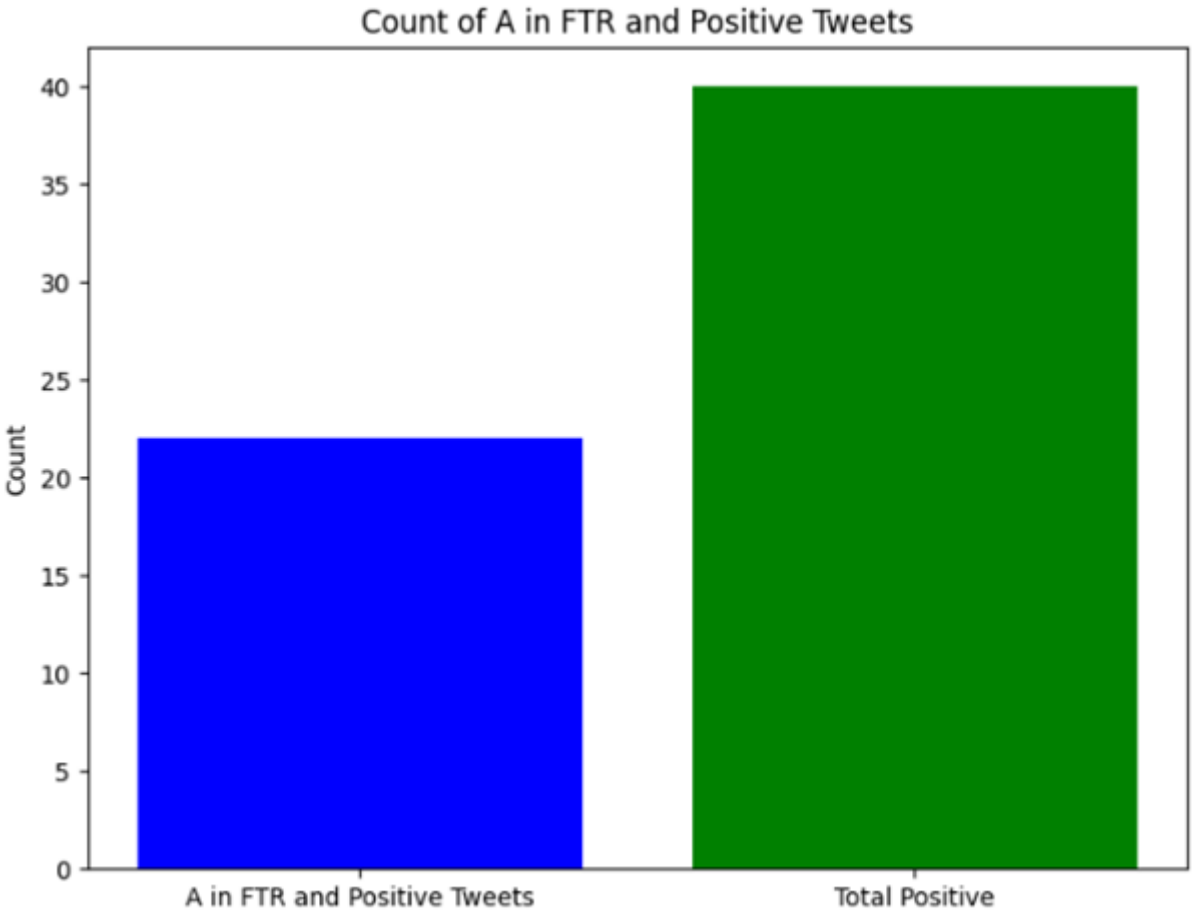
Εικόνα 5.12 Καταμέτρηση Αγώνων με Θετική Υποστήριξη & με Νίκη της Γηπεδούχας Ομάδας



(Αριθμός αγώνων με γηπεδούχα νίκη (H) και θετική υποστήριξη)/(Συνολικός αριθμός αγώνων με θετική υποστήριξη)=18/36=1/2= 50%

2. **Σχετικά με το dataset AwayTeam**

Εικόνα 5.13 Καταμέτρηση Αγώνων με Θετική Υποστήριξη & με Νίκη της Φιλοξενούμενης Ομάδας



(Αριθμός αγώνων με φιλοξενούμενη νίκη (Α) και θετική υποστήριξη)/(Συνολικός αριθμός αγώνων με θετική υποστήριξη)=22/40=11/20= 55%

Η ανάλυση των δεδομένων μας σχετικά με την έκφραση θετικής υποστήριξης από τους οπαδούς και τα αποτελέσματα των ποδοσφαιρικών αγώνων αποκαλύπτει ενδιαφέροντα ευρήματα. Συγκεκριμένα, από το σύνολο των 36 αγώνων όπου οι οπαδοί εξέφρασαν θετική υποστήριξη ανεξαρτήτως του αποτελέσματος, διαπιστώθηκε ότι η ομάδα κέρδισε σε 18 από αυτούς τους αγώνες. Αυτό αντιστοιχεί σε ποσοστό νικών 50% όταν η θετική υποστήριξη εκφράζεται υπέρ της γηπεδούχου ομάδας.

Στην περίπτωση των αγώνων όπου η θετική υποστήριξη εκφράστηκε υπέρ της φιλοξενούμενης ομάδας, παρατηρήθηκε ότι σε 22 από τους 40 συνολικούς αγώνες, η ομάδα κατάφερε να κερδίσει. Αυτό μεταφράζεται σε ποσοστό νικών 55%.

5.2 Παραδείγματα Γλώσσας

Θα εξετάσουμε τυχαία κάποια παραδείγματα από το σύνολο των δεδομένων που έχουν περάσει από τις συναρτήσεις που έχουμε αναφέρει, έτσι ώστε να μπορούμε να παρατηρήσουμε πώς αντιδρά το μοντέλο.

Επίσης, θα δούμε γενικότερες πληροφορίες που βγάζουν τα παραδείγματα που έχουν γραφτεί από τους φανς. Παρακάτω επισυνάπτονται, τα οποία είναι τυχαία επιλεγμένα από το αρχικό σύνολο των δεδομένων μας, πριν αφαιρέσουμε τα records στα οποία οι τιμές του roberta_negative ή roberta_positive δεν υπερβαίνουν το 0,3 (Εικόνα 5.14) . Με αυτόν τον τρόπο, θα μπορέσουμε να διακρίνουμε αν στο πείραμά μας έχουμε πάρει τη σωστή απόφαση να εξαιρέσουμε παραδείγματα όπου οι λέξεις δεν εκφράζουν επαρκώς συναίσθημα.

