Εξαμηνιαία Εργασία στην Μηχανική Μάθηση

Δι-ιδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών (Δ.Π.Μ.Σ.), Χειμερινό Εξάμηνο 2025 - 2026

Καράτσαλος Χρήστος - ΑΜ: $mtn2510$

Φαίδων Γραμματικόπουλος - ΑΜ: $mtn2505$

Φεβρουάριος 2026

Περιεχόμενα

1 Εισαγωγή

Αναμφισβήτητα, τα τελευταία χρόνια, το Twitter ανήκει στην λίστα με τα πιο δημοφιλή κοινωνικά δίκτυα παγκοσμίως. Δεν είναι λίγες οι φορές, όμως, που έχει κατηγορηθεί για το γεγονός ότι δεν προστατεύει επαρκώς τους χρήστες του, και ειδικότερα τις γυναίκες, από πιθανές παρενοχλήσεις και ύβρεις. Αρκετά συνηθισμένη είναι πλέον η εμφάνιση σχολίων, τα οποία μπορούν να χαρακτηριστούν ξεκάθαρα ως ρατσιστικά ή μισογυνικά, προβάλλοντας πόσο επιτακτική είναι η ανάγκη άμεσου εντοπισμού και διαχείρισής τους. Δεδομένου, όμως, ότι το Twitter αποτελεί ένα από τα μεγαλύτερα κοινωνικά δίκτυα των ημερών μας, γίνεται εμφανές ότι το συγκεκριμένο ζήτημα δεν πρέπει να θεωρείται καθόλου ως τετριμένο. Ιδανικά, η ύπαρξη ενός μηχανισμού αυτόματης ανίχνευσης και διαχείρισης σχολίων τέτοιου είδους, χωρίς την εμπλοκή του ανθρώπινου παράγοντα, κρίνεται κάτι παραπάνω από αναγκαία.

Για τον λόγο αυτό, στο πλαίσιο της συγκεκριμένης εργασίας επιλέξαμε να ασχοληθούμε με τον αντίστοιχο διαγωνισμό του ECML PKDD 2019 συνεδρίου, βασικό θέμα του οποίου ήταν η ταξινόμηση μηνυμάτων από το Twitter, σε υβριστικά/προσβλητικά ή μη, καθώς επίσης και ο περαιτέρω χαρακτηρισμός των προσβλητικών σχολίων ως εμμέσως προσβλητικά, σεξιστικά καθώς και σχόλια απειλής της σωματικής ακεραιότητας. Το dataset, του διαγωνισμού περιλαμβάνει κείμενα από διάφορα tweets, έχοντας για κάθε ένα από αυτά το αντίστοιχο label.

Στην συγκεκριμένη εργασία, βασικός μας στόχος είναι να προσδιοριστεί η επίδοση κλασικών μοντέλων μηχανικής μάθησης, σε ένα πρόβλημα ιδιαίτερα απαιτητικό, αλλά και η σύγκριση αυτών με μία αρχιτεκτονική βαθιάς μηχανικής μάθησης, που επιτυγχάνει ιδιαίτερα καλές επιδόσεις. Ειδικότερα, αρχικά επιδιώκουμε την επίλυση του προβλήματος με την χρήση κλασικών μοντέλων μάθησης, χρησιμοποιώντας Naive Bayes, Logistic Regression, SVM ταξινομητές. Έπειτα παρουσιάζεται η επίδοση μιας αρχιτεκτονικής βαθιάς μάθησης, καθώς και ορισμένες παραλαγές της, με βάση Recurrent Neural Networks (RNNs), σε συνδυασμό με attention μηχανισμούς.

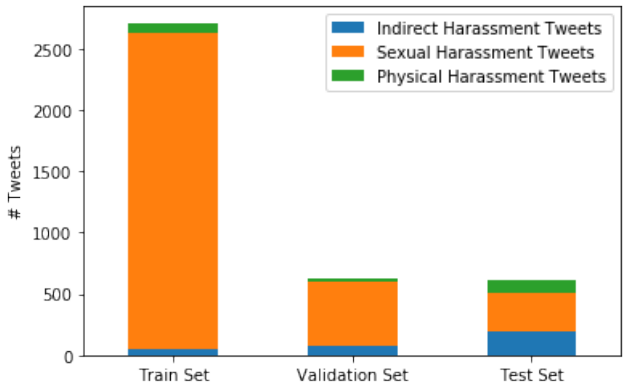
2 Dataset

Το δοθέν dataset αποτελείται από 10.622 posts, συλλεγμένα από το Twitter, στην αγγλική γλώσσα. Χωρίζεται σε train, validation και test set, ενώ επίσης τα κείμενα που το απαρτίζουν έχουν χαρακτηριστεί εκ των προτέρων ως υβριστικά/προσβλητικά ή

μη. Επιπλέον, εκείνα που ανήκουν στην πρώτη κατηγορία, διαχωρίζονται περαιτέρω σε εμμέσως προσβλητικά, σεξιστικά καθώς και σχόλια απειλής της σωματικής ακεραιότητας.

Στα ακόλουθα διαγράμματα απεικονίζονται τα στατιστικά στοιχεία του συγκεκριμένου dataset για τα δύο task του διαγωνισμού.

Σχήμα 1: Στατιστικά του dataset για το πρώτο task



Σχήμα 2: Στατιστικά του dataset για το δεύτερο task

Εύκολα γίνεται αντιληπτό, παρατηρώντας τα παραπάνω διαγράμματα, ότι το συγκεκριμένο dataset είναι imbalanced. Επιπλέον, παρατηρούμε ότι τα tweets που χαρακτηρίζονται ως σεξιστικά ή ότι απειλούν την σωματική ακεραιότητα, είναι λιγότερα στο train set, σε σχέση με τα validation και test sets. Οι δύο αυτοί παράγοντες, επηρεάζουν σημαντικά τα αποτελέσματα αλλά και την επιλογή της μετρικής αξιολόγησης αυτών. Κάτι τέτοιο, φυσικά, είναι ιδιαίτερα συχνό σε datasets που δίνονται σε διάφορους διαγωνισμούς, όπου τα θέματα που συναντώνται βασίζονται συνήθως σε κάποιο real world πρόβλημα.

Σχετικά με την διαδικασία προεπεξεργασίας του dataset, επειδή ακολουθείται διαφορετική προσέγγιση για τα μοντέλα κλασικής μηχανικής μάθησης, σε σχέση με εκείνη που

ακολουθείται για τα μοντέλα βαθιάς μηχανικής μάθησης, λεπτομέρειες δίνονται στο *Κεφ. 4.*

3 Θεωρητικό Υπόβαθρο Μεθόδων Κλασικής Μάθησης

Σε αυτή την ενότητα, θα πραγματοποιηθεί μια σύντομη συζήτηση ([1, 2]), σχετικά με τα κλασικά μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν στην συγκεκριμένη εργασία. Στόχος μας είναι η υπενθύμηση βασικών ιδεών που διέπουν το εκάστοτε μοντέλο, καθώς και η εισαγωγή του απαραίτητου συμβολισμού που θα ακολουθηθεί.

