ΕΕΛΛΑΚ

Εφαρμογές της ΕΛΛΑΚ με τη RoBERTa

Δημήτρης Καστρίτης, *ML/Sysadmin Intern*
Νίνα Γιαλλούση, *Data Product Developer*

Ανίχνευση ψευδών λογαριασμών χρήστη

- Δική μας βάση δεδομένων
- Ντετερμινιστικά φίλτρα προεπεξεργασίας
- Διαστρωματωμένη δειγματοληψία
- Ανθρώπινη επισημείωση
- Επιπτώσεις περιττής πληροφορίας στην εκπαίδευση
- Καθαρισμός δεδομένων

Εντοπισμός κατηγοριών περιεχομένου

- Κατηγορίες μεταδεδομένων σχετικά με ΕΤΑΚ
- Έρευνα, Τεχνολογία, Ανάπτυξη, Καινοτομία
- Δοκιμή με τρείς ετικέτες, τελική μορφή με έξι
- Δειγματοληπτική έρευνα σε κείμενα του εταίρου (Google Scholar)
- Εμπλουτισμός με συνθετικά δεδομένα (GPT-4)
- Ενίσχυση tags χαμηλής επίδοσης με περισσότερα συνθετικά δεδομένα

Παρατηρήσεις

- Πειραματικό σχέδιο για καθορισμό ρυθμού μάθησης
- Διαστρωματωμένη δειγματοληψία για ανθρώπινη επισημείωση
- Δομή προς εκμάθηση έργου (συνδυαστική ανάλυση) και επάρκεια δεδομένων εκπαίδευσης
- Τα δεδομένα εκπαίδευσης πρέπει να αποτυπώνουν την κατανομή των δυνατών εισροών που το μοντέλο θα συναντήσει σε πραγματικές περιστάσεις χρήσης του

Παράδειγμα Πειραματικού Σχεδίου