Η Επεξεργασία Φυσικής Γλώσσας για την πρόβλεψη της επίδοσης: Μια διερευνητική αξιολόγηση δύο συνόλων χαρακτηριστικών

Παπαδήμας Χαράλαμπος, Ραγάζου Βασιλική, Καρασαββίδης Ηλίας
papadimas@uth.gr, ragazou@uth.gr, ikaras@uth.gr
Παιδαγωγικό Τμήμα Προσχολικής Εκπαίδευσης, Πανεπιστήμιο Θεσσαλίας

Περίληψη

Η παρούσα εργασία παρουσιάζει μια μελέτη η οποία εστίασε στην Επεξεργασία Φυσικής Γλώσσας για την πρόβλεψη της επίδοσης φοιτητών σε περιβάλλον ηλεκτρονικής μάθησης. Παρόλο που οι εξελίξεις στο πεδίο της ΕΦΓ είναι εντυπωσιακές, μέχρι σήμερα απουσιάζουν συστηματικές διερευνήσεις του δυναμικού που έχει η ΕΦΓ για την εξαγωγή χαρακτηριστικών που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της επίδοσης. Στη μελέτη συμμετείχαν 201 φοιτητές οι οποίοι παρακολούθησαν 6 βιντεοδιαλέξεις για κάθε μία εκ των οποίων κλήθηκαν να συντάξουν μια μικρή περίληψη. Με βάση την επίδοση τους σε τεστ δηλωτικής γνώσης κατηγοριοποιήθηκαν σε δύο κλάσεις, υψηλής και χαμηλής επίδοσης αντίστοιχα. Η εργασία εστιάζει στην εξαγωγή δύο συνόλων χαρακτηριστικών από τις περιλήψεις και εξετάζει τόσο τη συνάφεια τους όσο και την προβλεπτική τους ισχύ για την ταξινόμηση των φοιτητών στις δύο κλάσεις. Τα αποτελέσματα δείχνουν αφενός ότι τα δύο χαρακτηριστικά αναπαριστούν διαφορετικό σήμα και αφετέρου ότι η ακρίβεια ταξινόμησης είναι υποσχόμενη.

Λέξεις κλειδιά: Επεξεργασία Φυσικής Γλώσσας, Μηχανική Μάθηση, Κειμενικά Χαρακτηριστικά, Ταξινόμηση, Επίδοση

Εισαγωγή

Το ενδιαφέρον για την αξιοποίηση της Επεξεργασίας Φυσικής Γλώσσας (ΕΦΓ) στην εκπαίδευση έχει αυξηθεί σημαντικά κατά την τελευταία δεκαετία. Η έρευνα έχει εστιάσει μεταξύ άλλων σε πρόβλεψη ολοκλήρωσης μαθήματος με βάση τις γραπτές απαντήσεις φοιτητών ως προς την αναμενόμενη χρησιμότητα του μαθήματος πριν την παρακολούθηση του (Robinson et al., 2016), στη βάση γραπτών αξιολογήσεων μετά την παρακολούθηση του μαθήματος (Peng & Xu, 2020), στην έγκαιρη διάγνωση της επίδοσης με βάση τα μηνύματα σε forum συζήτησης (Hung et al., 2020), στην κατηγοριοποίηση βιντεοδιαλέξεων με βάση το περιεχόμενο τους (Dessì et al., 2019), στην κατηγοριοποίηση μηνυμάτων forum συζήτησης ως προς τη γνωστική παρουσία (Neto et al., 2021), στον προσδιορισμό του κατά πόσο ένα μήνυμα σε forum συζήτησης απαιτεί την άμεση παρέμβαση των εκπαιδευτών (Almatrafi et al., 2018), στον εντοπισμό του βαθμού στον οποίο ένα μήνυμα σε forum συζήτησης είναι δηλωτικό σύγχυσης ή όχι (Agrawal et al., 2015; Atapattu et al., 2019), στην ταξινόμηση μηνυμάτων forum συζήτησης σε συναφή ή μη (Wise et al., 2017), στην ταξινόμηση αξιολογήσεων μαθημάτων σε θεματικά πεδία (Hew et al., 2018) και στη χρήση αναστοχαστικών σχολίων στο πλαίσιο ψηφιακού παιχνιδιού (Geden et al., 2021). Παρά την αυξανόμενη αυτή έμφαση, μέχρι σήμερα το ζήτημα της πρόβλεψης της μάθησης από βιντεοδιαλέξεις δεν έχει προσελκύσει το ερευνητικό ενδιαφέρον. Η παρούσα εργασία συνεισφέρει στο πεδίο αυτό συνδυάζοντας χαρακτηριστικά κειμένου από στατιστικές και νευρωνικές προσεγγίσεις επιχειρώντας την πρόβλεψη του επιπέδου μάθησης από βιντεοδιαλεξεις.