3.1 Naive Bayes Ταξινομητής

Ο Naive Bayes ταξινομητής ήταν ιδιαίτερα δημοφιλής στην κατηγοριοποίηση κειμένου, για αρκετά χρόνια, λόγω της απλότητάς του καθώς και του χαμηλού υπολογιστικού κόστους που συνδέεται με αυτόν. Σήμα κατατεθέν του είναι η υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών, όπου κάποιες φορές ισχύει και στην πραγματικότητα ή τις περισσότερες είναι απλά υπόθεση του συγκεκριμένου μοντέλου.

Δεδομένου ότι

$$p(C_k|\mathbf{x}) \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{1}$$

και χρησιμοποιώντας το γνωστό *MAP* κανόνα απόφασης, το αποτέλεσμα της ταξινόμησης προκύπτει ως εξής

$$\hat{y} = \underset{k \in \{1,2,...,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{2}$$

με $k \in \{1, 2, ..., K\}$ οι διαφορετικές κλάσεις του προβλήματος και $i \in \{1, 2, ..., n\}$ τα χαρακτηριστικά.

Το μόνο που απομένει, είναι η κατάλληλη επιλογή της κατανομής $p(x_i|C_k)$. Η πιο απλή περίπτωση είναι η $p(x_i|C_k)$ να ακολουθεί κατανομή Bernoulli. Μία άλλη επιλογή είναι να ακολουθεί κανονική κατανομή, συνθέτοντας στην περίπτωση αυτή τον Gaussian Naive Bayes ταξινομητή.

3.2 Logistic Regression Ταξινομητής

Ας επικεντρωθούμε αρχικά στο πρόβλημα ταξινόμησης δύο κλάσεων, όπου κάτω από ορισμένες γενικές υποθέσεις, μπορεί να δειχτεί ότι ισχύει

$$p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}), \text{ όπου } \sigma(\alpha) = \frac{1}{1+e^{-\alpha}} \tag{3}$$

$$\text{και προφανώς } p(C_2|\mathbf{x}) = 1 - p(C_1|\mathbf{x})$$

Ακόμη και να μην ισχύει κάτι τέτοιο, στην πράξη, τις περισσότερες φορές θεωρείται ως υπόθεση. Επιχειρώντας να μεγιστοποιηθεί η πιθανοφάνεια, βασιζόμενοι στα δεδομένα εκπαίδευσης, καταλήγουμε στην ελαχιστοποίηση της συνάρτησης διεντροπίας

$$E(\mathbf{w}) = -\sum_{i=1}^{n} y_i \; ln(p(C_1|\mathbf{x}_i)) + (1-y_i) \; ln(p(C_2|\mathbf{x}_i)) \tag{4}$$

με $\mathbf{x}_i$, $\text{i} \in \{1, 2, ..., n\}$ τα διανύσματα του συνόλου εκπαίδευσης. Έπειτα μπορεί να χρησιμοποιηθεί κάποιου είδους επαναληπτικός αλγόριθμος, όπως για παράδειγμα μία κατάλληλη μορφή της μεθόδου κλίσης, προκειμένου να ελαχιστοποιηθεί η (4). Σε συνδυασμό δε με το γεγονός ότι η $E(\mathbf{w})$ είναι κυρτή, γνωρίζουμε εκ των προτέρων ότι το προκύπτον διάνυσμα $\mathbf{w}$, θα αποτελεί ολικό ελάχιστο της $E(\mathbf{w})$.

Στην περίπτωση όπου έχουμε πρόβλημα ταξινόμησης περισσότερων από δύο κλάσεων, υποθέτουμε αρχικά ότι ισχύει

$$p(C_k|\mathbf{x}) = \frac{e^{\mathbf{w}_k^T\mathbf{x}}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T\mathbf{x}}}, k = 1, 2, ..., K \tag{5}$$

και ο στόχος μας είναι η εύρεση των κατάλληλων $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K$, τα οποία μεγιστοποιούν την συνάρτηση πιθανοφάνειας. Παρομοίως με την απλή περίπτωση, χρησιμοποιώντας όμως 1-από-K κωδικοποίηση, καταλήγουμε ότι η συνάρτηση κόστους είναι

$$E(\mathbf{w}_1, ..., \mathbf{w}_K) = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \; ln(p(C_k|\mathbf{x}_i)) \tag{6}$$

Ομοίως με πριν, χρησιμοποιώντας κάποια επαναληπτική αριθμητική μέθοδο, όπως για παράδειγμα μία κατάλληλη εκδοχή της μεθόδου κλίσης, επιτυγχάνεται η εύρεση των βαρών $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K$, που ελαχιστοποιούν την $E(\cdot)$.

3.3 SVM Ταξινομητής

Ο απλός SVM ταξινομητής ανήκει στην κατηγορία των γραμμικών ταξινομητών, επιχειρώντας να διαχωρίσει δύο γραμμικώς διαχωρίσιμες κλάσεις, αφήνοντας τον περισσότερο

¨χώρο' σε κάθε του πλευρά, με στόχο δεδομένα και από τις δύο κλάσεις να μπορούν να κινηθούν λίγο πιο ελεύθερα, με μικρότερο ρίσκο να προκαλέσουν κάποιο σφάλμα. Στην γενικότερη περίπτωση των μη γραμμικών διαχωρίσιμων κλάσεων, το προαναφερθέν υπερεπίπεδο προκύπτει από την επίλυση του ακόλουθου προβλήματος βελτιστοποίησης

$$\max_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\} \tag{7}$$

$$\text{s.t. } 0 \leq \lambda_i \leq C, i \in \{1, 2, ..., n\} \text{ και } \sum_{i=1}^{n} \lambda_i y_i = 0$$

όπου λ_i οι συντελεστές Lagrange των αντίστοιχων δειγμάτων εκπαίδευσης $\mathbf{x}_i$ και y_i οι αντίστοιχες ετικέτες τους.