Εικόνα 5.14 Δειγματοληπτικές Τυχαίες Τιμές των Δεδομένων για να Ερευνήσουμε την Συμπεριφορά του Roberta

```
import random
random_row_index = random.randint(0, len(tweets) - 1)
random_row = tweets.iloc[random_row_index]

for column, value in random_row.items():
  print(f"{column}: {value}")
```

(i) ***Παράδειγμα Τυχαίο #1***

Id: 32816

roberta_negative: 0.09277902

roberta_neutral: 0.86268955

roberta_positive: 0.044531416

Team_Name: TottenhamHotspur

Original_Text: @THFCMill/Fans #thfc #coys https://t.co/sW2MTuTzBE

Text_after_aplying_stopwords_function: thfcmill fan thfc coy http co sw mtutzb

Σχόλιο : Το συγκεκριμένο tweet, όπως φαίνεται με την πρώτη ματιά, δεν μπορεί να προσδιορίσει ούτε θετικό ούτε αρνητικό συναίσθημα, καθώς ο χρήστης επισυνάπτει απλώς έναν σύνδεσμο χωρίς να παρέχει περαιτέρω πληροφορίες. Για τα tweets που δεν έχουμε τιμές συναισθήματος κοντά στο 0,3 και πάνω, δεν μπορούμε να πούμε ότι δηλώνουν συναίσθημα. Επιπλέον, όποιο μοντέλο και να χρησιμοποιούσαμε, είναι πολύ λογικό το σκορ neutral να είναι σχεδόν κοντά στο 1.

(ii) ***Παράδειγμα Τυχαίο #2***

Id: 8992

roberta_negative: 0.9010026

roberta_neutral: 0.09066438

roberta_positive: 0.008333075

Team_Name: Chelsea

Original_Text: Frank #Lampard is just sitting there sulking and doing fuck all... #LampardOut #CFC | #PL | #PremierLeague | #WBACHE | #Chelsea

Text_after_aplying_stopwords_function: frank lampard sit sulk fuck lampardout cfc pl premierleagu wbach Chelsea

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), δείχνει ότι οι λέξεις που διαγράφονται έχουν έναν ουδέτερο χαρακτήρα, οπότε δεν έχουν και μεγάλη βαρύτητα στο συνολικό συναίσθημα του κειμένου. Επίσης, είναι εμφανές ότι ο χρήστης εκφράζει την απογοήτευσή του για την απραγία του προπονητή, κάτι που αναγνωρίζεται από το μοντέλο RoBERTa ως αρνητικός τόνος. Αυτά τα δύο σημεία συμβάλλουν στη γενικότερη αρνητική διάθεση του tweet.

(iii) ***Παράδειγμα Τυχαίο #3***

Id: 41898

roberta_negative: 0.05743489

roberta_neutral: 0.9033388

roberta_positive: 0.039226398

Team_Name: Burnley

Original_Text: Here's our second U18 @premierleague review of the season with @dcfcofficial the early leaders in the north group. ▶https://t.co/sM1FpgRP4v #twitterclarets #BurnleyFC #UTC

Text_after_aplying_stopwords_function: second u premierleagu review season dcfcoffici earli leader north group http co sm fpgrp v twitterclaret burnleyfc utc

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), δεν περιέχει λέξεις ή εκφράσεις που να δηλώνουν κάποιο συναίσθημα του χρήστη για την ομάδα του. Ο χρήστης παραθέτει απλώς μια πηγή που παρέχει γενικές πληροφορίες για τη σεζόν. Το μοντέλο RoBERTa αναγνωρίζει πολύ καλά ότι το περιεχόμενο έχει έναν πιο ουδέτερο τόνο.

(iv) ***Παράδειγμα Τυχαίο #4***

Id: 22112

roberta_negative: 0.122282445

roberta_neutral: 0.81765044

roberta_positive: 0.060067087

Team_Name: Chelsea

Original_Text: Watching #Enslaved by Samuel L Jackson on @BBCiPlayer and the ending of the episode brings a crew to #Galibi #Suriham to a #Maroon village and the main chief /captain who greets them is wearing a #Chelsea football top. We are spirited freedom fighters. Don't mess with us. #cfc

Text_after_aplying_stopwords_function: watch enslav samuel l jackson bbciplay end episod bring crew galibi suriham maroon villag main chief captain greet wear chelsea footbal top spirit freedom fighter mess us cfc

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), προφανώς δεν είχε ως στόχο το αγγλικό πρωτάθλημα ή κάποια συγκεκριμένη ομάδα. Είναι εμφανές ότι ο χρήστης μιλάει για μια τηλεοπτική σειρά, στην οποία αναφέρεται η ομάδα για κάποιον λόγο. Καταλαβαίνουμε ότι μπορεί να υπάρχουν παρόμοια tweets, τα οποία δεν θα ήταν σωστό να συμπεριλάβουμε στην ανάλυσή μας. Για αυτόν τον λόγο, θέσαμε έναν περιορισμό στις τιμές του συναισθήματος (είτε θετικό είτε αρνητικό) πάνω από 0,3, ώστε να αποφύγουμε περιπτώσεις όπου τα tweets μπορεί να περιέχουν κατά λάθος hashtags ομάδων χωρίς να εκφράζουν πραγματικό συναίσθημα.

(v) ***Παράδειγμα Τυχαίο #5***

Id: 12199

roberta_negative: 0.6207983

roberta_neutral: 0.3380991

roberta_positive: 0.041102655

Team_Name: ManchesterCity

Original_Text: Ooh crap ... Kevin has come off ... Not sure why but could have a problem #MCFC #Worried

Text_after_aplying_stopwords_function: ooh crap kevin come sure could problem mcfc worri

Σχόλιο : Η εκδήλωση του χρήστη, επειδή κάποιος ποδοσφαιριστής χρειάζεται να βγει από το παιχνίδι, παραθέτει το σχόλιο ότι η απουσία του μπορεί να προκαλέσει πρόβλημα. Το αρχικό σχόλιο "oh crap" εκφράζει μια έντονη αρνητική αύρα, καθώς και η λέξη "worried" υποδηλώνει ανησυχία. Όπως ακριβώς ένας άνθρωπος μπορεί να χαρακτηρίσει αυτό το tweet

ως ανησυχητικό, το μοντέλο RoBERTa επίσης αναγνωρίζει ότι υπάρχει σημαντική αρνητική χροιά, και επομένως η τιμή του negative αναμένεται να είναι αρκετά υψηλή.