Ι. Καζανίδης, Α. Τσινάκος (Επιμ.), Πρακτικά Εργασιών 13ου Πανελλήνιου και Διεθνούς Συνεδρίου «Οι ΤΠΕ στην Εκπαίδευση», Διεθνές Πανεπιστήμιο της Ελλάδας, Καβάλα, 29 Σεπτεμβρίου – 1 Οκτωβρίου 2023. ISSN: 2529-0916, ISBN: 978-618-83186-8-7

Θεωρητικό Πλαίσιο

Η ανυσματική αναπαράσταση κειμένου υλοποιείται με διαφορετικούς τρόπους στο πλαίσιο των στατιστικών και νευρωνικών προσεγγίσεων. Στα πλαίσια της στατιστικής αναπαράστασης κειμένου, το κείμενο μοντελοποιείται ως ένα άνυσμα συχνοτήτων, με τη συχνότητα της κάθε λέξης (Term Frequency) να είναι δηλωτική της βαρύτητας που έχει η λέξη για το συγκεκριμένο έγγραφο. Η προσέγγιση αυτή είναι γνωστή ως σάκος λέξεων (Bag of Words) και βασίζεται στην απλουστευτική υπόθεση ότι ένα κείμενο συγκροτείται από λέξεις οι οποίες είναι ανεξάρτητες μεταξύ τους. Θα πρέπει να σημειωθεί πως οι διαστάσεις των παραγόμενων ανυσμάτων είναι συνάρτηση του μεγέθους του λεξιλογίου, ενώ τα ανύσματα αυτά είναι αραιά καθώς οι περισσότερες τιμές τους είναι μηδενικές.

Η πρόσφατη μετάβαση από τις συμβολικές και στατιστικές προσεγγίσεις στις νευρωνικές οδήγησε σε καταιγιστικές εξελίξεις στον τομέα της ΕΦΓ, αρχικά με την εφαρμογή Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) (π.χ. Word2Vec Mikolov et al., 2013) και στη συνέχεια με την αρχιτεκτονική των Transformers (Vaswani et al., 2017; Devlin et al., 2019). Οι εξελίξεις αυτές επέτρεψαν πιο λεπτομερείς τρόπους αναπαράστασης κειμένων όπως είναι οι ενσωματώσεις λέξης (word embeddings). Στην περίπτωση αυτή, το κείμενο μοντελοποιείται ως ένα άνυσμα σχέσεων μεταξύ μιας λέξης και των γειτονικών της λέξεων (δηλαδή του πλαισίου). Η νευρωνική προσέγγιση βασίζεται στην Κατανεμημένη Υπόθεση (Distributional Hypothesis) (Harris, 1954) σύμφωνα με την οποία λέξεις με παρόμοια σημασία τείνουν να εμφανίζονται σε παρόμοια κειμενικά πλαίσια. Συνεπώς, στη διάρκεια της εκπαίδευσης των μοντέλων Τεχνικών Νευρωνικών Δικτύων (ΤΝΔ), οι ενσωματώσεις λέξης δημιουργούνται με τέτοιο τρόπο ώστε οι λέξεις που έχουν παρόμοια σημασία να βρίσκονται κοντά στον πολυδιάστατο χώρο. Τα παραγόμενα ανύσματα από τα ΝΔ για κάθε λέξη έχουν πολύ μικρότερες διαστάσεις (συνήθως 300) ενώ είναι πυκνά καθώς δεν περιλαμβάνουν μηδενικές τιμές.