Πέρα από την απλή έκδοση του SVM ταξινομητή, βασιζόμενοι στο γνωστό *θεώρημα του Mercer* και υποθέτοντας έναν κατάλληλο πυρήνα, το πρόβλημα βελτιστοποίησης της (7) παίρνει την μορφή

$$\max_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right\} \tag{8}$$

$$\text{s.t. } 0 \leq \lambda_i \leq C, i \in \{1, 2, ..., n\} \text{ και } \sum_{i=1}^{n} \lambda_i y_i = 0$$

όπου $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j)$. Στο πλαίσιο της συγκεκριμένης άσκησης, πέρα από την απλή εκδοχή του ταξινομητή, χρησιμοποιούμε

- γραμμικό πυρήνα, $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z} + c$,
- πολυωνυμικό πυρήνα, $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + 1)^q, q > 0$

3.4 Random Forest Ταξινομητής

Ο αλγόριθμος **Random Forest** (Τυχαίο Δάσος) είναι μια ισχυρή και ευέλικτη μέθοδος μηχανικής μάθησης που χρησιμοποιείται τόσο για προβλήματα ταξινόμησης (*classification*) όσο και για προβλήματα παλινδρόμησης (*regression*).

Ανήκει στην κατηγορία των μεθόδων **ensemble learning** (μάθηση συνόλου), η οποία λειτουργεί συνδυάζοντας τις προβλέψεις από πολλαπλά μοντέλα — στην προκειμένη περίπτωση, πολλά δέντρα απόφασης (*decision trees*) — για την παραγωγή ενός πιο ακριβούς και σταθερού αποτελέσματος.

3.4.1 Πώς Λειτουργεί (Βασικές Αρχές)

- **Δημιουργία 'Δάσους' (Ensemble):** Αντί για ένα μόνο δέντρο απόφασης, ο αλγόριθμος δημιουργεί εκατοντάδες ή χιλιάδες δέντρα.
- **Bootstrapping** (Δειγματοληψία): Κάθε δέντρο εκπαιδεύεται σε ένα διαφορετικό, τυχαίο υποσύνολο των δεδομένων εκπαίδευσης με επανόθεση (*with replacement*).
- **Random Feature Selection** (Τυχαία Επιλογή Χαρακτηριστικών): Σε κάθε διαχωρισμό (*split*) ενός δέντρου, ο αλγόριθμος επιλέγει μόνο ένα τυχαίο υποσύνολο των διαθέσιμων χαρακτηριστικών (*features*). Αυτό μειώνει τη συσχέτιση μεταξύ των δέντρων.

3.4.2 Τελική Πρόβλεψη

Η ενσωμάτωση των αποτελεσμάτων γίνεται ως εξής:

1. **Ταξινόμηση:** Εφαρμόζεται 'πλειοψηφική ψήφος' (*majority vote*) μεταξύ όλων των δέντρων.
2. **Παλινδρόμηση:** Λαμβάνεται ο μέσος όρος (*average*) των προβλέψεων όλων των δέντρων.

3.4.3 Πλεονεκτήματα

- **Υψηλή Ακρίβεια:** Παρέχει συνήθως ανώτερες προβλέψεις σε σύγκριση με μεμονωμένα δέντρα απόφασης.
- **Αντοχή σε Overfitting:** Λόγω του συνδυασμού πολλών δέντρων, μειώνεται ο κίνδυνος υπερπροσαρμογής (*overfitting*).
- **Διαχείριση Ελλιπών Δεδομένων:** Μπορεί να διαχειριστεί αποτελεσματικά σύνολα δεδομένων με ελλιπείς τιμές (*missing values*).
- **Αυτόματη Επιλογή Χαρακτηριστικών:** Παρέχει μέτρα για τη σημασία των χαρακτηριστικών (*feature importance*), διευκολύνοντας την ανάλυση των δεδομένων.

4 Προτεινόμενες Μέθοδοι

Στο παρόν κεφάλαιο παρουσιάζονται ορισμένα μοντέλα κλασικής και βαθιάς μάθησης, ειδικά προσαρμοσμένα για το πρόβλημα κατηγοριοποίησης tweets, πράγμα το οποίο αποτελεί και την βασική ενασχόληση μας στο πλαίσιο της συγκεκριμένης εργασίας.

4.1 Κλασική Μηχανική Μάθηση

Στην συγκεκριμένη ενότητα παρουσιάζονται αναλυτικά όλα τα βήματα, με την σειρά που ακολουθήθηκαν, προκειμένου να προσαρμοστούν τα δεδομένα που διαθέτουμε σε κατάλληλη μορφή, ώστε να είναι συμβατά με τα μοντέλα κλασικής μάθησης, ενώ επίσης γίνεται και μία σύντομη αναφορά στον τρόπο αντιμετώπισης ορισμένων πρακτικών δυσκολιών που προέκυψαν στην πορεία. Ειδικότερα, αρχικά γίνεται μία συζήτηση σχετικά με την απαραίτητη προεπεξεργασία των δεδομένων καθώς και με την αναπαράσταση αυτών, ενώ έπειτα πραγματοποιείται ειδική αναφορά στον τρόπο διαχείρισης των ανισοκατανεμημένων κλάσεων που παρουσιάζονται.

4.1.1 Προεπεξεργασία Δεδομένων

Χρησιμοποιώντας κλασικά μοντέλα μηχανικής μάθησης, η φάση της προεπεξεργασίας των δεδομένων είναι ιδιαίτερα σημαντική. Για τον λόγο αυτό, αρχικά, πραγματοποιούμε tokenization, προκειμένου να προσδιοριστούν τα tokens του εκάστοτε tweet, ενώ ταυτόχρονα απομακρύνονται υπερσύνδεσμοι, νούμερα, hashtags ή διάφορα σημεία στίξης. Έπειτα μετασχηματίζεται κάθε λέξη σε μικρά γράμματα και απομακρύνονται stop-words, μιας και αποτελούν λέξεις με αρκετά μεγάλη συχνότητα εμφάνισης, χωρίς όμως ουσιαστικά να προσφέρουν κάποια χρήσιμη, προς εκμάθηση, πληροφορία στο εκάστοτε μοντέλο.

Ύστερα εφαρμόζεται η διαδικασία αποκατάληξης (stemming) χρησιμοποιώντας τον SnowBall stemmer, με στόχο να αναγνωριστούν οι ρίζες των λέξεων που απαρτίζουν τα σχόλια, ανεξάρτητα από την πτώση ή τον χρόνο στον οποίο βρίσκονται. Κάτι τέτοιο έχει ως αποτέλεσμα την περαιτέρω μείωση των όρων που εν τέλει θα χρησιμοποιηθούν για την αναπαράσταση των tweets που ανήκουν στο σύνολο εκπαίδευσης, βελτιώνοντας σημαντικά την επίδοση των μοντέλων κλασικής μάθησης. Σε αυτό το σημείο αξίζει να σημειωθεί ότι η απαλοιφή των stop-words πρέπει να γίνει πριν την εφαρμογή του stemming, καθώς υπάρχει ο κίνδυνος επεξεργασίας ορισμένων λέξεων, οι οποίες εν συνεχεία δεν θα μπορούν να αναγνωριστούν ως stop-words.