(vi) ***Παράδειγμα Τυχαίο #6***

Id: 38

roberta_negative: 0.06924998

roberta_neutral: 0.88011837

roberta_positive: 0.050631665

Team_Name: ManchesterUnited

Original_Text: What does Sandra Bullock have to do with Manchester United's EBITDA limit being set at £65m by banks? Not much, aside from a tortured analogy. So of course the smart guys in graphics made it into this quote card. #MUFC @TheAthleticUK More 👇 https://t.co/UYdApJipZB https://t.co/ANwC0snHzT

Text_after_aplying_stopwords_function: sandra bullock manchest unit ebitda limit set bank much asid tortur analog cours smart guy graphic made quot card mufc theathleticuk http co uydapjipzb http co anwc snhzt

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), δεν περιέχει λέξεις ή εκφράσεις που να δηλώνουν σαφές συναίσθημα. Ο χρήστης αναφέρει μια αναλογία που συνδέει την Sandra Bullock με τον περιορισμό του EBITDA της Manchester United, καθιστώντας το tweet περισσότερο πληροφοριακό παρά συναισθηματικό. Ο σαρκαστικός τόνος και η αναφορά στους "έξυπνους τύπους" στα γραφικά ενισχύουν αυτόν τον πληροφοριακό χαρακτήρα. Το μοντέλο RoBERTa αναγνωρίζει ότι το περιεχόμενο έχει έναν πιο ουδέτερο τόνο, καθώς το συναίσθημα δεν είναι έντονα θετικό ή αρνητικό.

(vii) ***Παράδειγμα Τυχαίο #7***

Id: 28788

roberta_negative: 0.04900994

roberta_neutral: 0.59917486

roberta_positive: 0.35181522

Team_Name: ManchesterUnited

Original_Text: @nalikajanari07 @ManUtd Yes I forgot you are right...thank you #ManchesterUnited #ManUnited

Text_after_aplying_stopwords_function: nalikajanari manutd ye forgot right thank manchesterunit manunit

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), πιθανότατα αποτελεί απάντηση σε κάποιο άλλο tweet που αφορά την ομάδα, καθώς οι λέξεις δείχνουν έναν θετικό χαρακτήρα. Το μοντέλο RoBERTa αναγνωρίζει κυρίως έναν ουδέτερο χαρακτήρα, αλλά επίσης δίνει ένα αισθητό σκορ ως θετικό λόγω των λέξεων "right" και "thank", που εκφράζουν συμφωνία και ευγνωμοσύνη.

(viii) ***Παράδειγμα Τυχαίο #8***

Id: 15355

roberta_negative: 0.011206162

roberta_neutral: 0.6955809

roberta_positive: 0.29321295

Team_Name: Arsenal

Original_Text: New: New signings give Arsenal U23s a major boost https://t.co/xFl2Nbgw5K #arsenal #afc https://t.co/RZGbpZ1Ley

Text_after_aplying_stopwords_function: new new sign give arsen u major boost http co xfl nbgw k arsen afc http co rzgbpz ley

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), αναφέρεται στην ομάδα Arsenal U23 και στις νέες μεταγραφές που τους δίνουν σημαντική ενίσχυση. Παρόλα αυτά, δεν υπάρχει κάποιο σχόλιο που να τοποθετεί το tweet υπό τη θετική ή την αρνητική ομπρέλα. Το περιεχόμενο παραμένει ουδέτερο, εστιάζοντας απλώς στην πληροφόρηση.

(ix) ***Παράδειγμα Τυχαίο #9***

Id: 15943

roberta_negative: 0.38172874

roberta_neutral: 0.46721578

roberta_positive: 0.15105553

Team_Name: Burnley

Original_Text: Swear to god tarkys toe better fucking look like this right now 🖐 #twitterclarets #BurnleyFC @TurfCastPodcast #garlickout #backdyche https://t.co/nxlYmv3sfA

Text_after_aplying_stopwords_function: swear god tarki toe better fuck look like right twitterclaret burnleyfc turfcastpodcast garlickout backdych http co nxlymv sfa

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), εκφράζει έναν έντονα αρνητικό τόνο. Η χρήση των λέξεων "swear", "god", "better", "fucking", και "look like" υποδηλώνουν έντονα συναισθήματα και ανησυχία για την κατάσταση του ποδιού του Tarky. Το μοντέλο RoBERTa αναγνωρίζει τον αρνητικό τόνο του tweet, καθώς οι λέξεις εκφράζουν ένταση και δυσαρέσκεια.

(x) ***Παράδειγμα Τυχαίο #10***

Id: 40736

roberta_negative: 0.12788689

roberta_neutral: 0.8235467

roberta_positive: 0.0485664

Team_Name: ManchesterCity

Original_Text: The @MumbaiCityFC boss! Read: https://t.co/IjQ4UV1KTr #MCFC #HeroISL #IndianFootball https://t.co/vmLS5GlbKr

Text_after_aplying_stopwords_function: mumbaicityfc boss read http co ijq uv ktr mcfc heroisl indianfootbal http co vml glbkr

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), δεν περιέχει λέξεις ή εκφράσεις που να δηλώνουν σαφές συναίσθημα. Αναφέρεται στον προπονητή της Mumbai City FC και παραθέτει έναν σύνδεσμο για διάβασμα, συνοδευόμενο από σχετικά hashtags. Το μοντέλο RoBERTa αναγνωρίζει ότι το περιεχόμενο είναι πληροφοριακό και διατηρεί έναν ουδέτερο τόνο, χωρίς να εκφράζει συγκεκριμένα θετικά ή αρνητικά συναισθήματα.