Οι νευρωνικές αναπαραστάσεις είναι πιο προηγμένες καθώς αποτυπώνουν τη σημασία των λέξεων λαμβάνοντας υπόψη τις σχέσεις τους. ΤΝΔ όπως το Word2Vec παράγουν πάντοτε στατικές αναπαραστάσεις της σημασίας των λέξεων, δηλαδή μια λέξη θα έχει πάντοτε το ίδιο άνυσμα ανεξαιρέτως του συγκείμενου στο οποίο εμφανίζεται. Ο περιορισμός αυτός αντιμετωπίζεται αποτελεσματικά από τα ΤΝΔ τύπου transformers (Devlin et al., 2019), τα οποία επιτρέπουν τη δημιουργία δυναμικών αναπαραστάσεων, όπου λαμβάνεται υπόψη η σημασία μιας λέξης δεδομένου του συγκειμένου. Θα πρέπει να σημειωθεί πως οι νευρωνικές προσεγγίσεις επιτρέπουν όχι μόνο την αναπαράσταση λέξεων αλλά και μεγαλύτερων κειμενικών ενοτήτων, όπως είναι ακολουθίες λέξεων ή προτάσεις. Για παράδειγμα, το μοντέλο Sentence BERT (SBERT) (Reimers & Gurevych, 2019) αποδεικνύεται καταλληλότερο από το BERT καθώς αξιοποιείται για τη δημιουργία ενσωματώσεων προτάσεων.

Ο υπολογισμός της κειμενικής ομοιότητας παρουσιάζει σημαντικά πλεονεκτήματα όταν χρησιμοποιούνται ενσωματώσεις λέξης. Ειδικότερα, ο υπολογισμός της ομοιότητας δύο κειμένων στην περίπτωση των στατιστικών προσεγγίσεων προϋποθέτει ότι υπάρχουν κοινές λέξεις στα δύο κείμενα. Εάν δεν συμβαίνει αυτό, η κειμενική ομοιότητα θα είναι πολύ μικρή ακόμα και εάν τα δύο κείμενα έχουν την ίδια επακριβώς σημασία. Αντίθετα, στην περίπτωση των νευρωνικών προσεγγίσεων, εάν δύο κείμενα έχουν την ίδια σημασία, η ομοιότητα τους θα είναι πολύ υψηλή - ακόμα και εάν δεν υπάρχουν κοινές λέξεις.

Εστίαση έρευνας

Παρόλο που η ΕΦΓ προσελκύει ολοένα και μεγαλύτερο ερευνητικό ενδιαφέρον σε επίπεδο εκπαίδευσης, η προηγούμενη έρευνα διακρίνεται από δύο σημαντικές ελλείψεις. Πρώτον, όπως προκύπτει από την παραπάνω επισκόπηση, η εστίαση είναι συνήθως σε δυαδική

ταξινόμηση διαφόρων μεταβλητών όπως π.χ. η εννοιολογική σύγχυση (Agrawal et al., 2015; Atapattu et al., 2019), η ανάγκη άμεσης παρέμβασης εκπαιδευτών (Almatrafi et al., 2018), η συνάφεια των μηνυμάτων με το περιεχόμενο του μαθήματος (Wise et al., 2017) κτλ. Συνεπώς, δεν έχει καταγραφεί μέχρι σήμερα ρητή εστίαση στην αξιοποίηση χαρακτηριστικών που εξάγονται από κείμενα που δημιουργούν φοιτητές για την πρόβλεψη της επίδοσης, είτε σε πλαίσια βιντεοδιαλέξεων είτε άλλα. Δεύτερον, με ελάχιστες εξαιρέσεις (π.χ. Atapattu et al., 2019, Geden et al., 2021), η προηγούμενη έρευνα έχει κατά κανόνα υιοθετήσει στατιστικές προσεγγίσεις στη μελέτη κειμένων - συχνά μάλιστα σε συνδυασμό με άλλους γλωσσικούς δείκτες και κλίμακες όπως LIWC και Coh-Metrix. Συνεπώς, απουσιάζει μια συστηματική αξιοποίηση κειμενικών χαρακτηριστικών που βασίζονται σε ανύσματα που αντιστοιχούν σε ενσωματώσεις λέξης, τα οποία αναπαριστούν τη σημασία λέξεων ή προτάσεων.