Τέλος, για ολόκληρη την φάση προεπεξεργασίας των δεδομένων χρησιμοποιούνται συναρτήσεις που παρέχονται από την NLTK βιβλιοθήκη.

4.1.2 Αναπαράσταση Δεδομένων

Σχετικά με την αναπαράσταση των δεδομένων, χρησιμοποιούμε το μοντέλο Bag-of-words, το οποίο για κάθε tweet, ορίζει μία απλοποιημένη διανυσματική αναπαράσταση, πράγμα που εμφανίζεται αρκετά συχνά σε προβλήματα που σχετίζονται με *επεξεργασία φυσικής γλώσσας*. Ειδικότερα, με βάση το συγκεκριμένο μοντέλο κάθε tweet αναπαρίσταται από ένα σύνολο λέξεων που εμπεριέχονται σε αυτό, με βάση ένα ευρύτερο σύνολο λέξεων που έχει ήδη σχηματιστεί (λεξιλόγιο). Για την αναπαράσταση του εκάστοτε κειμένου υπό την μορφή διανύσματος υπάρχουν διάφορες εναλλακτικές, όπως για παράδειγμα ότι κάθε θέση του διανύσματος σηματοδοτεί την συχνότητα εμφάνισης της λέξης, ή απλά την ύπαρξη ή όχι εκείνης εντός του κειμένου. Η επικρατέστερη επιλογή με βάση την βιβλιογραφία, καθώς και εκείνη που πηγαίνει καλύτερα στα δεδομένα που έχουμε εμείς στην διάθεση μας, είναι η χρήση του term frequency-inverse document frequency (TF-IDF) διανύσματος, με κάθε τιμή του να εκφράζει, κατά ένα τρόπο, πόσο σημαντική είναι η συγκεκριμένη λέξη του λεξιλογίου για το αντίστοιχο tweet. μή ά tweets μή μέ TF-IDF (Term Frequency – Inverse Document Frequency), η οποία μετατρέπει κάθε κείμενο σε ένα αραιό (sparse) διάνυσμα. Κάθε διάσταση αντιστοιχεί σε ένα n-gram και η τιμή της αντανακλά τη σημασία του για το συγκεκριμένο tweet: όροι που εμφανίζονται συχνά στο tweet αλλά σπάνια στο σύνολο του corpus λαμβάνουν υψηλότερο βάρος. Η εξαγωγή χαρακτηριστικών βασίστηκε σε δύο παράλληλους TF-IDF vectorizers μέσω FeatureUnion. Ο πρώτος λειτουργεί σε επίπεδο λέξεων (word-level) με unigrams και bigrams, αποτυπώνοντας τόσο μεμονωμένους όρους όσο και φράσεις δύο λέξεων. Ο δεύτερος λειτουργεί σε επίπεδο χαρακτήρων (character-level, analyzer='char_wb') με n-grams μήκους 3 έως 5, γεγονός που του επιτρέπει να αναγνωρίζει μορφολογικά μοτίβα και ορθογραφικές παραλλαγές που διαφεύγουν από την ανάλυση σε επίπεδο λέξεων. Κάθε vectorizer παράγει έως 10.000 χαρακτηριστικά, δίνοντας ένα συνολικό διάνυσμα περίπου 20.000 διαστάσεων ανά tweet. Και στους δύο vectorizers εφαρμόστηκε λογαριθμική κλιμάκωση (sublinear_tf) καθώς και φιλτράρισμα πολύ σπάνιων (min_df=2) και πολύ κοινών (max_df=0.95) όρων. Το fit πραγματοποιήθηκε αποκλειστικά στο training set ώστε να αποφευχθεί διαρροή πληροφορίας (data leakage) προς τα validation και test sets. Η στατιστική μέτρηση TF-IDF χρησιμοποιείται για την αξιολόγηση της σπουδαιότητας μιας λέξης σε ένα έγγραφο μέσα

σε μια συλλογή εγγράφων (σώμα κειμένων).

1. Term Frequency (TF) Μετρά πόσο συχνά εμφανίζεται ένας όρος t σε ένα έγγραφο d. Ο απλούστερος τύπος είναι:

$$tf(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \tag{9}$$

Όπου $n_{t,d}$ είναι ο αριθμός των εμφανίσεων του όρου t στο έγγραφο d, και ο παρονομαστής είναι το σύνολο των λέξεων στο έγγραφο.

2. Inverse Document Frequency (IDF) Μετρά τη σπουδαιότητα του όρου σε ολόκληρο το σώμα κειμένων D. Οι σπάνιοι όροι έχουν υψηλό IDF:

$$idf(t,D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{10}$$

Όπου:

- N: Το συνολικό πλήθος των εγγράφων στο σώμα κειμένων.
- $|\{d \in D : t \in d\}|$: Ο αριθμός των εγγράφων που περιέχουν τον όρο t.

3. TF-IDF Score Το τελικό βάρος προκύπτει από το γινόμενο των δύο παραπάνω:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{11}$$

4.1.3 Αντιμετώπιση Ανισοκατανεμημένων Κλάσεων

Όπως έχει προαναφερθεί, το γεγονός ότι οι κλάσεις είναι ανισοκατανεμημένες μπορεί να οδηγήσει στην μέτρια ή κακή επίδοση των μοντέλων που προτείνονται. Προκειμένου να αποφευχθεί, έως ένα βαθμό τουλάχιστον, η κυριαρχία των κλάσεων με αρκετά δείγματα, εφαρμόζουμε την μέθοδο Synthetic Minority Over-sampling Technique (SMOTE), βασική ιδέα της οποίας είναι η αύξηση των δειγμάτων μίας κλάσης, όπου αρχικά διαθέτει πολύ λίγα δείγματα, μέσα από την δημιουργία ενός επιθυμητού αριθμού συνθετικών δειγμάτων [3]. Η συγκεκριμένη μέθοδος ανήκει στην ευρύτερη οικογένεια μεθόδων παραγωγής συνθετικών δεδομένων, όπου με την βοήθεια τεχνητών επιχειρείται η εξάλειψη της υπάρχουσας ανισορροπίας που εντοπίζεται.

SMOTE (Synthetic Minority Over-sampling Technique)

Ο αλγόριθμος SMOTE είναι μια τεχνική υπερδειγματοληψίας που χρησιμοποιείται για την αντιμετώπιση του προβλήματος των μη ισορροπημένων συνόλων δεδομένων (imbalanced datasets). Αντί να αναπαράγει υπάρχουσες παρατηρήσεις, δημιουργεί νέα συνθετικά δείγματα για τη μειοψηφική κλάση.