(xi) ***Παράδειγμα Τυχαίο #11***

Id: 23599

roberta_negative: 0.05529677

roberta_neutral: 0.8981958

roberta_positive: 0.04650742

Team_Name: ManchesterUnited

Original_Text: As bankers running the club, you'd think it would be their job to know all financial aspect of a deal they were to have apparently worked on for over a year before THEY called it off. I won't believe anything coming out of the club as they try and spin the PR on this deal. #MUFC https://t.co/JzDHhLScJO

Text_after_aplying_stopwords_function: banker run club think would job know financi aspect deal appar work year call believ anyth come club tri spin pr deal mufc http co jzdhhlscjo

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), εκφράζει έντονη δυσαρέσκεια και δυσπιστία προς τους τραπεζίτες που διαχειρίζονται τον σύλλογο. Ο χρήστης αναφέρεται στην υποχρέωση των τραπεζιτών να γνωρίζουν όλες τις οικονομικές πτυχές μιας συμφωνίας που εργάζονταν για πάνω από ένα χρόνο, πριν τελικά την ακυρώσουν οι ίδιοι. Η χρήση λέξεων όπως "won't believe anything" καθώς ενισχύει τον περισσοτερο τον αρνητικό τόνο παρά τον θετικό και την καχυποψία του χρήστη προς τον σύλλογο. Το μοντέλο RoBERTa αναγνωρίζει τον αρνητικό χαρακτήρα του tweet, καθώς το σύνολο των λέξεων δεν εκφράζουν έντονη απογοήτευση ή έλλειψη εμπιστοσύνης έτσι ώστε να παραθέσει ένα μεγαλύτερο νούμερο. Συνεπώς το ουδετερό value είναι αρκετά μεγάλο και τέτοια είδους tweets δεν τα συμπεριλαμβάνουμε στην ανάλυση μας.

(xii) ***Παράδειγμα Τυχαίο #12***

Id: 4559

roberta_negative: 0.05531987

roberta_neutral: 0.8898093

roberta_positive: 0.05487083

Team_Name: Arsenal

Original_Text: #OTD 1996, Arsène Wenger's first match, a 0-2 victory at Blackburn. Up to this moment he had been seeing out his contract in Japan. Ian Wright scored the first Arsene Wenger goal for #Arsenal #Throwback #PremierLeague #MerciArsene https://t.co/1QQFd1voo3

Text_after_aplying_stopwords_function: otd ar ne wenger first match victori blackburn moment see contract japan ian wright score first arsen wenger goal arsen throwback premierleagu merciarsen http co qqfd voo

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), αναφέρεται σε ένα ιστορικό γεγονός για τον Arsène Wenger και την ομάδα Arsenal. Ο χρήστης μιλάει για τον πρώτο αγώνα του Wenger ως προπονητή, ο οποίος

κατέληξε σε νίκη 0-2 εναντίον της Blackburn. Αναφέρεται επίσης στην πορεία του Wenger πριν από αυτόν τον αγώνα, ενώ βρισκόταν στην Ιαπωνία. Το tweet περιλαμβάνει λέξεις και φράσεις όπως "victory", "Ian Wright scored" και "Merci Arsène", που εκφράζουν θετικά συναισθήματα. Ωστόσο, το μοντέλο RoBERTa μπορεί να αναγνωρίσει κυρίως έναν ουδέτερο, δεδομένου ότι το περιεχόμενο είναι περισσότερο πληροφοριακό.

(xiii) ***Παράδειγμα Τυχαίο #13***

Id: 17398

roberta_negative: 0.027907928

roberta_neutral: 0.6213801

roberta_positive: 0.35071194

Team_Name: ManchesterUnited

Original_Text: Goodbye for now 😑#MUFC https://t.co/6l3FSYGKvq

Text_after_aplying_stopwords_function: goodby mufc http co l fsygkvq

Σχόλιο : Στο συγκεκριμένο παράδειγμα, όταν περνάμε το κείμενο από τη συνάρτηση αφαίρεσης λέξεων-κλειδιών (stopwords), δεν επιτυγχάνεται το επιθυμητό αποτέλεσμα, καθώς η λέξη "goodbye" αλλάζει σε "goodby". Αυτή η αλλοίωση επηρεάζει την ακρίβεια της ανάλυσης συναισθήματος, με αποτέλεσμα το μοντέλο RoBERTa να κατανοεί το περιεχόμενο ως θετικό. Το γεγονός αυτό αντικατοπτρίζεται στην υψηλή θετική τιμή (πάνω από 0,3) που αποδίδει το μοντέλο, οδηγώντας το tweet να θεωρηθεί ως θετικό και να συμπεριληφθεί στην ανάλυση μας. Αυτό το παράδειγμα αναδεικνύει μια περίπτωση όπου το συνολικό μας σύστημα δεν λειτουργεί σωστά, καθώς η αφαίρεση λέξεων-κλειδιών μπορεί να αλλοιώσει το νόημα και να επηρεάσει την αξιοπιστία των αποτελεσμάτων.

(xiv) ***Παράδειγμα Τυχαίο #14***

Id: 17354

roberta_negative: 0.0040853913

roberta_neutral: 0.11529823

roberta_positive: 0.88061637

Team_Name: ManchesterUnited

Original_Text: Excited to see some training pics tomorrow you know. Still made transfers and we'll be stronger for it. Let's get behind the lads #MUFC

Text_after_aplying_stopwords_function: excit see train pic tomorrow know still made transfer stronger let get behind lad mufc

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), εκφράζει ενθουσιασμό και αισιοδοξία για την ομάδα Manchester United. Η χρήση λέξεων όπως "excited", "stronger", και "get behind the lads" δείχνει θετικά συναισθήματα και υποστήριξη προς την ομάδα. Το μοντέλο RoBERTa αναγνωρίζει κυρίως έναν θετικό τόνο, με την τιμή του θετικού συναισθήματος να είναι πολύ υψηλή ενώ οι ουδέτερες και αρνητικές τιμές είναι πολύ μικρές. Αυτό αντικατοπτρίζει τη συνολική θετική διάθεση του χρήστη προς την ομάδα και τις μελλοντικές προοπτικές της.