Η παρούσα διερευνητική μελέτη επιχειρεί να καλύψει αυτό το ερευνητικό κενό εξετάζοντας τον βαθμό στον οποίο οι περιλήψεις που συντάσσουν φοιτητές μετά την παρακολούθηση βιντεοδιαλέξεων μπορούν να αξιοποιηθούν για την πρόβλεψη της επίδοσης τους. Ειδικότερα, η εργασία επιδιώκει τη διεύρυνση της συμβολής της ΕΦΓ στην εξαγωγή χαρακτηριστικών που δυνητικά ερμηνεύουν τη μάθηση από βιντεοδιαλέξεις. Τα ερευνητικά ερωτήματα στα οποία εστίασε η μελέτη είναι τα παρακάτω:

RQ#1: *Ποια είναι η σχέση μεταξύ των στατιστικών και νευρωνικών χαρακτηριστικών κειμένου;*

RQ#2: *Ποιος συνδυασμός χαρακτηριστικών κειμένου επιφέρει υψηλότερη ακρίβεια ταξινόμησης;*

Μέθοδος

Συμμετέχοντες και πλαίσιο έρευνας

Στη μελέτη συμμετείχαν 201 φοιτήτριες και φοιτητές Παιδαγωγικών Τμημάτων σε περιφερειακό ΑΕΙ (95% γυναίκες, 5% άνδρες). Οι ηλικίες των συμμετεχόντων κυμαίνονταν μεταξύ 18 και 45 ετών (M= 20.37, SD = 4). Η συμμετοχή στη μελέτη ήταν εθελοντική ενώ δόθηκε σχετικό βαθμολογικό κίνητρο συμμετοχής.

Υλικά

Στο πλαίσιο της έρευνας, χρησιμοποιήθηκε το Σύστημα Διαχείρισης Μάθησης (ΣΔΜ) Moodle, το οποίο προσαρμόστηκε κατάλληλα. Δημιουργήθηκαν έξι βιντεοδιαλέξεις που κάλυπταν θεμελιώδεις πτυχές των ψηφιακών μέσων (Manovich, 2013). Η θεματική των διαλέξεων περιλάμβανε αντικείμενα όπως η μετάβαση από τα αναλογικά στα ψηφιακά μέσα, η νέα υβριδική οπτική γλώσσα της κινούμενης εικόνας και η ψηφιακή σύνθεση.

Μετρήσεις

Η συλλογή δεδομένων περιλάμβανε (α) γραπτές περιλήψεις και (β) μετρήσεις δηλωτικής γνώσης. Ειδικότερα, μετά την παρακολούθηση κάθε βιντεοδιάλεξης, ζητήθηκε από τους συμμετέχοντες να γράψουν μια σύντομη περίληψη, η οποία κατά την κρίση τους αντικατόπτριζε τις κύριες έννοιες που είχαν παρουσιαστεί. Με βάση το περιεχόμενο της κάθε βιντεοδιάλεξης δημιουργήθηκε ένα τεστ δηλωτικής γνώσης που απαρτίζονταν από 10 ερωτήσεις κλειστού τύπου (Σ-Λ). Μετά την ολοκλήρωση της συγγραφής κάθε περίληψης οι φοιτητές κλήθηκαν να απαντήσουν το εκάστοτε τεστ. Για κάθε σωστή απάντηση η βαθμολογία ήταν 1 βαθμός, ενώ η συνολική μέτρηση της επίδοσης για κάθε βιντεοδιάλεξη υπολογιζόταν αθροίζοντας τις τιμές για κάθε τεστ.

Διαδικασία

Αρχικά, οι φοιτητές ενημερώθηκαν για τον σκοπό της έρευνας και τις προϋποθέσεις συμμετοχής. Μετά την εκδήλωση ενδιαφέροντος και τη λήψη συναίνεσης, δημιουργήθηκαν ατομικοί λογαριασμοί για τους συμμετέχοντες και αποστάλθηκαν οι απαιτούμενες οδηγίες τόσο για την αλληλεπίδραση με το ΣΔΜ και όσο και για την ακολουθούμενη διαδικασία της έρευνας. Η συνολική διάρκεια της μελέτης ήταν 3 περίπου ώρες και η διεξαγωγή της ήταν εξ αποστάσεως. Οι συμμετέχοντες συνδέονταν στο ΣΔΜ και ακολουθούσαν μια γραμμή μάθησης η οποία περιλάμβανε διαδοχικά (α) τη βιντεοδιάλεξη, (β) τη φόρμα σύνταξης της περίληψης και (γ) το τεστ δηλωτικής γνώσης. Οι συμμετέχοντες μπορούσαν να προχωρήσουν στο επόμενο βήμα μόνο αφού ολοκλήρωναν το προηγούμενο (π.χ. δεν μπορούσαν να απαντήσουν το τεστ δηλωτικής γνώσης πριν γράψουν την περίληψη). Μετά την ολοκλήρωση της πρώτης βιντεοδιάλεξης επαναλήφθηκε η ίδια διαδικασία για τις υπόλοιπες 5 βιντεοδιαλέξεις.