Μηχανισμός Λειτουργίας

Η διαδικασία δημιουργίας συνθετικών δειγμάτων περιλαμβάνει τα εξής βήματα:

- **Επιλογή Δείγματος:** Για κάθε δείγμα x_i της μειοψηφικής κλάσης, υπολογίζονται οι k κοντινότεροι γείτονες (k-nearest neighbors).
- **Επιλογή Γείτονα:** Επιλέγεται τυχαία ένας από τους k γείτονες, έστω x_{zi}.
- **Γραμμική Παρεμβολή:** Το νέο συνθετικό δείγμα x_{new} δημιουργείται τοποθετώντας το σε ένα τυχαίο σημείο στο ευθύγραμμο τμήμα που συνδέει τα δύο αρχικά σημεία.

Μαθηματική Διατύπωση

Ο υπολογισμός του συνθετικού δείγματος δίνεται από τον τύπο:

$$x_{new} = x_i + \lambda \cdot (x_{zi} - x_i) \tag{12}$$

Όπου:

- x_i: Το αρχικό διάνυσμα χαρακτηριστικών της μειοψηφικής κλάσης.
- x_{zi}: Το διάνυσμα ενός από τους k κοντινότερους γείτονες.
- λ: Ένας τυχαίος αριθμός στο διάστημα $[0, 1]$.

Πλεονεκτήματα

1. **Μείωση του Overfitting:** Σε αντίθεση με το Random Oversampling, το SMOTE δεν δημιουργεί αντίγραφα, άρα το μοντέλο μαθαίνει γενικότερα πρότυπα.
2. **Διεύρυνση της Περιοχής Απόφασης:** Επεκτείνει τον χώρο χαρακτηριστικών που καταλαμβάνει η μειοψηφική κλάση, διευκολύνοντας τον ταξινομητή.

4.1.4 Αλγόριθμοι Μάθησης

Έχοντας προηγηθεί η απαραίτητη διαδικασία προεπεξεργασίας και κατάλληλης αναπαράστασης των δεδομένων, καθώς και η εν μέρει αντιμετώπιση της ανισοκατανομής των κλάσεων, χρησιμοποιούμε (Simple) SVM, Linear SVM, Polynomial SVM, Multinomial Naive Bayes και Logistic Regression αλγορίθμους, όπως είναι αυτοί υλοποιημένοι στο sci-kit learn, προκειμένου να διαπιστωθεί η επίδοσή τους (βλ. υποενότητα 5.1). Τέλος έχουμε προσπαθήσει να βελτιστοποιήσουμε τις παραμέτρους του εκάστοτε αλγορίθμου, άλλοτε χειροκίνητα και άλλοτε χρησιμοποιώντας Grid-Search, όπου αυτό ήταν εφικτό.

4.2 Βαθιά Μάθηση

Στην συγκεκριμένη ενότητα παρουσιάζεται αναλυτικά η αρχιτεκτονική βαθιάς μάθησης, καθώς και οι παραλλαγές αυτής, που συνιστούμε, για την επίλυση του προβλήματος κατηγοριοποίησης tweets, πράγμα που αποτελεί τον σκοπό της συγκεκριμένης εργασίας. Αρχικά γίνεται μία σύντομη συζήτηση σχετικά με την προεπεξεργασία των δεδομένων, καθώς και την αντιμετώπιση της ανισοκατανομής μεταξύ των κλάσεων, το οποίο αποτελεί βασικό χαρακτηριστικό του συγκεκριμένου dataset, ενώ έπειτα περιγράφεται λεπτομερέστατα η βασική αρχιτεκτονική βαθιάς μάθησης που ακολουθήθηκε, η οποία κατά κύριο λόγο βασίζεται σε ιδέες που συζητήθηκαν στο [4].

4.2.1 Προεπεξεργασία & Αντιμετώπιση Ανισοκατανεμημένων Κλάσεων

Αρχικά πριν την εκπαίδευση οποιασδήποτε αρχιτεκτονικής, τα δεδομένα υφίστανται κάποιου είδους προεπεξεργασία, χρησιμοποιώντας έναν tweet pre-processor[1], έχοντας ως στόχο τον καθαρισμό και το tokenization του dataset.

Έπειτα, λόγω του γεγονότος ότι υπάρχουν πολύ λιγότερα δείγματα από tweets , στο σύνολο εκπαίδευσης, τα οποία ανήκουν στην κατηγορία των εμμέσως προσβλητικών σχολίων ή στην κατηγορία των σχολίων απειλής της σωματικής ακεραιότητας, σε σχέση με τα αντίστοιχα δείγματα που υπάρχουν στο validation και test set, εφαρμόζεται μία back-translation μέθοδος, προκειμένου να αντιμετωπιστεί το ζήτημα αυτό. Ως αποτέλεσμα, τα δείγματα από tweets των συγκεκριμένων κατηγοριών, που ανήκουν στο σύνολο εκπαίδευσης, μεταφράζονται από τα αγγλικά στα γερμανικά, τα γαλλικά και τα ελληνικά και έπειτα

[1] https://pypi.org/project/tweet-preprocessor/

μεταφράζονται πάλι ξανά στα αγγλικά. Με αυτόν τον απλό τρόπο, αυξάνονται σημαντικά τα δείγματα που διαθέτουμε για τις συγκεκριμένες κατηγορίες, πράγμα που βελτιώνει τις συνθήκες εκπαίδευσης των μοντέλων βαθιάς μάθησης τα οποία συζητούνται στην συγκεκριμένη εργασία, ενισχύοντας περαιτέρω την επίδοσή τους. Στην πραγματικότητα, ως ιδέα είναι αρκετά παραπλήσια με την μέθοδο SMOTE, με μοναδική διαφορά τον απλούστερο και πιο πρακτικό τρόπο υλοποίησής της.

Επιπλέον, προκειμένου να βελτιωθεί και άλλο η απόδοση των μοντέλων, χρησιμοποιούνται pre-trained word embeddings από δύο δισεκατομμύρια δεδομένα παρμένα από το Twitter [5], έχοντας εικοσιεπτά δισεκατομμύρια tokens, για την αρχικοποίηση του embedding επιπέδου.

4.2.2 Βασική Αρχιτεκτονική & Παραλλαγές

Η βασική αρχιτεκτονική βαθιάς μάθησης που ακολουθήθηκε στην συγκεκριμένη εργασία, αφορά τον συνδυασμό ενός RNN με μία attention-based προσέγγιση, για την κατηγοριοποίηση διαφόρων tweets. Γενικά μιλώντας, ως μία παρένθεση, η βαθιά μάθηση προσφέρει αρκετές διαφορετικές αρχιτεκτονικές, οι οποίες επιτυγχάνουν αρκετά καλή επίδοση όταν καλούνται να διαχειριστούν ακολουθιακά δεδομένα. Αυτός ακριβώς είναι και ο λόγος, όπου τα τελευταία χρόνια σε αρκετές εφαρμογές, συμπεριλαμβανομένης και της κατηγοριοποίησης κειμένου, πλέον η ερευνητική κοινότητα έχει στρέψει το ενδιαφέρον της αποκλειστικά σε τέτοιου είδους αρχιτεκτονικές, με τις πιο κλασικές μεθόδους να έχουν περάσει στο παρασκήνιο, προς το παρόν.