(xv) ***Παράδειγμα Τυχαίο #15***

Id: 30454

roberta_negative: 0.8392109

roberta_neutral: 0.14477018

roberta_positive: 0.01601892

Team_Name: ManchesterUnited

Original_Text: All those people who think OGS isn't qualified enough to be the manager of Man United y'all are some sort of a highly retarded disgraceful bastards. Get Well Soon Clowns. #MUFC #OleIn

Text_after_aplying_stopwords_function: peopl think og qualifi enough manag man unit sort highli retard disgrac bastard get well soon clown mufc olein

Σχόλιο : Το συγκεκριμένο tweet, μετά την εφαρμογή της συνάρτησης αφαίρεσης λέξεων-κλειδιών (stopwords), εκφράζει έντονα αρνητικά συναισθήματα προς τους επικριτές του προπονητή της Manchester United, OGS (Ole Gunnar Solskjær). Η χρήση λέξεων όπως "retard", "disgrac bastard", και "clown" υποδηλώνει προσβολή και έντονη δυσαρέσκεια. Το μοντέλο RoBERTa αναγνωρίζει τον αρνητικό χαρακτήρα του tweet, με την αρνητική τιμή να είναι πολύ υψηλή ενώ οι ουδέτερες και θετικές τιμές είναι πολύ χαμηλές Αυτό αντικατοπτρίζει τη σαφή και έντονη αρνητική διάθεση του χρήστη απέναντι σε αυτούς που αμφισβητούν τα προσόντα του προπονητή.

1. **Σύνοψη:**

Τα παραπάνω παραδείγματα παρουσιάζουν μια ποικιλία σχολίων χρηστών στα social media, τα οποία αναλύονται με τη χρήση της τεχνικής αφαίρεσης λέξεων-κλειδιών (stop words) και την αξιολόγηση από το μοντέλο RoBERTa.

Η παράλειψη λέξεων-κλειδιών μπορεί να έχει αρνητικές συνέπειες: Σε ορισμένες περιπτώσεις, όπως η αλλαγή της λέξης "goodbye" σε "goodby", μπορεί να παραμορφωθεί το νόημα του κειμένου και να πληγεί η ακρίβεια της ανάλυσης συναισθήματος. Το RoBERTa μοντέλο μπορεί επιτυχώς να ανιχνεύσει το συνολικό συναισθηματικό ύφος των σχολίων, είτε είναι θετικό (π.χ., χαρά για μια ομάδα), αρνητικό (π.χ., απογοήτευση για μια τράπεζα) ή ουδέτερο (π.χ., πληροφοριακά σχόλια για μια ιστορική γεγονότα).Η ανάλυση του συναισθήματος είναι πολύπλοκη λόγω της γλώσσας και των πολλών παραμέτρων που επηρεάζουν το νόημα των λέξεων.

Συνολικά, τα παραδείγματα αποδεικνύουν τη σημασία της ανάλυσης συναισθημάτων στην κατανόηση των κοινωνικών μέσων και τις προκλήσεις που προκύπτουν σε αυτήν την διαδικασία. Η αποτελεσματική χρήση της μεθόδου αποτελεί ζήτημα επιλογής κατάλληλων εργαλείων και αντίληψης των περιορισμών τους.

6. Σύνοψη και Συμπεράσματα

Η παρούσα μελέτη αφορά την Ανάλυση Συναισθήματος (Sentiment Analysis) στις αναρτήσεις χρηστών σε μέσα κοινωνικής δικτύωσης (Social Media) στο πεδίο του Ποδοσφαίρου (Football, USA: Soccer). Πρόκειται για αναρτήσεις οπαδών και ποδοσφαιρόφιλων στην πλατφόρμα κοινωνικής δικύωσης «X» - πρώην «Twitter «, με σύνολων δεδομένων 1.000.562 από το Πρωτάθλημα English Premier League («Αγγλικό Ποδόσφαιρο»). Τα δεδομένα προήλθαν από την πλατφόρμα του Kaggle, συγκεκριμένα από το σύνολο δεδομένων "EPL Teams Twitter Sentiment Dataset" που μπορείτε να βρείτε εδώ: https://www.kaggle.com/datasets/wjia26/epl-teams-twitter-sentiment-dataset/data. Στόχο της παρούσας ανάλυσης και μελέτης αποτελεί, κατ' αρχήν, η έρευνα συσχέτισης αποτελεσμάτων ομάδων με συναισθηματικά δεδομένα οπαδών από το Twitter. Ταυτοχρόνως, στοχεύεται η δημιουργία και αξιολόγηση μοντέλων μηχανικής μάθησης για πρόβλεψη αποτελεσμάτων αγώνων με υψηλή ακρίβεια. και η ανάπτυξη δεικτών για ενσωμάτωση στους αλγόριθμους μηχανικής μάθησης.

Η διαδικασία μπορεί να συνοψιστεί σε 14 βασικά στάδια, όπως θα αναφερθεί παρακάτω. Κάθε στάδιο συμβάλλει στην συνολική ανάλυση, αρχίζοντας από την επιλογή των δεδομένων και ολοκληρώνοντας με την τελική παρουσίαση των αποτελεσμάτων, επιδιώκοντας μια ολοκληρωμένη και συστηματική εξερεύνηση του θέματος.

Βήμα 1. Εύρεση Δεδομένων: Επιλογή κατάλληλου συνόλου δεδομένων για την ανάλυση.

Βήμα 2. Μοντέλα Πρόβλεψης: Δημιουργία και αξιολόγηση μοντέλων μηχανικής μάθησης για πρόβλεψη αποτελεσμάτων αγώνων με υψηλή ακρίβεια.

Βήμα 3. Προγνωστικά και Δείκτες: Ανάπτυξη δεικτών για ενσωμάτωση στους αλγόριθμους μηχανικής μάθησης.

Βήμα 4. Σχέση Οπαδών-Αποτελεσμάτων: Έρευνα συσχέτισης αποτελεσμάτων ομάδων με συναισθηματικά δεδομένα οπαδών από το Twitter.

Βήμα 5. Ανάλυση Δεδομένων: Εξερεύνηση και αξιοποίηση διαθέσιμων δεδομένων.

Βήμα 6. Προ επεξεργασία Tweets: Απλοποίηση tweets χωρίς αλλοίωση περιεχομένου για ανάλυση συναισθήματος.

Βήμα 7. Επιλογή Μοντέλου: Χρήση του RoBERTa, εκπαιδευμένου σε δεδομένα Twitter.

Βήμα 8. Χρονικό Εύρος: Περιορισμός ανάλυσης για ακρίβεια αποτελεσμάτων.