Ανάλυση

Εξαγωγή χαρακτηριστικών από περιλήψεις

Το πρωτογενές υλικό της μελέτης αποτέλεσαν (α) οι περιλήψεις των βιντεοδιαλέξεων που έγραψαν οι συμμετέχοντες και (β) τα κείμενα από τις απομαγνητοφωνήσεις των βιντεοδιαλέξεων. Χρησιμοποιώντας τον διαχωριστή της spaCy (Honnibal & Montani, 2017), τα κείμενα αυτά διαχωρίστηκαν σε στοιχεία (tokens) που αντιστοιχούσαν (α) σε λέξεις και (β) σε προτάσεις. Λόγω της διερευνητικής φύσης της έρευνας, δεν υιοθετήθηκαν αυστηρές προδιαγραφές για την κανονικοποίηση των κειμένων όπως είναι η αφαίρεση στοπ λέξεων ή η λημματοποίηση (Παναγιωτακόπουλος κ.α., 2023).

Για τις ανάγκες της μελέτης υιοθετήσαμε (α) στατιστικές και (β) νευρωνικές αναπαραστάσεις κειμένου. Αναφορικά με το πρώτο, δημιουργήθηκαν ανύσματα από συχνότητες λέξεων υιοθετώντας την προσέγγιση σάκος με λέξεις (BoW). Δεδομένου του συνολικού μεγέθους του λεξιλογίου (βιντεοδιαλέξεις και περιλήψεις), τα ανύσματα είχαν περισσότερες από 1.5K διαστάσεις. Αναφορικά με το δεύτερο, δημιουργήθηκαν ανύσματα από ενσωματώσεις λέξης και πρότασης που βασίζονται στην Κατανεμημένη Υπόθεση. Οι ενσωματώσεις λέξης προήλθαν από τη βιβλιοθήκη spaCy (Honnibal & Montani, 2017) και είχαν 300 διαστάσεις ενώ οι αντίστοιχες ενσωματώσεις πρότασης προήλθαν από το μοντέλο SBERT (Reimers & Gurevych, 2019) και είχαν 768 διαστάσεις. Θα πρέπει να σημειωθεί ότι και στις δύο περιπτώσεις χρησιμοποιήθηκαν έτοιμες ενσωματώσεις λέξης ή πρότασης που είχαν εκπαιδευτεί από τους δημιουργούς των αντίστοιχων μοντέλων για την ελληνική γλώσσα. Ειδικότερα, για τις ενσωματώσεις λέξης χρησιμοποιήθηκε το μεγάλο μοντέλο της spaCy 3.5 (el_core_news_lg), που περιλάμβανε 500K μοναδικά ανύσματα και είχε εκπαιδευτεί σε ειδησιογραφικά κυρίως κείμενα (spaCy, 2023). Αντίστοιχα, για τις ενσωματώσεις πρότασης χρησιμοποιήθηκε το μοντέλο sentence-transformers για ομοιότητα προτάσεων της symanto από το αποθετήριο των transformers στο Huggingface (symanto, 2023), το οποίο βασίζεται στο μοντέλο XML-RoBERTa (Conneau et al., 2019).

Τα χαρακτηριστικά που εξήχθηκαν από τις περιλήψεις ήταν τα εξής:

(α) μέσος όρος ομοιότητας συνημιτόνου μεταξύ κάθε πρότασης της περίληψης που είχαν συντάξει οι φοιτητές με κάθε πρόταση της αντίστοιχης βιντεοδιάλεξης, χρησιμοποιώντας τα ανύσματα συχνοτήτων λέξεων (BoW)

(β) μέσος όρος ομοιότητας συνημιτόνου για κάθε πρόταση της περίληψης που είχαν συντάξει οι φοιτητές με κάθε πρόταση της αντίστοιχης βιντεοδιάλεξης, χρησιμοποιώντας τα ανύσματα από ενσωματώσεις πρότασης (SBERT)

(γ) μέσος όρος ομοιότητας συνημιτόνου των 8 υψηλότερων ομοιοτήτων για κάθε πρόταση της περίληψης που είχαν συντάξει οι φοιτητές με κάθε πρόταση της αντίστοιχης βιντεοδιάλεξης, χρησιμοποιώντας τα ανύσματα από ενσωματώσεις πρότασης (SBERT).