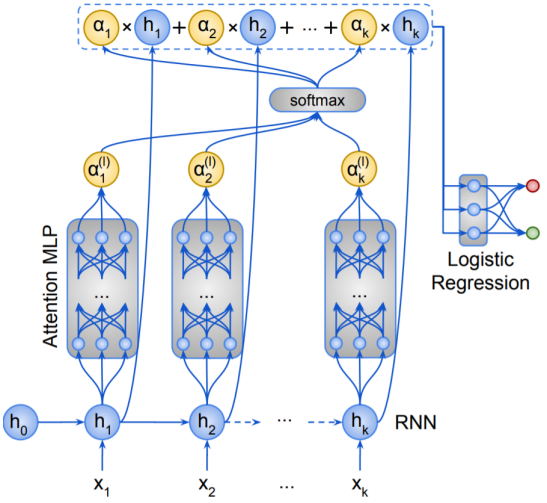
Ειδικότερα, στο πλαίσιο του συγκεκριμένου προβλήματος, είναι προτιμητέα η χρήση RNN εξαιτίας της αποτελεσματικής διαχείρισης ακολουθιακών δεδομένων. Επιλέγεται το RNN μοντέλο να αποτελείται ως μία αλυσίδα από GRU cells (GRUs), τα οποία μετατρέπουν τα tokens $w_1, ..., w_k$ του εκάστοτε tweet σε hidden states $h_1, ..., h_k$, ακολουθούμενα από ένα Logistic Regression (LR) επίπεδο, το οποίο χρησιμοποιεί το τελευταίο hidden state, h_k, προκειμένου να κατηγοριοποιήσει το σχόλιο στην κατάλληλη κλάση. Θεωρώντας ότι έχουν υπολογιστεί τα h_k, πρακτικά το LR επίπεδο, στο τέλος, υπολογίζει την υπό συνθήκη πιθανότητα ενός σχολίου να ανήκει σε κάποια από τις πιθανές κατηγορίες. Για μία πιο λεπτομερή επισκόπηση των μεθόδων βαθιάς μάθησης, που παρουσιάζονται στην συγκεκριμένη εργασία, ο αναγνώστης παραπέμπεται στο [6].

Πέρα από αυτή την απλή σχετικά αρχιτεκτονική, είναι επιθυμητή και η προσθήκη κάποιου attention μηχανισμού, παρόμοιου με εκείνον που παρουσιάζεται στο [4]. Ειδι-

κότερα, στην περίπτωση αυτή, το LR επίπεδο δέχεται ως είσοδο ένα βεβαρημένο άθροισμα των hidden states, της μορφής

$$h_{sum} = \sum_{t=1}^{k} \alpha_t h_t \tag{13}$$

έναντι απλώς του h_k, ενώ οι συντελεστές $\alpha_t, t \in \{1, 2, ..., k\}$, υπολογίζονται με βάση τον attention μηχανισμό που περιγράφεται αναλυτικά στο [4] και απεικονίζεται σχηματικά ακολούθως. Είναι αναγκαίο στο σημείο αυτό να επισημανθεί ότι ο συγκεκριμένος attention μηχανισμός, διαφέρει από άλλους που είχαν αναφερθεί στην βιβλιογραφία, καθώς αναθέτει μεγαλύτερα βάρη, α_t, σε hidden states, τα οποία αντιστοιχούν στην πράξη, σε σημεία του σχολίου όπου υπάρχει εντονότερη απόδειξη ότι ανήκει σε κάποια από τις δεδομένες κλάσεις.



Σχήμα 3: Attention RNN

Σε σχέση με το [4], επιχειρήθηκαν δύο επιπλέον πράγματα. Πρώτον, η χρήση ενός multiattention μηχανισμού, ο οποίος αποτελείται ουσιαστικά από τέσσερις επιμέρους attention μηχανισμούς, έναν για κάθε κατηγορία σχολίου. Δεύτερον, χρησιμοποιήθηκε ένα projection επίπεδο για τα word embeddings κατά το οποίο η διάσταση των pre-trained word embeddings [5] προσαρμόζεται στην διάσταση του προβλήματος, δηλαδή στο πλήθος, k, των hidden states.

Όσον αφορά το πολυπλοκότερο μοντέλο βαθιάς μάθησης που προτείνεται, συνδυάζει ένα projection επίπεδο καθώς και multiattention μηχανισμό. Στην συγκεκριμένη περίπτωση, εν συντομία, αρχικά το embedding επίπεδο αρχικοποιείται χρησιμοποιώντας pre-trained word embeddings ([5]), όπως έχει περιγραφεί σε προηγούμενη υποενότητα. Έπειτα χρησιμοποιείται ένα spatial dropout επίπεδο, προκειμένου να αγνοηθεί ένα συγκεκριμένο

ποσοστό των αρχικών διαστάσεων για κάθε διάνυσμα-λέξη στο σύνολο εκπαίδευσης. Ύστερα, τα word embeddings εισέρχονται σε ένα νευρωνικό δίκτυο ενός επιπέδου, όπου κάθε ένας από τους 128 νευρώνες του, έχει ως συνάρτηση ενεργοποίησης την συνάρτηση υπερβολικής εφαπτομένης. Το συγκεκριμένο δίκτυο χρησιμοποιείται ως projection επίπεδο, καθώς έπειτα τα embeddings εισέρχονται σε ένα GRU μεγέθους 128 (δηλαδή 128 GRU cells)[2]. Έπειτα το output state από τα GRUs εισέρχεται σε τέσσερις attention μηχανισμούς, όπως περιγράφθηκε προηγουμένως. Εν τέλει, υπάρχει ένα νευρωνικό δίκτυο ενός επιπέδου, το οποίο διαθέτει 128 νευρώνες, στο οποίο χρησιμοποιείται ως συνάρτηση ενεργοποίησης η ReLU, προκειμένου να υπολογιστεί το τελικό σκορ για κάθε κατηγορία.

Στο επόμενο κεφάλαιο (βλ. υποενότητα 5.2) δοκιμάζονται διάφοροι συνδυασμοί μοντέλων, για την επίλυση του προβλήματος κατηγοριοποίησης των tweets, έχοντας όλα ως αφετηρία το πολυπλοκότερο μοντέλο που μόλις περιγράφθηκε. Τα απλούστερα μοντέλα τα οποία παρουσιάζονται εκεί, προκύπτουν με παράλειψη του projection επιπέδου ή/και παράλειψη του simple/multi-attention μηχανισμού.