Βήμα 9. Αξιοποίηση Retweets: Χρήση retweets για εκτίμηση συμφωνίας με το συναίσθημα.

Βήμα 10. Περιορισμοί Δεδομένων: Ελλιπής διαθεσιμότητα δεδομένων για όλες τις ομάδες/ημερομηνίες.

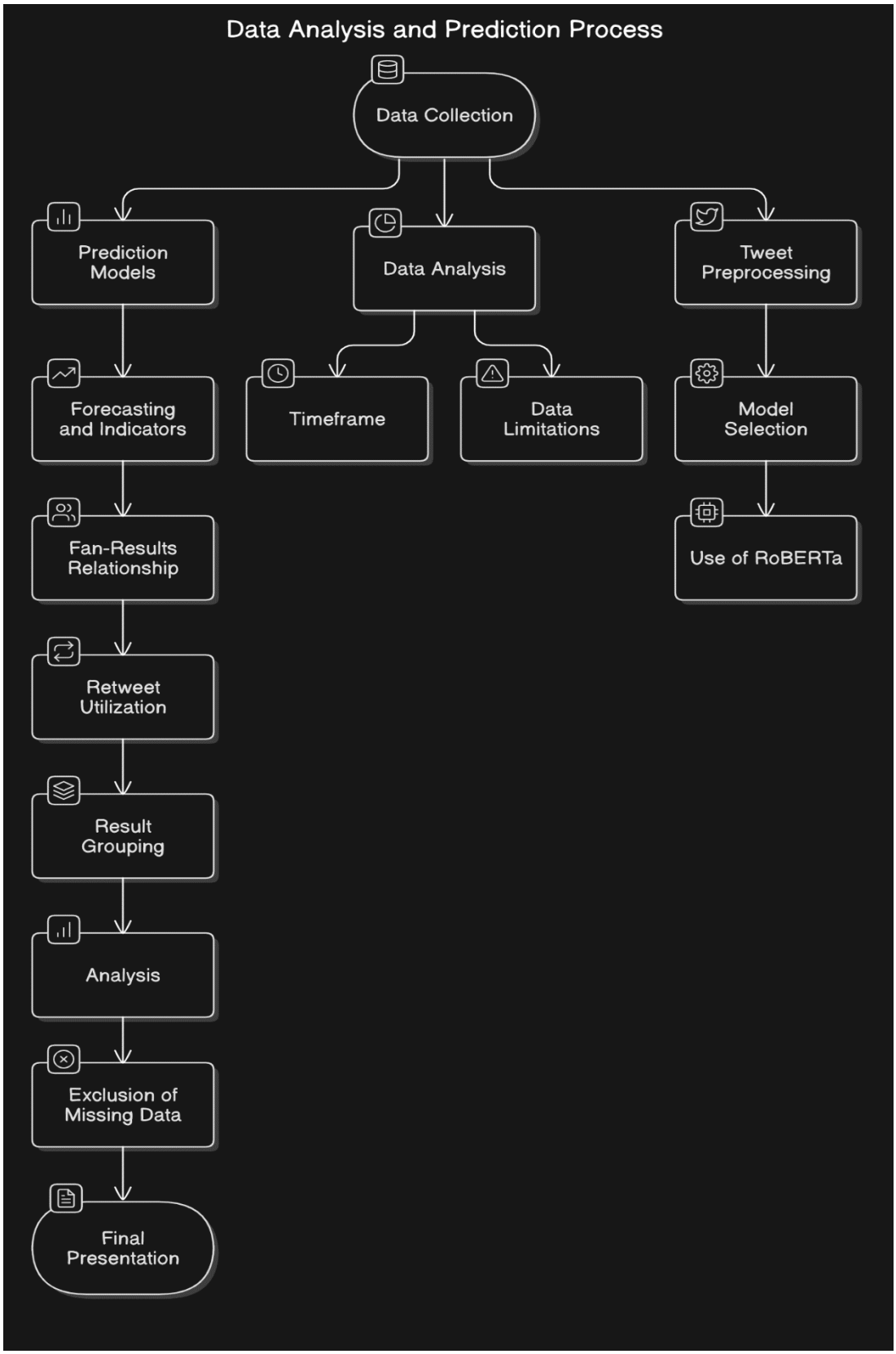
Βήμα 11. Ομαδοποίηση Αποτελεσμάτων: Ανά ομάδα και ημερομηνία για υπολογισμό θετικών/αρνητικών vibes.

Βήμα 12. Ανάλυση Εντός/Εκτός Έδρας: Ανά ομάδα για περιορισμένο σύνολο δεδομένων.

Βήμα 13. Αποκλεισμός Ελλιπών Δεδομένων: Αφαίρεση ομάδων χωρίς διαθέσιμα tweets.

Βήμα 14. Τελική Παρουσίαση: Δημιουργία πινάκων με θετικά/αρνητικά vibes ανά αγώνα, συσχέτιση αποτελεσμάτων με υποστήριξη οπαδών.

Εικόνα 6.1 Διάγραμμα Μεθόδολογίας



1. Στοιχεία που ενισχύουν τη σύνδεση

Η ανάλυση δεδομένων που παρουσιάστηκε φέρνει στο φως μια ενδιαφέρουσα σύνδεση μεταξύ της θετικής υποστήριξης από τους οπαδούς και της αγωνιστικής επίδοσης των ομάδων. Η μελέτη υποδεικνύει ότι η θετική ατμόσφαιρα στο γήπεδο μπορεί να λειτουργήσει ως καταλύτης για την επίτευξη νίκης, ενισχύοντας το ηθικό των παικτών και τροφοδοτώντας την ορμή τους.

Τα στοιχεία που ενισχύουν τη σύνδεση μεταξύ οπαδών, ομάδας και αθλητικών επιδόσεων μέσω των social media περιλαμβάνουν: τα ποσοστά νίκης (1.1), τη συναισθηματική επίδραση (1.2) και την ομαδική δυναμική (1.3).