Η προκαταρκτική εξέταση της κειμενικής ομοιότητας κάθε πρότασης της περίληψης με κάθε πρόταση της βιντεοδιάλεξης έδειξε ότι στις περιπτώσεις που η σημασία των δύο ήταν πολύ κοντινή, η ομοιότητα έτεινε να είναι υψηλή, γεγονός που ήταν απολύτως αναμενόμενο. Ωστόσο, στις περισσότερες άλλες περιπτώσεις που οι προτάσεις είχαν διαφορετική σημασία, η ομοιότητα μεταξύ των δύο έτεινε να είναι μεσαία ή μικρή. Συνεπώς υποθέσαμε πως ο υπολογισμός του μέσου όρου μεταξύ όλων των προτάσεων (περίληψης και βιντεοδιάλεξης) πιθανόν να μην αντικατοπτρίζει το υφιστάμενο σήμα. Για τον λόγο αυτό δημιουργήθηκε ένα επιπλέον χαρακτηριστικό που βασίζονταν στον μέσο όρο των 8 υψηλότερων τιμών ομοιότητας.

Για τον υπολογισμό της κειμενικής ομοιότητας, χρησιμοποιήθηκε το μέτρο της ομοιότητας συνημιτόνου (cosine similarity). Δεδομένων δύο ανυσμάτων σε έναν χώρο, ο συγκεκριμένος δείκτης τείνει προς το 1 όταν η μεταξύ τους γωνία τείνει να είναι 0, όταν δηλαδή τα ανύσματα αυτά βρίσκονται κοντά στον χώρο αυτό. Αντίστοιχα, όσο τα ανύσματα απομακρύνονται στον χώρο τόσο ο συγκεκριμένος δείκτης θα μειώνεται, φτάνοντας στο 0 όταν τα ανύσματα γίνουν κάθετα μεταξύ τους (Παναγιωτακόπουλος κ.α., 2023).

Η επίδοση στο τεστ δηλωτικής γνώσης σε κάθε βιντεοδιάλεξη χρησιμοποιήθηκε για τη δημιουργία δυαδικών μεταβλητών επίδοσης με τη χρήση της διαμέσου. Με τον τρόπο αυτό δημιουργήθηκαν δύο κλάσεις, μια χαμηλής επίδοσης (μικρότερη από τη διάμεσο) και μια υψηλής επίδοσης (μεγαλύτερη από τη διάμεσο). Δεδομένης της ανισορροπίας των κλάσεων στο σύνολο των βιντεοδιαλέξεων, επιλέξαμε δύο μέτρα για την αξιολόγηση της ταξινόμησης: ακρίβεια (accuracy) και F1 score.

Αποτελέσματα-Συζήτηση

Για την εξέταση του πρώτου ερευνητικού ερωτήματος υπολογίστηκαν οι συνάφειες Pearson μεταξύ της κειμενικής ομοιότητας των 3 βασικών χαρακτηριστικών κειμένου. Οι συνάφειες αυτές παρατίθενται στον Πίνακα 1. Τα αποτελέσματα δείχνουν την ύπαρξη δύο συστηματικών μοτίβων για κάθε βιντεοδιάλεξη. Πρώτο, διαπιστώνεται η ύπαρξη σχεδόν τέλειας συνάφειας (0.89-0.97) μεταξύ των χαρακτηριστικών (α) του μέσου όρου της ομοιότητας κάθε πρότασης της περίληψης με κάθε πρόταση της αντίστοιχης βιντεοδιάλεξης και (β) του μέσου όρου των οκτώ υψηλότερων ομοιοτήτων της κάθε πρότασης της περίληψης με κάθε πρόταση της βιντεοδιάλεξης. Το γεγονός αυτό δείχνει ότι τα δύο αυτά χαρακτηριστικά αναπαριστούν πανομοιότυπο σήμα, το οποίο δεν αλλοιώνεται από τη χρήση μέσου όρου. Συνεπώς, δεν συντρέχει λόγος να υπολογιστεί ο μέσος όρος μόνο για τις υψηλές τιμές ομοιότητας αντί για όλες τις τιμές ομοιότητας. Δεύτερο, οι αντίστοιχες συνάφειες μεταξύ της ομοιότητας των ανυσμάτων ενσωματώσεων πρότασης και ανυσμάτων συχνότητας είναι πολύ μικρότερες, στο εύρος 0.51-0.74. Το στοιχείο αυτό δείχνει ότι τα ανύσματα συχνοτήτων έχουν μερική επικάλυψη με τα ανύσματα ενσωματώσεων πρότασης, δηλαδή αναπαριστούν διαφορετικό σήμα.