5 Πειράματα

Στο συγκεκριμένο κεφάλαιο παρουσιάζεται η επίδοση των μοντέλων, κλασικής και βαθιάς μάθησης, τα οποία χρησιμοποιήθηκαν, σε συνδυασμό με ένα σύντομο σχολιασμό, αλλά και τις συγκρίσεις μεταξύ τους.

5.1 Κλασική Μάθηση

Καθένας από τους αλγορίθμους κλασικής μάθησης, επιλύει εν τέλει ξεχωριστά τέσσερα προβλήματα δυαδικής ταξινόμησης. Αρχικά, τα δύο task του διαγωνισμού αντιμετωπίζονταν ως ένα πρόβλημα δυαδικής ταξινόμησης, καθώς και ένα πρόβλημα ταξινόμησης περισσοτέρων κλάσεων[3], αντίστοιχα. Παρόλα αυτά, έπειτα από διάφορες δοκιμές, καταλήξαμε ότι οι αλγόριθμοι μας παρουσίαζαν σημαντικά καλύτερη επίδοση στην επίλυση δυαδικών προβλημάτων ταξινόμησης, έναντι προβλημάτων περισσοτέρων κλάσεων.

[2] Σε αυτό το σημείο, είναι σημαντικό να επισημανθεί το γεγονός ότι προτιμάται η χρήση GRUs έναντι LSTMs, εξαιτίας του αισθητά χαμηλότερου υπολογιστικού κόστους το οποίο χαρακτηρίζει τα πρώτα. Επιπλέον, το βασικό πλεονέκτημα των LSTM, το οποίο είναι η διατήρηση μνήμης μεγάλων κειμένων, είναι δύσκολο να εκμεταλλευτεί, δεδομένου ότι τα tweets δεν έχουν τόσο μεγάλο μέγεθος.

[3] Τεσσάρων κλάσεων για την ακρίβεια.

Πίνακας 1: Πίνακας Αποτελεσμάτων Αλγορίθμων Κλασικής Μάθησης

Model	**sexual_f1**	**indirect_f1**	**physical_f1**	**harassment_f1**	**f1_macro**
Logistic Regression	0.81	0.53	0.51	0.73	**0.64**
Linear SVM	0.81	0.51	0.50	0.73	0.64
SVM RBF	0.80	0.48	0.49	0.72	0.62
Multinomial Naive Bayes	0.68	0.50	0.52	0.73	0.61
Random Forest	0.82	0.48	0.49	0.72	0.63
Voting Ensemble	0.81	0.48	0.50	0.72	0.63

Ο πίνακας 1 παρουσιάζει την επίδοση του κάθε αλγορίθμου κλασικής μάθησης στο test set, με τα TF-IDF διανύσματα, όμως, να έχουν υπολογιστεί επάνω σε όλα τα διαθέσιμα δεδομένα. Είναι σημαντικό σε αυτό το σημείο να επισημανθεί ότι, σε περιπτώσεις όπου περιοριζόμασταν μόνο στο σύνολο εκπαίδευσης, για την δημιουργία των TF-IDF διανυσμάτων, οι επιδόσεις των κλασικών μοντέλων, προέκυπταν αισθητά χαμηλότερες, όπως είναι διαισθητικά αναμενόμενο.

5.2 Βαθιά Μάθηση

Όσον αφορά τα μοντέλα βαθιάς μάθησης, καθένα από αυτά παράγει τέσσερα σκορ, τα οποία είναι η πιθανότητα το δεδομένο tweet να ανήκει στην αντίστοιχη κατηγορία. Για κάθε από αυτά, αρχικά ελέγχεται εάν το σκορ που προκύπτει για την πρώτη κλάση (είναι υβριστικό/προσβλητικό - harassment) είναι μεγαλύτερο ή ίσο από 1/3. Στην περίπτωση αυτή, το label harassment παίρνει την τιμή 1, ενώ το είδος του είναι εκείνο το label με το υψηλότερο σκορ, ανάμεσα στις τρεις πιθανές κατηγορίες. Διαφορετικά το label harassment παίρνει την τιμή 0, μαζί με τα υπόλοιπα label. Επιπλέον, δεδομένου ότι κάθε μοντέλο βαθιάς μάθησης επιτελεί ταυτόχρονα και τα δύο task, κρίνεται αναγκαία η επιλογή της loss function, έτσι ώστε να εμπεριέχει κατάλληλα βεβαρημένα τα συστατικά στοιχεία των δύο tasks. Ως συνέπεια, έχει επιλεγεί

$$L = \frac{1}{2}BCE(H) + \frac{1}{2}\left\{\frac{1}{5}BCE(sexualH) + \frac{2}{5}BCE(indirectH) + \frac{2}{5}BCE(physicalH)\right\} \tag{14}$$

με BCE η συνάρτηση διεντροπίας. Παρατηρούμε ότι επιλέγοντας την συγκεκριμένη συνάρτηση, τα δύο task θεωρούνται το ίδιο σημαντικά, ενώ επιπλέον στο δεύτερο task δίνεται μεγαλύτερο βάρος στις δύο κατηγορίες με τα λιγότερα δείγματα. Με αυτόν τον τρόπο

Table 2: Πίνακας Αποτελεσμάτων Μοντέλων Βαθιάς Μάθησης

Model	**sexual_f1**	**indirect_f1**	**physical_f1**	**harassment_f1**	**f1_macro**
attentionRNN	0.674975	0.296320	0.087764	0.709539	0.442150
MultiAttentionRNN	0.693460	0.325338	0.145369	0.700354	0.466130
MultiProjectedAttentionRNN	0.714094	0.355600	0.126848	0.686694	**0.470809**
ProjectedAttentionRNN	0.692316	0.315336	0.019372	0.694082	0.430276
AvgRNN	0.637822	0.175182	0.125596	0.688122	0.40668
LastStateRNN	0.699117	0.258402	0.117258	0.710071	0.446212
ProjectedAvgRNN	0.655676	0.270162	0.155946	0.675745	0.439382
ProjectedLastStateRNN	0.696184	0.334655	0.072691	0.707994	0.452881

επιδιώκεται να περιοριστεί κάπως η επίδραση της ανισοκατανομής των κλάσεων.

Όσον αφορά ορισμένα τεχνικά ζητήματα, το batch size είναι ίσο με 32, ενώ έχουμε επιλέξει 20 εποχές, χρησιμοποιώντας early stopping με patience ίσο με 10. Ως κριτήριο για το early stopping έχει επιλεγεί το μέσο AUC, μιας και τα δεδομένα που έχουμε στην διάθεση μας είναι ανισοκατανεμημένα.

Τα μοντέλα των οποίων δοκιμάζεται η απόδοση είναι

- LastStateRNN: Το τελευταίο hidden state διέρχεται σε ένα πολυεπίπεδο νευρωνικό δίκτυο και έπειτα το LR επίπεδο εκτιμάει τις αντίστοιχες πιθανότητες. Προφανώς πρόκειται για το πιο απλό μοντέλο.

- AvgRNN: Εν αντιθέσει με προηγουμένως, στην συγκεκριμένη περίπτωση λαμβάνεται ο μέσος όρος από όλα τα hidden states.

- attentionRNN: Πρόκειται για το μοντέλο που βασίζεται στο [4] και παρουσιάστηκε διεξοδικά σε προηγούμενη ενότητα.

- MultiAttentionRNN: Όπως έχει προαναφερθεί, το συγκεκριμένο μοντέλο είναι παρόμοιο με το attentionRNN, με μόνη διαφορά ότι περιλαμβάνει τέσσερις ξεχωριστούς attention μηχανισμούς, έναν για κάθε κλάση σχολίων.

Επιπλέον για κάθε ένα από αυτά τα μοντέλα, προστίθεται και η projected έκδοση τους, η οποία επιπρόσθετα περιλαμβάνει ένα projection επίπεδο, η λειτουργία του οποίου έχει περιγραφεί στο προηγούμενο κεφάλαιο.

Τέλος, τα μοντέλα βαθιάς μάθησης αξιολογήθηκαν με βάση το f1 score, το οποίο αποτελεί τον αρμονικό μέσο των precision, recall. Ο πίνακας με τα αποτελέσματα των

μοντέλων προέκυψε ως ο μέσος όρος των αντίστοιχων σκορ, για κάθε μοντέλο, έπειτα από δέκα επαναλήψεις.

6 Σύνοψη

Στο πλαίσιο της συγκεκριμένης εργασίας, ασχοληθήκαμε με το πρόβλημα κατηγοριοποίησης διαφόρων tweets, χρησιμοποιώντας εργαλεία από δύο εντελώς διαφορετικούς κόσμους, εκείνον της κλασικής αλλά και αυτόν της βαθιάς μηχανικής μάθησης. Μεταξύ των δύο αυτών προσεγγίσεων, προκαλεί σε πρώτο πλάνο αίσθηση η διαφορά της επίδοσης που επιτυγχάνεται. Οι κλασικοί αλγόριθμοι φαίνεται να υπερνικούν τα μοντέλα βαθιάς μάθησης.

Γενικά μιλώντας, είναι γεγονός ότι η βαθιά μάθηση, προσφέροντας περισσότερη ευελιξία, επιτυγχάνει καλύτερη απόδοση, έναντι της κλασικής μάθησης, ειδικά όταν αναφερόμαστε σε τέτοιου είδους προβλήματα. Δεν είναι τυχαίο άλλωστε, ότι η κατηγοριοποίηση κειμένου, που αποτελεί υποπεριοχή της *επεξεργασίας φυσικής γλώσσας*, γνωρίζει σημαντική πρόοδο τα τελευταία χρόνια, σε συνδυασμό με την άνθιση της βαθιάς μάθησης. Παρόλα αυτά, γίνεται φανερό ότι, προκειμένου οι αλγόριθμοι κλασικής μάθησης να επιτυγχάνουν εξίσου αρκετά καλή απόδοση, είναι κομβικής σημασίας αφενώς η αντιμετώπιση των ανισοκατανομών μεταξύ των κλάσεων, αλλά κυρίως η ύπαρξη σχετικά περιορισμένης ποικιλίας σημαντικών λέξεων μεταξύ των κειμένων. Δυστυχώς η ευρωστία δεν αποτελεί χαρακτηριστικό των μεθόδων κλασικής μάθησης, καθώς αδυνατούν να λαμβάνουν αποφάσεις με σιγουριά, σε τέτοιες περιπτώσεις, αυξάνοντας με αυτόν τον τρόπο το υφιστάμενο σφάλμα. Εν αντιθέσει, τα μοντέλα βαθιάς μάθησης είναι αρκετά πιο εύρωστα, καθώς δεν εξετάζουν το εκάστοτε κείμενο σε επίπεδο λέξης, αλλά λαμβάνουν υπόψη τους και την ακολουθιακή δομή που το χαρακτηρίζει. Αυτός είναι και ο λόγος που σε προβλήματα του πραγματικού κόσμου, όπου προφανώς διατίθεται μόνο ένα σύνολο εκπαίδευσης, στην καλύτερη περίπτωση, τα μοντέλα βαθιάς μάθησης αναμένεται να ανταποκριθούν επαρκέστερα.

Τα μοντέλα κλασικής μηχανικής μάθησης υπερτερούν σε αυτό το task για αρκετούς λόγους. Πρώτον, το μέγεθος του dataset είναι σχετικά μικρό, γεγονός που ευνοεί τα παραδοσιακά μοντέλα τα οποία γενικεύουν καλύτερα με λιγότερα δεδομένα, ενώ τα νευρωνικά δίκτυα απαιτούν τυπικά μεγαλύτερο όγκο δειγμάτων για να εκπαιδευτούν αποτελεσματικά. Δεύτερον, η χειροποίητη αναπαράσταση μέσω TF-IDF — και ιδιαίτερα τα character n-grams — αποδεικνύεται ιδιαίτερα κατάλληλη για ανίχνευση παρενόχλησης,

καθώς πιάνει ορθογραφικές παραλλαγές και σκόπιμες αλλοιώσεις λέξεων που ξεφεύγουν από τα pre-trained word embeddings. Τρίτον, η προσέγγιση του traditional ML χρησιμοποιεί ξεχωριστό binary classifier ανά target με SMOTE oversampling και εξαντλητικό GridSearchCV tuning, ενώ η deep learning προσέγγιση εκπαιδεύει ένα κοινό δίκτυο με βαρυτημένη multi-task loss, στην οποία τα δύσκολα targets (IndirectH, PhysicalH) ανταγωνίζονται για χωρητικότητα μοντέλου με τα ευκολότερα. Τέλος, τα frozen GloVe embeddings, αν και γενικής χρήσης, δεν αποτυπώνουν εξ ορισμού τη σημασιολογία της παρενόχλησης — σε αντίθεση με τα TF-IDF features, τα οποία μαθαίνουν βάρη απευθείας από τη διανομή του συγκεκριμένου corpus.

Αναφορές

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern recognition; 4th ed.* Elsevier, Burlington, 2008.

[3] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *CoRR*, 2011.

[4] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135. Association for Computational Linguistics.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[6] Christos Karatsalos and Yannis Panagiotakis. Attention-based method for categorizing different types of online harassment language. *CoRR*, 2019.