(1.1) Ποσοστά νίκης

Τα ποσοστά νίκης αποτελούν έναν κρίσιμο παράγοντα που συνδέεται με την υποστήριξη των οπαδών στα social media. Η μελέτη δείχνει ότι οι ομάδες που λαμβάνουν έντονη και θετική υποστήριξη από τους οπαδούς τους έχουν αυξημένα ποσοστά επιτυχίας. Συγκεκριμένα, παρατηρήθηκαν ποσοστά νίκης που φτάνουν το 50% για τους γηπεδούχους και το 55% για τους φιλοξενούμενους σε αγώνες όπου η υποστήριξη αυτή ήταν έντονη. Τα δεδομένα υποδηλώνουν ότι η διαρκής αλληλεπίδραση με τους οπαδούς μπορεί να προσφέρει στους παίκτες ένα ψυχολογικό πλεονέκτημα που μεταφράζεται σε καλύτερες επιδόσεις στο γήπεδο.

(1.2) Συναισθηματική επίδραση

Όσον αφορά τη συναισθηματική επίδραση, η θετική ενέργεια που μεταφέρεται μέσω των μέσων κοινωνικής δικτύωσης από ένα ενθουσιώδες πλήθος οπαδών έχει σημαντική επιρροή. Αυτή η ενέργεια μπορεί να λειτουργήσει ψυχολογικά ενισχυτικά, τόσο στους οπαδούς που παρακολουθούν όσο και στους ίδιους τους παίκτες. Παρόμοια με τη δυναμική ενός ζωντανού πλήθους σε ένα αθλητικό γεγονός, τα μηνύματα υποστήριξης και ο ενθουσιασμός στα social media δημιουργούν ένα περιβάλλον που ενισχύει την αυτοπεποίθηση και την ψυχική ανθεκτικότητα των αθλητών.

(1.3) Ομαδική δυναμική

Σε σχέση με την ομαδική δυναμική, αξίζει να σημειωθεί ότι η έντονη και θετική υποστήριξη από τους οπαδούς συμβάλλει σημαντικά στη σύσφιξη των σχέσεων μεταξύ των παικτών. Η ενέργεια αυτή λειτουργεί ως καταλύτης για την ενίσχυση του ομαδικού πνεύματος και της

αλληλεγγύης μέσα στην ομάδα. Οι παίκτες, νιώθοντας ότι αποτελούν μέρος μιας μεγαλύτερης κοινότητας που τους στηρίζει, τείνουν να συνεργάζονται πιο αποτελεσματικά, να καταβάλλουν μεγαλύτερη προσπάθεια και να επιδιώκουν με περισσότερη αποφασιστικότητα το κοινό όφελος της ομάδας.

2. Εφαρμογές και οφέλη

Η παρούσα προσέγγιση μπορεί να θεωρηθεί ότι συνδέεται με τις εξής εφαρμογές και οφέλη: βελτίωση στρατηγικής εμπλοκής οπαδών (2.1), ενίσχυση ψυχολογίας παικτών (2.2) και δυναμωμένη σχέση με τους υποστηρικτές (2.3).

(2.1) Βελτίωση στρατηγικής εμπλοκής οπαδών

Ειδικότερα, για τη βελτίωση στρατηγικής εμπλοκής οπαδών (2.1), έχει παρατηρηθεί ότι η χρήση ανάλυσης συναισθημάτων μπορεί να καθοδηγήσει στοχευμένες δράσεις που ενθαρρύνουν τη θετική συμμετοχή. Τέτοιες δράσεις περιλαμβάνουν τη διοργάνωση διαδραστικών εκδηλώσεων, διαγωνισμών και προωθητικών ενεργειών που διαμορφώνουν ένα πιο ελκυστικό περιβάλλον για τους οπαδούς, ενισχύοντας την αλληλεπίδραση και τη σύνδεση με την ομάδα.

(2.2) Ενίσχυση ψυχολογίας παικτών

Επιπλέον, για την ενίσχυση ψυχολογίας παικτών (2.2), θεωρείται ότι η κατανόηση του συναισθηματικού αντίκτυπου της υποστήριξης των οπαδών επιτρέπει στις ομάδες να υιοθετήσουν αποτελεσματικές στρατηγικές ψυχολογικής ενίσχυσης. Αυτές μπορεί να περιλαμβάνουν εμπνευσμένες ομιλίες, προβολές βίντεο που τονίζουν την αξία της ομαδικότητας, καθώς και πρακτικές που ενισχύουν τη συνοχή και το ηθικό της ομάδας, βοηθώντας τους παίκτες να αποδώσουν στο μέγιστο των δυνατοτήτων τους.

(2.3) Δυναμωμένη σχέση με τους υποστηρικτές

Σε ό,τι αφορά τη δυναμωμένη σχέση με τους υποστηρικτές (2.3), αξίζει να αναφερθεί ότι η καλλιέργεια θετικών συναισθημάτων τόσο εντός όσο και εκτός γηπέδου συμβάλλει σημαντικά στην ενίσχυση της αφοσίωσης και της ταύτισης των οπαδών με την ομάδα. Αυτό έχει ως αποτέλεσμα τη δημιουργία μιας μακροχρόνιας και αμοιβαία ωφέλιμης σχέσης, που ενισχύει τη συνολική υποστήριξη και την αίσθηση κοινότητας γύρω από την ομάδα.

3. Συμπέρασμα

Η ανάλυση δεδομένων και συναισθημάτων έχει αποδείξει την καθοριστική σημασία της θετικής υποστήριξης των οπαδών στην επίτευξη αγωνιστικής επιτυχίας. Οι οπαδοί, μέσω της εμπλοκής τους τόσο στο γήπεδο όσο και στα μέσα κοινωνικής δικτύωσης, λειτουργούν ως πηγή ενέργειας και ενίσχυσης για τους αθλητές και τις ομάδες. Η αξιοποίηση της ανάλυσης συναισθημάτων και της παρακολούθησης των αντιδράσεων του κοινού επιτρέπει τη δημιουργία στοχευμένων στρατηγικών που βελτιώνουν τις αθλητικές επιδόσεις και εμβαθύνουν τη σχέση με τους υποστηρικτές.

Η έρευνα αποκαλύπτει ισχυρή συσχέτιση ανάμεσα στην ενεργή υποστήριξη των οπαδών και τη βελτίωση των αποτελεσμάτων. Η υποστήριξη αυτή εκτείνεται πέρα από τις εμφανείς εκφράσεις, όπως το χειροκρότημα στο γήπεδο ή τα σχόλια στα κοινωνικά δίκτυα, και συμβάλλει στη δημιουργία μιας διαρκούς σχέσης εμπιστοσύνης και αφοσίωσης.