Αναφορικά με το δεύτερο ερευνητικό ερώτημα, προχωρήσαμε στη διερευνητική σύγκριση του σήματος των διαφορετικών κειμενικών χαρακτηριστικών τα αποτελέσματα της οποίας παρατίθενται στον Πίνακα 2.

Πίνακας 1. Συνάφειες μεταξύ των δεικτών ομοιότητας

	μ.ο. ανυσμάτων - ενσωματώσεις πρότασης	**μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης**	**βιντεοδιάλεξη**
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.97		**1η**
μ.ο. ανυσμάτων συχνότητας	0.51	0.49	
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.96		**2η**
μ.ο. ανυσμάτων συχνότητας	0.60	0.54	
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.89		**3η**
μ.ο. ανυσμάτων συχνότητας	0.74	0.67	
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.92		**4η**
μ.ο. ανυσμάτων συχνότητας	0.62	0.59	
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.97		**5η**
μ.ο. ανυσμάτων συχνότητας	0.51	0.50	
μ.ο. - 8 υψηλότερων ανυσμάτων - ενσωματώσεις πρότασης	0.97		**6η**
μ.ο. ανυσμάτων συχνότητας	0.59	0.62	

Πίνακας 2. Δείκτες ταξινόμησης με τον αλγόριθμο Λογιστικής Παλινδρόμησης*

	Συχνότητες				**Ενσωματώσεις πρότασης**				**Συνδυασμός**			
ΒΔ	A	P	R	F	A	P	R	F	A	P	R	F
1η	0.62	0.60	**1.00**	0.75	0.59	0.58	**0.97**	0.73	0.61	0.59	**1.00**	0.74
2η	0.52	0.61	0.59	0.60	0.61	0.66	0.73	0.69	0.52	0.62	0.57	0.59
3η	0.66	**1.00**	0.05	0.09	0.64	0.00	0.00	0.00	0.64	0.00	0.00	0.00
4η	**0.74**	0.73	0.92	**0.82**	**0.70**	**0.70**	0.95	**0.80**	**0.74**	0.73	0.92	**0.82**
5η	0.62	0.74	0.44	0.55	0.49	0.67	0.06	0.11	0.64	**0.75**	0.47	0.58
6η	0.64	0.73	0.69	0.71	0.66	0.69	0.85	0.76	0.64	0.73	0.69	0.71

**ΒΔ: Βιντεοδιάλεξη, A: Accuracy, P: Precision, R: Recall, F: F1 score*
Οι υψηλότερες τιμές ανά μέτρο έχουν έντονη μορφοποίηση

Η χρήση της Λογιστικής Παλινδρόμησης (Logistic Regression), του απλούστερου δηλαδή δυνατού ταξινομητή, δεν έδειξε ότι το ένα κειμενικό χαρακτηριστικό υπερτερεί έναντι του άλλου. Ειδικότερα, χρησιμοποιώντας ως προβλεπτική μεταβλητή τα ανύσματα συχνοτήτων ο μέσος όρος ακρίβειας ταξινόμησης για όλες τις βιντεοδιαλέξεις ήταν 0.62. Αντίστοιχα, η

χρήση των ανυσμάτων ενσωματώσεων πρότασης ως προβλεπτική μεταβλητή έδωσε ακρίβεια ταξινόμησης 0.63.

Ο συνδυασμός των 3 αυτών χαρακτηριστικών κειμένου για την πρόβλεψη της επίδοσης δεν οδήγησε σε βελτίωση της ακρίβειας ταξινόμησης καθώς ο μέσος όρος για όλες τις βιντεοδιαλέξεις ήταν 0.63. Συνεπώς η χρήση ενσωματώσεων πρότασης και ο συνακόλουθος υπολογισμός της κειμενικής ομοιότητας μεταξύ της περίληψης και των βιντεοδιαλέξεων δεν επιφέρει βελτίωση της ακρίβειας ταξινόμησης.

Θα πρέπει να σημειωθεί ότι παρόλο που φαινομενικά οι συγκεκριμένοι μέσοι όροι ταξινόμησης αντιστοιχούν σε μικρό κέρδος (~10%) σε σχέση αυτό που θα αναμέναμε από μια απολύτως τυχαία ταξινόμηση (0.50), σε κάποιες από τις βιντεοδιαλέξεις οι τιμές των μέτρων ακρίβειας και F1 είναι υψηλότερες καθώς κυμαίνονται από 0.70-0.80. Από την άποψη αυτή, οι τιμές αυτές είναι αντίστοιχες άλλων που αναφέρονται στη βιβλιογραφία όπου χρησιμοποιούνται ανύσματα ενσωματώσεων λέξης (Geden et al., 2021).

Δεδομένης της διερευνητικής φύσης της μελέτης, χρησιμοποιήσαμε ως προβλεπτικές μεταβλητές αποκλειστικά και μόνο τις κειμενικές ομοιότητες. Δεν προχωρήσαμε στη συστηματική κανονικοποίηση των κειμένων, στη χρήση διαφορετικών αλγορίθμων, στην προσαρμογή των σχετικών υπερπαραμέτρων και στη χρήση των ανυσμάτων ενσωματώσεων λέξεων ή προτάσεων για την πρόβλεψη της επίδοσης. Συνολικά, τα αρχικά αποτελέσματα της παρούσας εργασίας δείχνουν μια πολύ υποσχόμενη εικόνα για το δυναμικό των διαφορετικών χαρακτηριστικών κειμένου για την πρόβλεψη της επίδοσης.

Το μελλοντικό πλάνο ανάλυσης περιλαμβάνει τη μεθοδικότερη κανονικοποίηση κειμένων, τον συνδυασμό χαρακτηριστικών κειμένου που βασίζονται σε ενσωματώσεις προτάσεων και τη χρήση διαφόρων αλγορίθμων μηχανικής μάθησης για την πρόβλεψη της επίδοσης.

Αναφορές

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. *International Educational Data Mining Society*.

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education*, 118, 1-9.

Atapattu, T., Thilakaratne, M., Vivian, R., & Falkner, K. (2019). Detecting cognitive engagement using word embeddings within an online teacher professional development community. *Computers & Education*, 140, 103594.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M. Zettlemoyer, L. & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv*:1911.02116.

Dessì, D., Fenu, G., Marras, M., & Recupero, D. R. (2019). Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections. *Computers in Human Behavior*, 92, 468-477.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805.

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education*, 31, 1-23.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.

Hew, K. F., Qiao, C., & Tang, Y. (2018). Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. *International Review of Research in Open and Distributed Learning*, 19(3).

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hung, J. L., Rice, K., Kepka, J., & Yang, J. (2020). Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content. *Information Discovery and Delivery*, 48(4), 199-212.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Neto, V., Rolim, V., Pinheiro, A., Lins, R. D., Gašević, D., & Mello, R. F. (2021). Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts. *IEEE Transactions on Learning Technologies*, 14(3), 299-312.

Peng, X., & Xu, Q. (2020). Investigating learners' behaviors and discourse content in MOOC course reviews. *Computers & Education*, 143, 103673.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv*:1908.10084.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016, April). Forecasting student achievement in MOOCs with natural language processing. *In Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 383-387).

spaCy. (2023, May 5). *Large Greek language model* (el_core_news_lg). https://spacy.io/models/el Symanto. (2023, May 5). *Siamese network model trained for zero-shot and few-shot text classification* (sn-xlm-roberta-base). https://huggingface.co/symanto/sn-xlm-roberta-base-snli-mnli-anli-xnli Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wise, A. F., Cui, Y., Jin, W., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, 32, 11-28.

Παναγιωτακόπουλος, Χ., Τσαλίδης, Χ., Γάκης, Π., & Κόκκινος, Θ. (2023). Υπολογιστική γλωσσολογία: Από τον προγραμματισμό μέχρι τη διδακτική πράξη [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/8638