Επιπλέον, η κατανόηση των συναισθημάτων και των προτιμήσεων του κοινού μέσω της ανάλυσης δεδομένων ανοίγει νέους δρόμους για καινοτόμες στρατηγικές μάρκετινγκ και αλληλεπίδρασης. Η ανάλυση των hashtags, των σχολίων και των συναισθημάτων επιτρέπει την προσαρμογή των εμπειριών στις ανάγκες του κοινού, ενισχύοντας τη σύνδεση με τους υποστηρικτές και την απόδοσή τους στο αγωνιστικό πεδίο.

Συνοψίζοντας, τα κύρια σημεία που προέκυψαν από την παρούσα ανάλυση είναι τα εξής:

- Η θετική υποστήριξη των οπαδών επηρεάζει άμεσα την αγωνιστική επιτυχία. Οι ομάδες με ενεργούς και αφοσιωμένους υποστηρικτές αποκτούν σημαντικό συγκριτικό πλεονέκτημα.
- Η ανάλυση δεδομένων οπαδών επιτρέπει τη δημιουργία στοχευμένων στρατηγικών. Αυτές οι στρατηγικές ενισχύουν τη σύνδεση με τους οπαδούς και την ψυχολογία των αθλητών.
- Η συσχέτιση μεταξύ υποστήριξης και αθλητικής επιτυχίας είναι ισχυρή. Μεγαλύτερη έρευνα μπορεί να αποκαλύψει πρακτικές που αξιοποιούν τη δύναμη των οπαδών για την επίτευξη βέλτιστων αποτελεσμάτων.

Οι οργανισμοί και οι αθλητές μπορούν να επωφεληθούν σημαντικά από την επένδυση στην ανάλυση δεδομένων και τη διαχείριση συναισθημάτων των οπαδών. Με σωστή διαχείριση, η θετική υποστήριξη μπορεί να μετατραπεί σε στρατηγικό πλεονέκτημα, τόσο σε

αγωνιστικό όσο και σε επιχειρησιακό επίπεδο, αναβαθμίζοντας την αθλητική εμπειρία στο σύνολό της.

Βιβλιογραφία

Ακολουθούν οι βιβλιογραφικές αναφορές (πηγές) της Εργασίας.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11), e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Aydin, Z. E., & Ozturk, Z. K. (2021). Performance Analysis of XGBoost Classifier with Missing Data. ResearchGate. https://www.researchgate.net/publication/350135431_Performance_Analysis_of_XGBoost_Classifier_with_Missing_Data

A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," Proceedings of the 6th International Conference on Information Technology and Multimedia, Putrajaya, Malaysia, 2014, pp. 212-216, doi: 10.1109/ICIMU.2014.7066632.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Pacak, Anna 2020/06/01 Sports In The Time Of Coronavirus Crisis: Social Media Response Strategies Of Professional English Football Clubs 10.13140/RG.2.2.10674.84169

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

Cheong, France Cheong, Christopher 2011/01/01 Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods15th Pacific Asia Conference on Information Systems (PACIS)

Deniz, A., Angin, M., & Angin, P. (2022). Understanding IMF Decision-Making with Sentiment Analysis. 2022 30th Signal Processing and Communications Applications Conference (SIU). https://doi.org/10.1109/siu55565.2022.9864926

Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

Giachanou, A., & Crestani, F. (2016). Like it or not. ACM Computing Surveys, 49(2), 1–41. https://doi.org/10.1145/2938640

Himelboim, I., Smith, MA, Rainie, L, Shneiderman, B, & Espina, C. (2017). Classifying Twitter topic-networks using social network analysis. Social Media+Society, 3(1), 1-13

IBM Developer. (n.d.). https://developer.ibm.com/tutorials/awb-stemming-text-porter-stemmer-algorithm-python/

International Journal of Scientific Research in Science, Engineering and Technology IJSRSET. (2016b). A Survey on Subjective Sentiment Analysis from Twitter Corpus. Technoscienceacademy. https://www.academia.edu/25502258/A_Survey_on_Subjective_Sentiment_Analysis_from_Twitter_Corpus

Juba, B., & Le, H. S. (2019). Precision-Recall versus accuracy and the role of large data sets. Proceedings of the . . . AAAI Conference on Artificial Intelligence, 33(01), 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Kaggle: your machine learning and data science community. (n.d.). https://www.kaggle.com

Kim, Y., Dwivedi, R., Zhang, J., & Jeong, S. R. (2016). Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5. Online Information Review, 40(1), 42–61. https://doi.org/10.1108/oir-03-2015-0068

Kolla, V. R. K. (2016, August 1). Analyzing the Pulse of Twitter: Sentiment Analysis using Natural Language Processing Techniques. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4413716

Larhgotra, A., & Walia, N. K. (2024). Sentiment Analysis Of Twitter Data SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5035972

Mahesh, B. (2020). Machine Learning Algorithms - a review. International Journal of Science and Research, 9(1), 381–386. https://doi.org/10.21275/art20203995

Masui, T. (2024, February 18). All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression. Medium. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Önden, A., Alnour, M., Simic, V., & Pamucar, D. (2024). The evolution of sentiment analysis across various scientific disciplines: A comprehensive review based on the bibliometric technique. Decision Making Advances, 2(1), 222–237. https://doi.org/10.31181/dma21202441

Popescu, M., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. E. (2009). Multilayer perceptron and neural networks. ResearchGate. https://www.researchgate.net/publication/228340819_Multilayer_perceptron_and_neural_networks

Rigatti, S. J. (2017). Random Forest. Journal of Insurance Medicine, 47(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

siebert/sentiment-roberta-large-english · Hugging Face. (n.d.). https://huggingface.co/siebert/sentiment-roberta-large-english

What is a multilayer perceptron (MLP) neural network? (2024, January 23). Shiksha Online. https://www.shiksha.com/online-courses/articles/understanding-multilayer-perceptron-mlp-neural-networks/

Wikipedia contributors. (2025, January 4). Twitter. Wikipedia. https://en.wikipedia.org/wiki/Twitter

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Systems With Applications, 42(7), 3508–3516. https://doi.org/10.1016/j.eswa.2014.12.006

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον.