ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Η Βάση Δεδομένων των Απεργιών και των Διαδηλώσεων στην Ευρωπαϊκή Ένωση

Αλκμήνη Σούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Θεόδωρος Τζουραμάνης
Επίκουρος Καθηγητής

Λαμία, έτος 2025

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Η Βάση Δεδομένων των Απεργιών και των Διαδηλώσεων στην Ευρωπαϊκή Ένωση

Αλκμήνη Σούλα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Θεόδωρος Τζουραμάνης
Επίκουρος Καθηγητής

Λαμία, έτος 2025

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 23/09/2025

Η Δηλ.

Αλκμήνη Σούλα

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η Βάση Δεδομένων των Απεργιών και των Διαδηλώσεων στην Ευρωπαϊκή Ένωση

Αλκμήνη Σούλα

Τριμελής Επιτροπή:

Θεόδωρος Τζουραμάνης, Επίκουρος Καθηγητής

Αθανάσιος Κακαρούντας, Καθηγητής

Σωτήριος Τασουλής, Επίκουρος Καθηγητής

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Περιεχόμενα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Ευρετήριο Εικόνων

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

1. Εισαγωγή και πλαίσιο της πτυχιακής εργασίας

1.1 Εισαγωγή

Η ραγδαία εξέλιξη της τεχνολογίας και η ευρεία διάδοση των ψηφιακών συσκευών τις τελευταίες δεκαετίες έχουν οδηγήσει στην εκθετική αύξηση της παραγωγής και ροής δεδομένων παγκοσμίως. Η καθημερινή χρήση του Διαδικτύου, των κοινωνικών δικτύων και των ηλεκτρονικών υπηρεσιών έχει δημιουργήσει έναν τεράστιο όγκο πληροφοριών, μετατρέποντας το Διαδίκτυο σε έναν βασικό πυλώνα επικοινωνίας, πληροφόρησης και κοινωνικής δραστηριότητας.

Κατά την περίοδο μεταξύ 2000 και 2009, το Διαδίκτυο μετατράπηκε από μια αναδυόμενη τεχνολογία επικοινωνιών σε ένα ώριμο και ευρέως διαδεδομένο μέσο μαζικής ενημέρωσης. Σύμφωνα με την εργασία (Yadav, 2010), που αξιοποίησε τα στατιστικά δεδομένα του Internet World Stats, έδειξε ότι η αύξηση της παγκόσμιας βάσης χρηστών του Διαδικτύου ήταν της τάξης του 380% κατά την προαναφερθείσα περίοδο, με τον αριθμό των χρηστών να ξεπερνά τα 1,8 δισεκατομμύρια μέχρι το τέλος του 2009.

Παρά την κυριαρχία του Διαδικτύου, οι έντυπες εφημερίδες διατηρούν ορισμένα πλεονεκτήματα. Όπως επισημαίνει ο Meyer (Meyer, 2004), λειτουργούν ως φυσικοί κατάλογοι, επιτρέποντας στους αναγνώστες να περιηγούνται ελεύθερα στις σελίδες για να αναζητήσουν πληροφορίες. Ωστόσο, η ψηφιακή ενημέρωση προσφέρει ταχύτερη, πιο διαδραστική και εξατομικευμένη πρόσβαση σε πληροφορίες, γεγονός που μειώνει σταδιακά τη σημασία του έντυπου τύπου. Η συνύπαρξη αυτών των δύο μέσων υποδηλώνει ότι η εξάπλωση του Διαδικτύου συνδέεται άμεσα με την υποχώρηση των έντυπων εφημερίδων, οι οποίες χάνουν σταθερά αναγνωστικό κοινό.

Οι διαμαρτυρίες, όπως οι απεργίες, οι διαδηλώσεις, οι πορείες και οι πολιτικές αναταραχές, αποτελούν κάποιες από τις πιο προφανείς και αναγνωρίσιμες εκφράσεις κοινωνικής σύγκρουσης. Δεν περιορίζονται σε μεμονωμένα περιστατικά, αλλά συνιστούν πολύπλοκα κοινωνικά φαινόμενα που περιλαμβάνουν διάφορες δυναμικές, από την ατομική πρωτοβουλία έως τη συλλογική οργάνωση. Για το λόγο αυτό, είναι αντικείμενο έρευνας όχι μόνο στον τομέα των κοινωνικών κινημάτων αλλά και σε ένα ευρύτερο πλαίσιο κοινωνιολογικών θεωρήσεων που εξετάζουν τις μορφές, τις αιτίες και τις συνέπειες των κοινωνικών συγκρούσεων.

Σε πολιτικό επίπεδο, οι διαμαρτυρίες λειτουργούν ως μέσα έκφρασης της δυσαρέσκειας των πολιτών, εκφράζοντας τις αντιρρήσεις και τα αιτήματά τους ενάντια στην πολιτική και οικονομική εξουσία. Επιπλέον, αποτελούν καίρια εργαλεία για την προώθηση αλλαγών και τον επανακαθορισμό των κοινωνικών σχέσεων, ιδιαίτερα όταν τα θεσμικά κανάλια εκπροσώπησης κρίνονται ανεπαρκή ή αναποτελεσματικά.

Σε κοινωνικό και πολιτισμικό επίπεδο, οι διαμαρτυρίες λειτουργούν ως καταλύτες κοινωνικής συνειδητοποίησης, ενδυναμώνοντας και καλλιεργώντας την αλληλεγγύη μεταξύ ατόμων και ομάδων που μοιράζονται κοινά συμφέροντα και αξίες. Η συχνότητα και η έντασή τους μπορούν να ερμηνευθούν ως δείκτες βαθμού

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

κοινωνικής ανησυχίας, του επιπέδου δημοκρατικής συμμετοχής, καθώς και της πολιτικής σταθερότητας ή αστάθειας μιας κοινωνίας. Αυτή η ερμηνευτική διάσταση καθιστά τη μελέτη τους ιδιαίτερα σημαντική, καθώς δεν περιορίζεται στην απλή καταγραφή γεγονότων, αλλά εκτείνεται στην κατανόηση των βαθύτερων κοινωνικών, πολιτικών και οικονομικών δομών που τις πυροδοτούν, και την καταγραφή των επιπτώσεών τους στη διαμόρφωση και την εξέλιξη της κοινωνίας.

Στο πλαίσιο αυτό, οι ψηφιακές εφημερίδες αποτελούν μια πολύτιμη πηγή για τη μελέτη των κοινωνικών κινητοποιήσεων, προσφέροντας πληροφορίες σε πραγματικό χρόνο που μπορούν να αναλυθούν συστηματικά. Η παρούσα πτυχιακή εργασία αξιοποιεί αυτόν τον πλούτο ψηφιακών δεδομένων, αναπτύσσοντας μια αυτοματοποιημένη μεθοδολογία για τον εντοπισμό, τη μετάφραση, την κατηγοριοποίηση και τη χωροχρονική ανάλυση περιστατικών διαμαρτυρίας, με έμφαση στις χώρες της Ευρωπαϊκής Ένωσης. Μέσω της συνδυαστικής χρήσης τεχνικών εξόρυξης δεδομένων, NER και τεχνητής νοημοσύνης, η εργασία στοχεύει να συμβάλει στην τεκμηριωμένη καταγραφή και στην εμβάθυνση της κατανόησης των σύγχρονων μορφών κοινωνικής δράσης.

1.2 Στόχος πτυχιακής εργασίας

Η παρούσα πτυχιακή εργασία στοχεύει στη συστηματική και αυτοματοποιημένη συλλογή, επεξεργασία και ανάλυση δεδομένων σχετικών με κοινωνικές κινητοποιήσεις, όπως απεργίες, πορείες και διαδηλώσεις, σε όλες τις χώρες της Ευρωπαϊκής Ένωσης. Ο πρωταρχικός σκοπός είναι η ανάπτυξη μιας ολοκληρωμένης βάσης δεδομένων, η οποία θα αποτυπώνει με ακρίβεια και πληρότητα τη χωρική, χρονική και θεματική διάσταση των σύγχρονων μορφών κοινωνικής διαμαρτυρίας.

Για την υλοποίηση αυτού του στόχου, η εργασία αξιοποιεί προηγμένες τεχνικές NER και μηχανικής μάθησης, οι οποίες επιτρέπουν τον αυτοματοποιημένο εντοπισμό και εξόρυξη ειδησεογραφικών αναφορών από ποικίλες ψηφιακές πηγές, όπως ηλεκτρονικές εφημερίδες και ειδησεογραφικά ιστοσελίδες. Η διαδικασία περιλαμβάνει στάδια όπως η συλλογή άρθρων, η προεπεξεργασία των κειμένων, η αυτόματη μετάφραση και η εξαγωγή βασικών χαρακτηριστικών. Αυτά τα χαρακτηριστικά καλύπτουν στοιχεία όπως η γεωγραφική τοποθεσία του γεγονότος, ο εκτιμώμενος αριθμός συμμετεχόντων, η παρουσία βίαιων ή έκρυθμων επεισοδίων (π.χ. συλλήψεις, τραυματισμοί ή χρήση βίας και όπλων), το χρονικό πλαίσιο διεξαγωγής, καθώς και κατηγοριοποίηση του είδους δράσης.

Η συγκέντρωση αυτών των δεδομένων σε μια ενιαία, τυποποιημένη και οργανωμένη βάση δεδομένων καθιστά εφικτή την εφαρμογή ποικίλων μεθόδων ανάλυσης, τόσο ποσοτικών όσο και ποιοτικών, διευκολύνοντας, έτσι, τον εντοπισμό μακροπρόθεσμων τάσεων, γεωγραφικών συγκεντρώσεων και συσχετίσεων με εξωτερικούς παράγοντες, όπως οικονομικές κρίσεις ή πολιτικές εξελίξεις, αποτυπώνοντας τη δυναμική των κοινωνικών κινητοποιήσεων στον ευρωπαϊκό χώρο.

Η σημασία αυτής της μελέτης ενισχύεται στο πλαίσιο του σύγχρονου κοινωνικοπολιτικού περιβάλλοντος, όπου παρατηρούνται αυξανόμενες κοινωνικές πιέσεις, διεύρυνση ανισοτήτων, κλιματική αβεβαιότητα και προκλήσεις στη λειτουργία δημοκρατικών θεσμών. Μέσω αυτής της προσέγγισης, η εργασία δεν

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

περιορίζεται σε ακαδημαϊκή ανάλυση, αλλά συνδέεται με πρακτικά ζητήματα, όπως η πολιτική ανάλυση, η προώθηση κοινωνικής δικαιοσύνης και η διαμόρφωση στρατηγικών για ενίσχυση της συμμετοχής και του δημοκρατικού διαλόγου.

Η παρούσα μελέτη επιδιώκει να εμβαθύνει στην ανάλυση και κατανόηση των σύγχρονων κοινωνικών κινημάτων, προσφέροντας ένα χρήσιμο εργαλείο για ερευνητές, πολιτικούς και υπεύθυνους λήψης αποφάσεων. Η καταγραφή και μελέτη των κοινωνικών αιτημάτων επιτρέπει την ανάδειξη φαινομένων που συχνά παραμένουν αθέατα στο δημόσιο διάλογο, φωτίζοντας έτσι πλευρές της κοινωνικής δυναμικής που δεν γίνονται πάντα αντιληπτές. Για να επιτευχθεί αυτή η συστηματική αποτύπωση, αξιοποιείται η αυτοματοποιημένη συλλογή και ανάλυση ειδησεογραφικών κειμένων σε μια ενιαία βάση δεδομένων, σχετικά με γεγονότα απεργιών, πορειών και διαδηλώσεων για όλα τα κράτη-μέλη της Ε.Ε. Με αυτόν τον τρόπο, εξασφαλίζεται όχι μόνο η συγκρισιμότητα σε διασυνοριακό επίπεδο, αλλά και η χρονική συνέχεια και η δυνατότητα αναπαραγωγής των ευρημάτων.

Η χρήση της αγγλικής γλώσσας διευκολύνει τη διαλειτουργικότητα και την διεθνή προσβασιμότητα, επιτρέποντας την ανάλυση χρονοσειρών, την εξέταση σχέσεων με κοινωνικοοικονομικούς δείκτες και τη δημιουργία αξιόπιστων ερευνητικών μεθόδων. Επιπλέον, οι οργανώσεις υπεράσπισης δικαιωμάτων, όπως είναι η Amnesty International και η Human Rights Watch, μπορούν να αξιοποιήσουν τους παραγόμενους δείκτες για την έγκαιρη ανίχνευση αλλαγών στην κινητοποίηση, την τεκμηρίωση περιστατικών και την προετοιμασία στοχευμένων παρεμβάσεων. Παράλληλα, οι φορείς πολιτικής ανάλυσης αποκτούν εργαλεία που ενισχύουν τη θεμελιωμένη αξιολόγηση μεταρρυθμίσεων και επιτρέπουν τη συστηματική χωροχρονική αποτύπωση προτύπων κοινωνικής συμμετοχής.

Τέλος, η αρχιτεκτονική του συστήματος έχει σχεδιαστεί ώστε να είναι συμβατή με προϋπάρχουσες πλατφόρμες ανάλυσης, επιτρέποντας την ενσωμάτωση διαφορετικών πηγών δεδομένων ενισχύοντας την ποιότητα μέσω διαδικασιών επικαιροποίησης και επαλήθευσης. Έτσι, το έργο αυτό δεν περιορίζεται μόνο στην επιστημονική ανάλυση, αλλά παρέχει ένα αξιόπιστο εργαλείο που μπορεί να αξιοποιηθεί τόσο στην επιστημονική έρευνα όσο και στη δημιουργία πολιτικών που θα ανταποκρίνονται στις πραγματικές ανάγκες της κοινωνίας, ανοίγοντας νέους δρόμους για μελλοντικές μελέτες και παρεμβάσεις.

1.3 Η έννοια της απεργίας, της διαδήλωσης και της πορείας

Οι απεργίες, οι διαδηλώσεις και οι πορείες αποτελούν βασικές μορφές δημόσιας κινητοποίησης και συλλογικής διεκδίκησης. Πρόκειται για διακριτές αλλά αλληλένδετες πρακτικές δημόσιας έκφρασης, μέσω των οποίων κοινωνικοί και επαγγελματικοί φορείς, κινήματα και πολίτες αρθρώνουν αιτήματα και ασκούν πίεση για τη διαμόρφωση δημόσιων πολιτικών, θεσμικών ρυθμίσεων και κοινωνικών πρακτικών.

Η απεργία (strike) είναι ένα φαινόμενο κατά το οποίο μια ομάδα εργαζομένων διακόπτει την εργασία με σκοπό να ασκήσει πίεση για την ικανοποίηση συγκεκριμένων αιτημάτων ή την επίτευξη βελτιώσεων στους όρους εργασίας (Rath, 2005). Αν και η έννοιά της φαίνεται απλή, στην πραγματικότητα, αποτελεί ένα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

εξαιρετικά σύνθετο φαινόμενο που περιλαμβάνει πολλαπλά αίτια, συνέπειες και διαδικασίες.

Η απεργία αποτελεί εργαλείο εκδήλωσης δυσαρέσκειας απέναντι σε εξαντλητικές συνθήκες εργασίας και εκμετάλλευση των εργαζομένων, καθώς και διεκδίκησης για την βελτίωση των συνθηκών αυτών και την απόκτηση περισσότερων δικαιωμάτων. Από την πλευρά των εργαζομένων, συνίσταται στη διακοπή της παραγωγής και στην άσκηση πίεσης για τη βελτίωση των όρων εργασίας. Διακρίνεται από άλλες συναφείς πρακτικές, όπως η στάση εργασίας (work stoppage), η οποία αφορά σε προσωρινή διακοπή της εργασίας, και άλλες μορφές κινητοποίησης που δεν περιλαμβάνουν πλήρη παύση αυτής, όπως η μειωμένη απόδοση στην εργασία.

Αντιθέτα, ο εργοδοτικός αποκλεισμός (lockout) αποτελεί μια διαφορετική πρακτική, διότι η απόφαση για τη διακοπή της εργασίας λαμβάνεται από τους εργοδότες, και όχι από τους εργαζόμενους. Ιστορικά, οι απεργίες έχουν διαδραματίσει καθοριστικό ρόλο στη βελτίωση των εργασιακών συνθηκών και στην κατοχύρωση κοινωνικών και εργατικών δικαιωμάτων.

Η διαδήλωση (demonstration, rally) αποτελεί μορφή δημόσιας και συλλογικής συγκέντρωσης σε ανοιχτό χώρο, με κύριο στόχο την έκφραση και προβολή κοινωνικών, πολιτικών ή οικονομικών αιτημάτων, καθώς και την άσκηση πίεσης στους φορείς λήψης αποφάσεων. Χαρακτηρίζεται από τη χρήση πανό, συνθημάτων και ομιλιών, ενώ λειτουργεί ως μέσο δημόσιας αντιπροσώπευσης και κοινωνικής συμμετοχής, θεμελιωμένο από τις αρχές ελευθερίας της συνάθροισης και έκφρασης.

Το ελάχιστο που απαιτείται για να θεωρηθεί μια διαμαρτυρία διαδήλωση είναι να περιλαμβάνει τέσσερα βασικά στοιχεία.

Το πρώτο στοιχείο είναι η προσωρινή κατάληψη ενός ανοιχτού φυσικού χώρου, είτε δημόσιου είτε ιδιωτικού. Αυτό εξαιρεί πολλές άλλες μορφές συναντήσεων ή συνελεύσεων, όπως οι πολιτικές συγκεντρώσεις σε κλειστούς χώρους ή οι πορείες σε εσωτερικούς χώρους μιας επιχείρησης.

Το δεύτερο στοιχείο είναι η έκφραση. Το κύριο χαρακτηριστικό της διαδήλωσης είναι η έκφραση κοινωνικών απαιτήσεων μέσω ορατών εκδηλώσεων μιας ομάδας, που μπορεί να υπάρχει ήδη ή να δημιουργείται για την εκδήλωση αυτών των απαιτήσεων. Αυτό μας επιτρέπει να διακρίνουμε τις διαδηλώσεις από άλλες συγκεντρώσεις, όπως το πλήθος σε μια αγορά ή ομάδες με άλλους σκοπούς.

Το τρίτο στοιχείο είναι ο αριθμός των συμμετεχόντων. Για να χαρακτηριστεί μια συγκέντρωση διαδήλωση, απαιτείται ένα ελάχιστο πλήθος ατόμων. Δεν υπάρχει αυστηρός αριθμός, αλλά είναι απαραίτητο να υπάρχει συλλογική δράση. Αυτή η διάκριση επισημαίνει τη διαφορά ανάμεσα σε ατομικές δράσεις και συλλογικές διαδηλώσεις.

Τέλος, το τέταρτο στοιχείο είναι η πολιτική φύση της διαδήλωσης. Μερικά γεγονότα που φαίνονται αρχικά μη πολιτικά μπορούν να εξελιχθούν σε πολιτικές εκδηλώσεις. Η διαδήλωση πρέπει να μεταφέρει ή να προωθεί απαιτήσεις πολιτικού ή κοινωνικού χαρακτήρα (Fillieule).

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η πορεία (march) αποτελεί μια κινητική μορφή διαδήλωσης, κατά την οποία οι συμμετέχοντες διανύουν μια προκαθορισμένη διαδρομή σε δημόσιο χώρο, συνήθως με σκοπό να ενισχύσουν την προβολή των αιτημάτων τους και να ζητήσουν πολιτική ή κοινωνική αλλαγή. Διακρίνεται από άλλες μορφές διαμαρτυρίας λόγω της οργανωμένης και δημόσιας φύσης της και έχει συχνά πολιτική ή στρατιωτική χροιά (Debouzy, 2003).

Η δυναμική της κίνησης ενισχύει την προβολή και την αναγνωρισιμότητα των αιτημάτων, ενώ συχνά επιφέρει προσωρινές επιπτώσεις στην καθημερινή δραστηριότητα, όπως στην κυκλοφορία, χωρίς όμως να συνεπάγεται αναγκαστικά τη διακοπή της παραγωγικής διαδικασίας, όπως συμβαίνει στην απεργία. Η πορεία μπορεί να περιλαμβάνει ανθρώπους που συμμετέχουν ατομικά ή συλλογικά, σε μεγάλες ή μικρές ομάδες, και συχνά χρησιμοποιεί συμβολικές κινήσεις, όπως οι διάφορες πορείες ανέργων που πραγματοποιήθηκαν στις Ηνωμένες Πολιτείες το 1894, με στόχο την πολιτική πίεση (Debouzy, 2003). Συναφείς αλλά διακριτές μορφές περιλαμβάνουν την καθιστική διαμαρτυρία (sit-in), την κατάληψη (occupation), τον αποκλεισμό (blockade) και τη συγκέντρωση ή συλλαλητήριο (rally).

Παρότι και οι τρεις μορφές έχουν ως κοινό σκοπό την συλλογική διεκδίκηση, διαφοροποιούνται ως προς τον φορέα υλοποίησης, τον τρόπο πίεσης και τον άμεσο αποδέκτη. Στην πράξη, συχνά απεργιακές κινητοποιήσεις πλαισιώνονται από διαδηλώσεις και πορείες, ενισχύοντας το εύρος και την ένταση της συλλογικής δράσης.

Η ευρωπαϊκή και διεθνής ιστορική πορεία αναδεικνύει ότι οι απεργίες, οι διαδηλώσεις και οι πορείες έχουν λειτουργήσει καθοριστικά στην κοινωνική και πολιτική αλλαγή. Από τις μαζικές κινητοποιήσεις των εργατών μέχρι τις σύγχρονες διαδηλώσεις στα μεγάλα αστικά κέντρα, εκφράζουν τη δυναμική των κοινωνικών αιτημάτων και διευκολύνουν τη σύνδεση μεταξύ της κοινωνίας και των θεσμών. Η σαφής διάκριση και η παράλληλη ανάλυση αυτών των φαινομένων επιτρέπουν την καταγραφή προτύπων κινητοποίησης, καθώς και της έντασης, της διάρκειας και της εξέλιξης των αιτημάτων στο χρόνο και στο χώρο. Στην παρούσα εργασία, αυτοί οι ορισμοί μετατρέπονται σε συγκεκριμένα κριτήρια για την επισήμανση ειδησεογραφικών αναφορών, με σκοπό την παραγωγή δομημένων, συγκρίσιμων και αναπαραγώγιμων δεδομένων που θα επιτρέψουν τη διαχρονική και διασυνοριακή ανάλυση.

Το ερευνητικό ενδιαφέρον για τις απεργίες και τις ευρύτερες κοινωνικές κινητοποιήσεις παραμένει διαχρονικό στις κοινωνικές επιστήμες, λόγω της αυξημένης έντασης των τελευταίων χρόνων, καθώς αναδεικνύει τις συνθήκες, τους τόπους και τους λόγους εκδήλωσής τους. Η προτεινόμενη συστηματοποίηση ενισχύει αυτή τη συζήτηση, παρέχοντας σαφή δεδομένα και αναλυτικά εργαλεία για την κατανόηση της κοινωνικής δυναμικής, της πολιτικής σταθερότητας και της ποιότητας της δημοκρατίας, συμβάλλοντας στην οργανωμένη καταγραφή, ανάλυση και ερμηνεία του φαινομένου.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

2. Πληροφοριακά συστήματα καταγραφής γεγονότων διαμαρτυρίας σε πραγματικό χρόνο

Η συλλογή και ανάλυση δεδομένων σχετικών με διαμαρτυρίες κατέχει κεντρική θέση στη μελέτη των κοινωνικών κινημάτων ήδη από τη δεκαετία του 1960. Στο πεδίο της κοινωνιολογίας, οι ερευνητές έχουν βασιστεί εκτενώς σε αναφορές των μέσων ενημέρωσης, προκειμένου να κατανοήσουν τη δυναμική των κινητοποιήσεων, διερευνώντας τα αίτια, τους μηχανισμούς και τις συνέπειες της συλλογικής δράσης. Από την άλλη πλευρά, η πολιτική επιστήμη έχει προσεγγίσει τις διαδηλώσεις ως πηγές εμπειρικών δεδομένων για τη μέτρηση και πρόβλεψη φαινομένων πολιτικής αστάθειας και συγκρούσεων.

Στο πλαίσιο αυτό, ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη υπαρχόντων έργων και βάσεων δεδομένων που έχουν αναπτυχθεί με σκοπό την καταγραφή και ανάλυση γεγονότων διαμαρτυρίας σε μεγάλη κλίμακα. Η αξιολόγηση αυτών των συστημάτων αποκαλύπτει τόσο τις τεχνολογικές και μεθοδολογικές προσεγγίσεις που έχουν υιοθετηθεί, όσο και τα όρια και τις προκλήσεις που παραμένουν. Παρακάτω παρουσιάζονται ορισμένες από τις πιο αναγνωρισμένες βάσεις δεδομένων και ερευνητικές πρωτοβουλίες του πεδίου, με έμφαση στον τρόπο με τον οποίο συμβάλλουν στην κατανόηση της κοινωνικής κινητοποίησης και στηρίζουν ερευνητικά έργα σε διάφορους επιστημονικούς τομείς.

2.1 Global Database of Events, Language and Tone - GDELT

Η GDELT (Global Database of Events, Language and Tone) (Leetaru, 2013) αποτελεί μία από τις μεγαλύτερες και πιο ολοκληρωμένες βάσεις δεδομένων παγκοσμίως για την καταγραφή και ανάλυση γεγονότων, συγκρούσεων και κοινωνικοπολιτικών τάσεων. Συλλέγει ειδησεογραφικά δεδομένα από χιλιάδες πηγές σε περισσότερες από 100 γλώσσες, καλύπτοντας ολόκληρο τον κόσμο, συμπεριλαμβανομένων υποεθνικών τοποθεσιών μέσω γεωγραφικών συντεταγμένων (lat/long). Η ανάλυση επεκτείνεται σε κάθε χώρα και περιοχή του πλανήτη, επιτρέποντας μελέτες με γεωχωρική ακρίβεια.

Η κωδικοποίηση των γεγονότων βασίζεται στο πρότυπο CAMEO (Deborah J. Gerner, 2009)και πραγματοποιείται μέσω του αυτοματοποιημένου συστήματος TABARI (Schrodt, 2012), το οποίο εξάγει πληροφορίες από ειδησεογραφικά κείμενα. Παράλληλα, χρησιμοποιούνται τεχνικές μηχανικής μάθησης (Machine Learning, ML) για την ταξινόμηση και φιλτράρισμα των γεγονότων, ενισχύοντας την ακρίβεια και την ευελιξία του συστήματος.

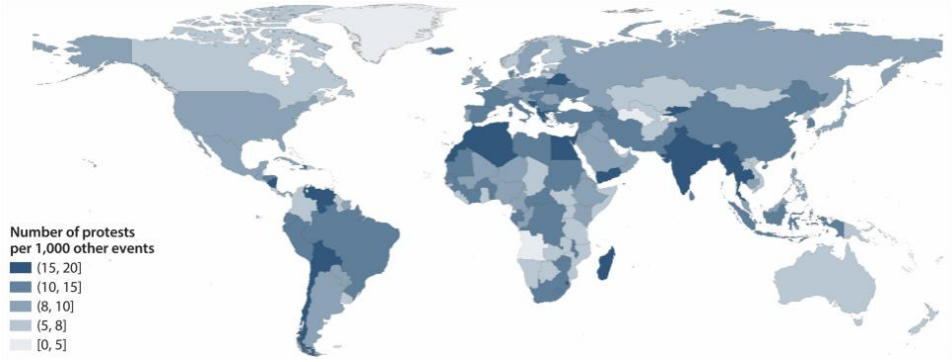
Η GDELT καταγράφει πληροφορίες όχι μόνο για τον τύπο του γεγονότος (π.χ. διαμαρτυρία, στρατιωτική επέμβαση, διπλωματική δήλωση), αλλά και για τους εμπλεκόμενους φορείς και δρώντες (actors), περιλαμβάνοντας ονόματα, χώρες προέλευσης και οργανωτική/πολιτική συσχέτιση (affiliation). Επιπλέον, αποθηκεύεται ο αριθμός των αναφορών ανά γεγονός, καθώς και μετρικές συναισθηματικής χροιάς (tone), προσφέροντας τη δυνατότητα παρακολούθησης της έντασης και του δημόσιου αντίκτυπου κάθε περιστατικού.

Τα ιστορικά δεδομένα εκκινούν από την 1η Ιανουαρίου 1979 και το σύστημα ενημερώνεται κάθε 15 λεπτά. Μόνο για το έτος 2015 καταγράφηκαν περίπου 750

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

δισεκατομμύρια συναισθηματικές στιγμές και πάνω από 1,5 δισεκατομμύριο αναφορές τοποθεσίας, με τα συνολικά του αρχεία να καλύπτουν περισσότερα από 215 χρόνια παγκόσμιας δραστηριότητας. Η GDELT ξεπερνά το παραδοσιακό επίκεντρο των δυτικών μέσων ενημέρωσης και επιδιώκει μια πραγματικά παγκόσμια οπτική, αποτυπώνοντας την πολυφωνία της ανθρώπινης κοινωνίας μέσα από τις ειδησεογραφικές της ροές.

Η εμβέλεια, η χρονική κάλυψη και η συνδυαστική προσέγγιση γεωχωρικών, θεματικών και συναισθηματικών δεδομένων καθιστούν τη GDELT ένα πρωτοποριακό εργαλείο για την έρευνα, την παρακολούθηση και την πρόβλεψη πολιτικών, κοινωνικών και ανθρωπιστικών εξελίξεων σε παγκόσμιο επίπεδο.



Εικόνα 1 Διαμαρτυρίες σε όλο τον κόσμο κατά την περίοδο 1980–2020, όπως καταγράφηκαν από τη βάση δεδομένων GDELT. Οι τιμές εκφράζουν τον αριθμό διαμαρτυριών ανά 1.000 καταγεγραμμένα γεγονότα. Η εικόνα απεικονίζει τον μέσο αριθμό διαμαρτυριών ανά 1.000 γεγονότα.

2.2 Armed Conflict Location & Event Data - ACLED

Το Armed Conflict Location & Event Data Project (ACLED, 2023) είναι ένας μη κερδοσκοπικός οργανισμός, ο οποίος ειδικεύεται στη συλλογή, ανάλυση και γεωχωρική αποτύπωση δεδομένων σε πραγματικό χρόνο που αφορούν την πολιτική βία και τις εκδηλώσεις διαμαρτυρίας παγκοσμίως. Ιδρύθηκε το 2005 ως ερευνητική πρωτοβουλία της πολιτικής επιστήμονα Clionadh Raleigh, και έχει εξελιχθεί σε μία από τις πιο αναγνωρισμένες πηγές για τη μελέτη συγκρούσεων και κοινωνικής αστάθειας.

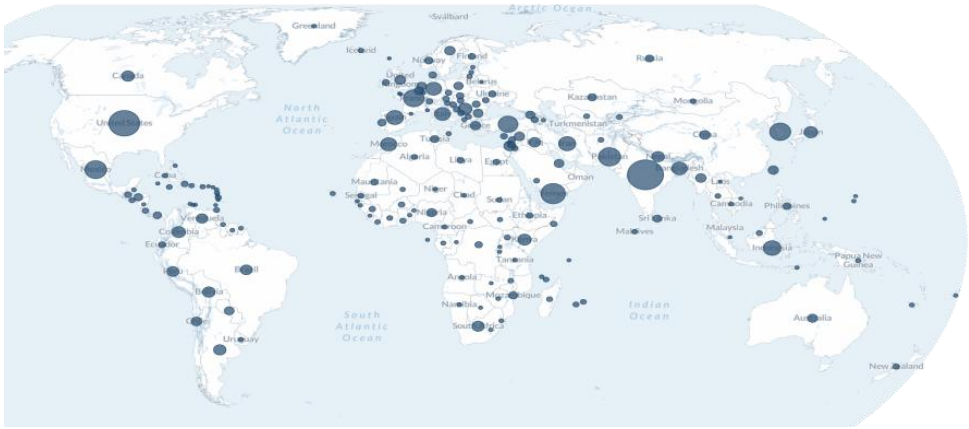
Κωδικοποιεί γεγονότα όπως ένοπλες συγκρούσεις, διαδηλώσεις, ταραχές, στρατηγικές εξελίξεις και απομακρυσμένες επιθέσεις**,** καταγράφοντας λεπτομερώς τη χρονολογία, τοποθεσία (συμπεριλαμβανομένων γεωγραφικών συντεταγμένων), τους εμπλεκόμενους δρώντες (ονόματα, χώρα, συσχέτιση), την ένταση του γεγονότος, και τυχόν θανάτους**.** Η συλλογή δεδομένων βασίζεται κυρίως σε ειδησεογραφικές πηγές**,** οι οποίες αξιολογούνται μέσω ανθρώπινης επιμέλειας (human review)**,** με εσωτερικό έλεγχο αξιοπιστίας (intracoder reliability) για την ενίσχυση της ακρίβειας.

Η γεωγραφική κάλυψη του ACLED ξεκίνησε από την Αφρική το 1997 και σταδιακά επεκτάθηκε: στη Μέση Ανατολή και τη Νοτιοανατολική Ασία από τα μέσα της δεκαετίας του 2010, στην Κεντρική Ασία, την Ανατολική Ευρώπη και τη Λατινική Αμερική από το 2018–2019, ενώ η παγκόσμια κάλυψη συμπληρώθηκε μετά το 2020. Η βάση δεδομένων παρέχει λεπτομερή χωρική πληροφόρηση, όχι μόνο σε εθνικό

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

αλλά και σε υποεθνικό επίπεδο, με γεωγραφική ακρίβεια μέσω συντεταγμένων γεωγραφικού πλάτους και μήκους (lat/long).

Μέχρι το 2022, η βάση δεδομένων περιλάμβανε πάνω από 1,3 εκατομμύρια καταγεγραμμένα γεγονότα, με τις ενημερώσεις να πραγματοποιούνται μέσω ενός παγκόσμιου δικτύου ερευνητών και τοπικών συνεργατών. Το ACLED συνεργάζεται με ακαδημαϊκά ιδρύματα, κυβερνήσεις, διεθνείς οργανισμούς και ανθρωπιστικές υπηρεσίες, ενώ τα δεδομένα του χρησιμοποιούνται ευρέως σε δημοσιογραφικά ρεπορτάζ, πολιτικές αναλύσεις, ερευνητικά έργα και συστήματα έγκαιρης προειδοποίησης, υποστηρίζοντας την τεκμηριωμένη λήψη αποφάσεων σε περιοχές με συγκρουσιακή δυναμική.



Εικόνα 2 Παγκόσμια γεωγραφική κατανομή διαδηλώσεων από τον Ιούλιο 2024 έως τον Ιούλιο 2025, βάσει δεδομένων ACLED. Καταγράφηκαν 155.594 γεγονότα, με κορυφαία συχνότητα στην Ινδία, τις ΗΠΑ και την Υεμένη. Το μέγεθος κάθε κύκλου αντιστοιχεί στη ένταση ανά περιοχή.

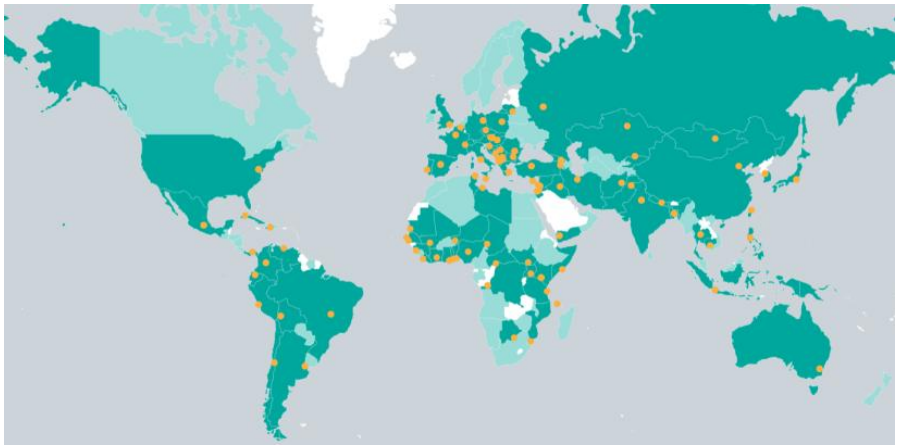
2.3 Carnegie Global Protest Tracker

Η Global Protest Tracker του Carnegie Endowment for International Peace (Peace)είναι μια διαδραστική βάση δεδομένων που καταγράφει και αναλύει σημαντικές αντικυβερνητικές διαδηλώσεις ανά τον κόσμο από το 2017 έως σήμερα. Εστιάζει αποκλειστικά σε αντικυβερνητικές κινητοποιήσεις, οι οποίες τεκμηριώνονται μέσα από έγκυρες αγγλόφωνες ειδησεογραφικές πηγές. Η κάλυψη είναι παγκόσμια, σε εθνικό επίπεδο, και η βάση επικαιροποιείται τακτικά με τη συνεισφορά ειδικών και πολιτικών αναλυτών.

Για κάθε διαμαρτυρία καταγράφονται βασικά χαρακτηριστικά, όπως η διάρκεια, ο εκτιμώμενος αριθμός συμμετεχόντων, η έκβαση (π.χ. αλλαγή πολιτικής ή ηγεσίας), οι βασικοί συμμετέχοντες (όπως εργατικά συνδικάτα, φοιτητές ή μειονοτικές ομάδες), καθώς και τα κίνητρα και οι αιτίες ενεργοποίησης της κινητοποίησης, οι οποίες παρατίθενται σε μορφή περιγραφικού κειμένου.

Η Global Protest Tracker απευθύνεται σε ερευνητές, φορείς χάραξης πολιτικής, δημοσιογράφους και την κοινωνία των πολιτών, προσφέροντας ένα πολύτιμο εργαλείο για την κατανόηση των πολιτικών, κοινωνικών και θεσμικών παραμέτρων που οδηγούν σε διαμαρτυρίες, καθώς και για τη μελέτη των επιπτώσεών τους στη διακυβέρνηση και τη σταθερότητα.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 3 Παγκόσμια απεικόνιση αντικυβερνητικών διαδηλώσεων, καταγεγραμμένων από το Carnegie Global Protest Tracker (2024–2025). Κάθε πορτοκαλί σημείο στον χάρτη αντιστοιχεί σε μία ξεχωριστή κινητοποίηση, ενώ η σκίαση δείχνει την ύπαρξη ενεργών γεγονότων.

2.4 Crowd Counting Consortium – CCC

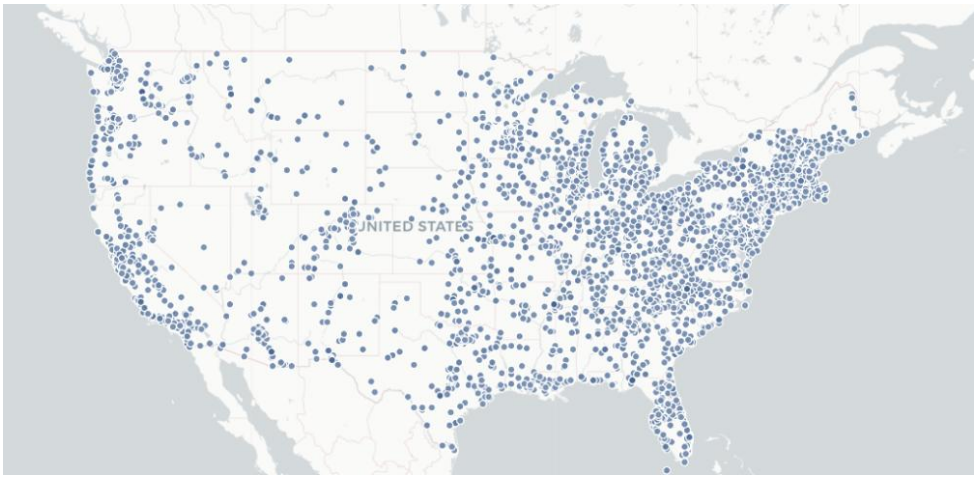
Το Crowd Counting Consortium (CCC) (Consortium) είναι ένα κοινό έργο του Harvard Kennedy School και του Πανεπιστημίου του Connecticut που συλλέγει δημόσια διαθέσιμα δεδομένα τα οποία περιλαμβάνουν πληροφορίες για πορείες, διαμαρτυρίες, απεργίες, διαδηλώσεις, ταραχές και άλλες μορφές κοινωνικών και πολιτικών δράσεων στις Ηνωμένες Πολιτείες.

Η πρωτοβουλία αυτή προέκυψε από μια συλλογική προσπάθεια για την ακριβή εκτίμηση του αριθμού των συμμετεχόντων στην Πορεία των Γυναικών στην Ουάσινγκτον. Αναγνωρίζοντας το αυξανόμενο δημόσιο ενδιαφέρον για αξιόπιστες και επικαιροποιημένες πληροφορίες σχετικά με τη συμμετοχή του κόσμου στις κοινωνικές κινητοποιήσεις και ανταποκρινόμενοι σε αιτήματα για συνέχιση της προσπάθειας πέρα από την Πορεία των Γυναικών, οι Pressman, Chenoweth και οι συνεργάτες τους ίδρυσαν το CCC.

Ο κύριος στόχος της πρωτοβουλίας είναι να δημοσιοποιεί συγκεντρωτικά δεδομένα για τον αριθμό των συμμετεχόντων σε κάθε εκδήλωση, προωθώντας την ανοιχτή πρόσβαση σε πληροφορίες σχετικά με τα πλήθη, προς όφελος του δημόσιου συμφέροντος και της ακαδημαϊκής έρευνας. Μέσω αυτής της συλλογής δεδομένων, το CCC συμβάλλει στην καλύτερη κατανόηση των κοινωνικών κινητοποιήσεων και στην αξιολόγηση της έντασης και της κλίμακας των δημόσιων εκδηλώσεων.

Πέρα από τον αριθμό συμμετεχόντων, το CCC τεκμηριώνει στοιχεία όπως η ημερομηνία και η τοποθεσία της εκδήλωσης, η θεματική στόχευση, το είδος της κινητοποίησης και, όταν είναι διαθέσιμα, πληροφορίες για το πλαίσιο και τα αιτήματα. Τα δεδομένα συλλέγονται μέσα από ειδησεογραφικές πηγές, δημόσια αρχεία και συνεισφορές πολιτών, ενώ δημοσιεύονται σε μορφή ανοιχτού συνόλου για χρήση από ερευνητές, δημοσιογράφους και ενδιαφερόμενους φορείς.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 4 Count Protest: Πόλεις των ΗΠΑ με τουλάχιστον μία αναφερόμενη διαμαρτυρία, 20 Ιανουαρίου 2017–31 Ιανουαρίου 2021. Κάθε κύκλος αντιστοιχεί σε μία πόλη. Τα πιο σκούρα σύμβολα δηλώνουν περισσότερες διαμαρτυρίες βάσει ειδησεογραφικών αναφορών.

2.5 Political Conflict and Democracy - PolDem

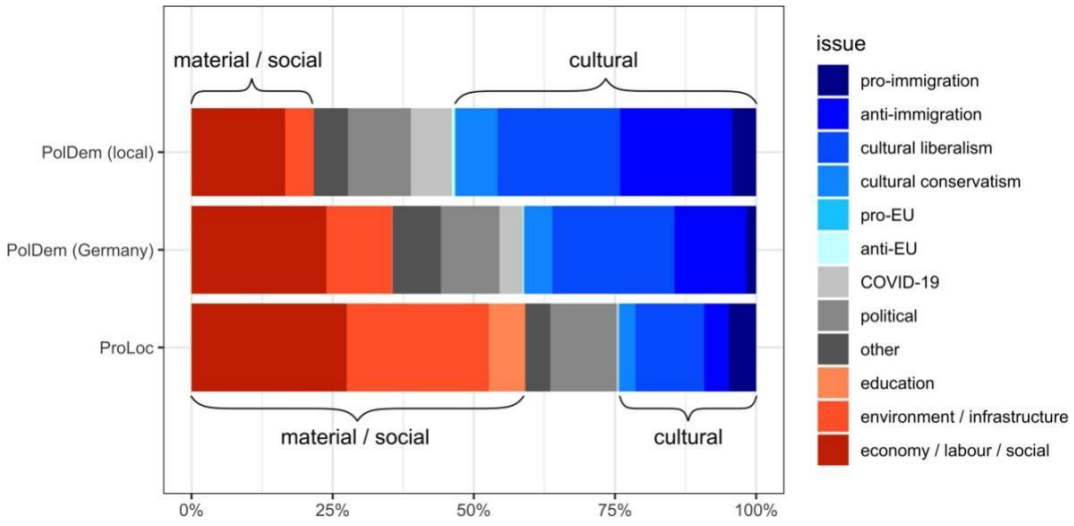
Ο φορέας Political Conflict and Democracy (PolDem) PolDem (Kriesi H. W., 2020) είναι μια συνεργατική πρωτοβουλία που δημιουργήθηκε για να προσφέρει αξιόπιστη πληροφόρηση για τις κοινωνικές και πολιτικές κινητοποιήσεις στην Ευρώπη. Ως συνέχεια των έργων των Hanspeter Kriesi, Edgar Grande και Swen Hutter (Kriesi H. W., 2020), ο PolDem απάντησε στην αυξανόμενη ανάγκη της έρευνας για συστηματικά και επικαιροποιημένα δεδομένα σχετικά με πορείες, διαδηλώσεις και, ειδικότερα, απεργίες. Κύριος στόχος του είναι η δημόσια τεκμηρίωση και διάθεση εναρμονισμένων στοιχείων για κάθε εκδήλωση συλλογικής δράσης, ώστε να υποστηρίζονται τόσο το δημόσιο συμφέρον όσο και η ακαδημαϊκή μελέτη της αμφιλεγόμενης πολιτικής.

Στο πλαίσιο αυτό, το σύνολο poldemprotest_30 καταγράφει γεγονότα διαμαρτυρίας σε 30 ευρωπαϊκές χώρες κατά την περίοδο 2000–2015, συγκροτώντας περίπου 31.000 εγγραφές. Για κάθε γεγονός αποτυπώνονται με ενιαίο σχήμα πεδίων η ημερομηνία και ο τόπος διεξαγωγής, η μορφή δράσης, οι φορείς που συμμετέχουν, ο στόχος ή αποδέκτης της διαμαρτυρίας, το περιεχόμενο/αίτημα και, όπου είναι διαθέσιμο, η κλίμακα συμμετοχής και δείκτες έντασης (π.χ. αστυνομική παρουσία, συλλήψεις, τραυματισμοί). Τα δεδομένα συγκεντρώνονται από δημοσίως διαθέσιμες ειδησεογραφικές πηγές και παράγονται μέσω υβριδικής διαδικασίας: μια εκτεταμένη δέσμη εργαλείων Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing, NLP) προεπιλέγει σχετικό υλικό από εκατομμύρια ειδησεογραφικά δελτία, ενώ εκπαιδευμένοι κωδικοποιητές πραγματοποιούν την τελική, λεπτομερή κωδικοποίηση. Για λόγους γλωσσικής ομοιογένειας και συγκρισιμότητας, η συλλογή εδράζεται σε δέκα αγγλόφωνα ειδησεογραφικά πρακτορεία που καλύπτουν τις ΕΕ-27 και τρεις επιπλέον χώρες (Ισλανδία, Νορβηγία, Ελβετία).

Όπως και στο Crowd Counting Consortium, ο PolDem επιδιώκει να καταστήσει ορατή την κλίμακα και την ποικιλία της συλλογικής δράσης, προσφέροντας διαχρονικά και διακρατικά συγκρίσιμες μετρήσεις. Η εστίαση στις απεργίες επιτρέπει την ανάλυση της συχνότητας και της έντασής τους ανά χώρα και κλάδο, τη σύνδεσή τους με θεσμικές ή οικονομικές μεταβολές και τη μελέτη των ρεπερτορίων δράσης

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

και των εκβάσεων. Με την ισορροπία που πετυχαίνει ανάμεσα σε μεγάλη κάλυψη και αυστηρή κωδικοποίηση, ο PolDem λειτουργεί ως σημείο αναφοράς για την εμπειρική τεκμηρίωση της αμφιλεγόμενης πολιτικής στην Ευρώπη (Kriesi H. W., 2020).



Εικόνα 5 Σύγκριση θεμάτων διαμαρτυριών της βάσης δεδομένων PolDem με την ProLoc.

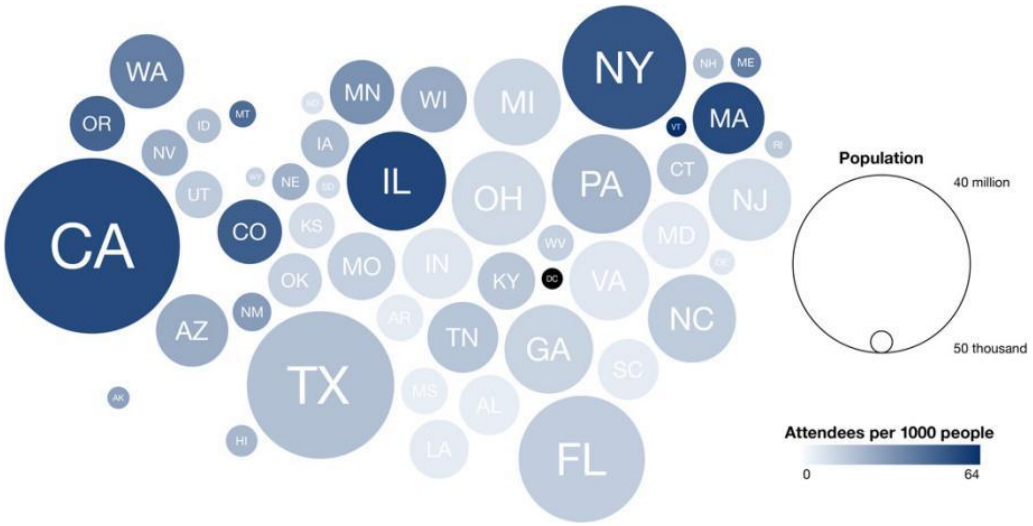
Η Εικόνα 5, αποτελεί αναφορά στην μελέτη των Daphi et al. (Daphi, 2024), κατά την οποία έγινε σύγκριση της βάσης δεδομέων Poldem με την ProLoc. Συγκρίνονται δεδομένα διαμαρτυριών από τη Λειψία, την Δρέσδη, τη Βρέμη και την Στουτγάρδη για την PolDem(local) και Proloc και με την PolDem για την Γερμανία. Με αυτόν τον τρόπο, γίνεται φανερή η διαφορετική εστίαση της κάθε βάσης.

2.6 Count Love

Η Count Love (Dana R. Fisher, 2019) είναι ανοικτή βάση δεδομένων που καταγράφει διαμαρτυρίες και διαδηλώσεις στις ΗΠΑ από το 2017 και μετά, με στόχο την τεκμηρίωση της έκτασης και της θεματολογίας της πολιτικής συμμετοχής. Καταχωρεί εκδηλώσεις όπως διαδηλώσεις, πορείες και απεργίες και για καθεμιά αποθηκεύει βασικά στοιχεία όπως τον αριθμό συμμετεχόντων, την ημερομηνία και τα αίτια/θεματικές. Τα δεδομένα είναι ελεύθερα προσβάσιμα και επιτρέπουν την ανάλυση τάσεων και τη μελέτη της εξέλιξης των κοινωνικών κινητοποιήσεων στον χρόνο.

Η τροφοδότηση της βάσης γίνεται από επιλεγμένο κατάλογο δημόσια διαθέσιμων τοπικών μέσων (εφημερίδες, ραδιόφωνα, τηλεοπτικοί σταθμοί) ανά πολιτεία, τον οποίο η Count Love συντηρεί και επεκτείνει. Ένας αυτόματος συλλέκτης (web crawler) επισκέπτεται τακτικά ειδησεογραφικούς ιστότοπους, ανασύρει σχετικά άρθρα και, με τη βοήθεια εργαλείων NLP και ML, φιλτράρει άσχετο περιεχόμενο, ομαδοποιεί παρόμοιες αναφορές και εντοπίζει διπλότυπα. Στη συνέχεια, εκπαιδευμένοι ερευνητές ελέγχουν και οριστικοποιούν τις εγγραφές. Τα αποτελέσματα δημοσιεύονται περιοδικά με μορφή CSV- Comma Separated Values και ως αναζητήσιμοι χάρτες και γραφήματα στο countlove.org.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 6 Count Love: Χάρτης που απεικονίζει τις διαμαρτυρίες ανά πολιτεία. Οι πολιτείες έχουν μέγεθος ανάλογο με τον πληθυσμό τους. Κάθε πολιτεία έχει χρώμα ανάλογο με τον αριθμό συμμετεχόντων στις διαμαρτυρίες. Βασίζεται σε δεδομένα που συλλέχθηκαν από τις 20 Ιανουαρίου 2017 έως τις 28 Δεκεμβρίου 2018.

Συνολικά, η μελέτη των πληροφοριακών συστημάτων που καταγράφουν και αναλύουν ειδησεογραφικά γεγονότα σε πραγματικό χρόνο υπογραμμίζει τη σημασία της ακριβούς, έγκαιρης και τεκμηριωμένης συλλογής δεδομένων για την κατανόηση των κοινωνικών και πολιτικών δυναμικών. Πλατφόρμες όπως το GDELT, το ACLED, το Carnegie Global Protest Tracker, το Crowd Counting Consortium, το PolDem και το Count Love αποδεικνύουν ότι οι τεχνολογικές εξελίξεις και οι σύγχρονες μεθοδολογίες, από την αυτοματοποιημένη εξαγωγή δεδομένων έως την ανθρώπινη επιμέλεια, μπορούν να συνδυαστούν αποτελεσματικά ώστε να παράγουν υψηλής ποιότητας πληροφορία.

Η αξιοποίηση γεωχωρικών δεδομένων, τεχνικών NLP, ML και αυτοματοποιημένων web crawlers, ενισχύει τη δυνατότητα παρακολούθησης κινητοποιήσεων, συγκρούσεων και κοινωνικών τάσεων σε παγκόσμια κλίμακα. Παράλληλα, η διαθεσιμότητα αυτών των δεδομένων σε ερευνητές, οργανισμούς, κυβερνήσεις και δημοσιογράφους δημιουργεί νέα περιθώρια για ανάλυση, πρόβλεψη και τεκμηριωμένη λήψη αποφάσεων.

Ωστόσο, η σύγκριση των διαφορετικών έργων αναδεικνύει σημαντικές προκλήσεις, καθώς υπάρχουν διαφοροποιήσεις στην κάλυψη, στην ακρίβεια των δεδομένων, στη μεθοδολογία συλλογής και στην εστίαση κάθε βάσης. Άλλες βάσεις δίνουν έμφαση στη γεωγραφική ανάλυση, άλλες στην ένταση ή στα αίτια των διαμαρτυριών. Επιπλέον, ζητήματα όπως η αξιοπιστία των πηγών, η γλωσσική πολυπλοκότητα και η ανάγκη ενοποίησης διαφορετικών προτύπων κωδικοποίησης παραμένουν ανοικτά.

Με αυτόν τον τρόπο, τελικά, τα πληροφοριακά συστήματα καταγραφής ειδησεογραφικών γεγονότων σε πραγματικό χρόνο αποτελούν αναπόσπαστο εργαλείο για την κατανόηση της σύγχρονης κοινωνικής και πολιτικής πραγματικότητας. Η συνεχής εξέλιξη των τεχνολογιών ανάλυσης δεδομένων και η διεύρυνση των διαθέσιμων πηγών υπόσχονται ακόμη μεγαλύτερη ακρίβεια και αξιοπιστία στο μέλλον, ενισχύοντας τόσο την επιστημονική έρευνα όσο και την κοινωνική κατανόηση των φαινομένων συλλογικής δράσης.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

3. Τεχνολογικό Υπόβαθρο της Πτυχιακής Εργασίας

3.1 Web Scrapping

Η απόκτηση πληροφοριών από ιστοσελίδες αποτελεί ένα κρίσιμο στάδιο σε έργα που βασίζονται στην επεξεργασία και ανάλυση μεγάλων κειμενικών όγκων δεδομένων. Ο όρος web scraping αναφέρεται στη διαδικασία αυτόματης εξαγωγής δεδομένων από ιστοσελίδες, με χρήση εξειδικευμένων εργαλείων ή αλγορίθμων, ώστε μη δομημένο περιεχόμενο, όπως άρθρα, αναρτήσεις σε κοινωνικά δίκτυα ή δεδομένα από πλατφόρμες δημόσιας πληροφόρησης, να μετατρέπεται σε δομημένη μορφή κατάλληλη για περαιτέρω ανάλυση.

Το web scraping έχει εξελιχθεί σε βασικό εργαλείο της επιστήμης δεδομένων, καθώς επιτρέπει την αποδοτική συλλογή μεγάλων ποσοτήτων πληροφορίας από το διαδίκτυο, αυτοματοποιώντας μια διαδικασία που διαφορετικά θα απαιτούσε χρονοβόρα χειροκίνητη αναζήτηση. Στην παρούσα ενότητα παρουσιάζονται οι βασικές αρχές, τα εργαλεία και τα στάδια της διαδικασίας του web scraping, καθώς και η τεχνολογική του σημασία στο πλαίσιο της ανάλυσης δεδομένων μεγάλης κλίμακας.

3.1.1 Ιστορική Εξέλιξη και Σημασία

Οι πρώτες ευρέως αναγνωρισμένες εφαρμογές της τεχνολογίας web scraping αναπτύχθηκαν από μηχανικούς μηχανών αναζήτησης, με χαρακτηριστικότερο παράδειγμα την Google (Khder, 2021)Τα προγράμματα scraping που χρησιμοποιούν οι μηχανές αυτές, οι οποίες εξερευνούν συνεχώς τον παγκόσμιο ιστό σαρώνοντας ιστοσελίδες, εξάγοντας δομημένες πληροφορίες και δημιουργώντας ένα εκτεταμένο ευρετήριο περιεχομένου. Με αυτόν τον τρόπο, διευκολύνεται η ταχεία και ακριβής ανάκτηση των πιο σχετικών αποτελεσμάτων για τους χρήστες.

3.1.2 Διαδικασία Web Scraping

Αν και συχνά συγχέονται, το web scraping δεν ταυτίζεται με το data mining (εξόρυξη δεδομένων). Η κύρια διαφορά έγκειται στο αντικείμενο κάθε διαδικασίας. Το web scraping επικεντρώνεται στη συλλογή και οργάνωση δεδομένων από το διαδίκτυο, ενώ το data mining εστιάζει στην ανάλυση υπαρχόντων δεδομένων, με στόχο την εξαγωγή προτύπων και γνώσης, αξιοποιώντας στατιστικές και υπολογιστικές τεχνικές.

Αναλυτικότερα, το web scraping αναφέρεται στη διαδικασία αυτόματης εξαγωγής δομημένων δεδομένων από ιστοσελίδες, με τη χρήση εξειδικευμένου λογισμικού. Στόχος αυτής της διαδικασίας είναι η συλλογή, ο καθαρισμός και η αποθήκευση πληροφοριών σε υπολογιστικά φύλλα ή βάσεις δεδομένων, με σκοπό να μπορεί να γίνει περαιτέρω ανάλυση. Πρόκειται για μια τεχνολογία που καθιστά δυνατή τη μαζική συλλογή και δομημένη οργάνωση πληροφοριών, ακόμη και σε πραγματικό χρόνο (Khder, 2021).

Η απόκτηση δεδομένων αποτελεί το πρώτο και θεμελιώδες στάδιο σε κάθε μελέτη ή έργο στον χώρο της επιστήμης δεδομένων. Τα απαραίτητα δεδομένα μπορούν να προέρχονται από ιδιωτικές πηγές, όπως εσωτερικά εταιρικά αρχεία και οικονομικές

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

αναφορές, από δημόσιες πηγές, όπως ιστότοποι, επιστημονικά περιοδικά και πλατφόρμες ανοικτών δεδομένων, καθώς και από εμπορικές βάσεις δεδομένων, που παρέχονται έναντι αμοιβής από τρίτους παρόχους.

Στο πλαίσιο της διαδικτυακής συλλογής δεδομένων, τρεις βασικές και αλληλένδετες διεργασίες διαδραματίζουν κρίσιμο ρόλο:

1. Web parsing (Ανάλυση ιστοσελίδων): Εξαγωγή δομημένων δεδομένων από HTML περιεχόμενο.

2. Web crawling (Ανίχνευση ιστοσελίδων): Αυτόματη πλοήγηση και ανακάλυψη νέων σελίδων μέσων συνδέσμων.

3. Οργάνωση δεδομένων: Προεπεξεργασία, καθαρισμός, αποθήκευση και προετοιμασία των συλλεγμένων πληροφοριών για περαιτέρω επεξεργασία.

Η εφαρμογή του web scraping επιτρέπει την εξαγωγή επιπλέον λεπτομερειών από ιστοσελίδες, οι οποίες ενδέχεται να μην είναι άμεσα προσβάσιμες μέσω απλών, μη αυτόματων μεθόδων. Έρευνες έχουν δείξει ότι η χρήση αυτοματοποιημένων εργαλείων scraping μπορεί να οδηγήσει σε σημαντικά πιο ακριβή, συνεπή και λεπτομερή δεδομένα σε σύγκριση με τη χειροκίνητη συλλογή, καθιστώντας το scraping αναπόσπαστο εργαλείο στην εποχή της πληροφορίας.

Η υλοποίηση ενός αποδοτικού συστήματος web scraping απαιτεί τη συνδυαστική χρήση τεχνολογιών, όπως η ανίχνευση (spidering) και η αντιστοίχιση προτύπων (pattern matching) (Khder, 2021). Παρότι η τεχνική υλοποίηση ενδέχεται να περιλαμβάνει πολύπλοκα στάδια, η ευρεία διαθεσιμότητα έτοιμων εργαλείων και βιβλιοθηκών το καθιστά προσβάσιμο ακόμη και σε μη εξειδικευμένους χρήστες. Ένα βασικό χαρακτηριστικό των περισσότερων εργαλείων scraping είναι η δυνατότητα αποστολής παραμετροποιημένων HTTP αιτημάτων, με προσαρμοσμένες επικεφαλίδες και περιεχόμενο, διευκολύνοντας την πρόσβαση σε δυναμικά ή προστατευμένα δεδομένα.

Σε τεχνικό επίπεδο, η συλλογή πραγματοποιείται μέσω αιτημάτων HTTP/HTTPS προς διακομιστές ιστού και αυτοματοποιείται με web crawlers, επιτρέποντας συστηματική, χαμηλού κόστους συγκέντρωση μεγάλου όγκου δεδομένων.

Η προεπεξεργασία και ο καθαρισμός των δεδομένων αποτελούν απαραίτητο βήμα πριν από οποιαδήποτε ανάλυση, καθώς απομακρύνουν τον θόρυβο, διορθώνουν λάθη και μετατρέπουν τα δεδομένα σε κατάλληλη μορφή για περαιτέρω επεξεργασία ή για χρήση σε μοντέλα ML.

3.1.3 Παραδείγματα μεθόδων συλλογής δεδομένων

Ενδεικτικές μέθοδοι συλλογής δεδομένων είναι τα παρακάτω (Sirisuriya, 2015):

- Παραδοσιακή τεχνική αντιγραφής και επικόλλησης: Η πιο απλή και προσιτή μέθοδος web scraping, κατάλληλη για μικρές ποσότητες δεδομένων. Παρότι είναι πρακτική σε περιορισμένες περιπτώσεις, είναι ιδιαίτερα χρονοβόρα και

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

μη αποδοτική για μεγάλα σύνολα πληροφοριών ή για επαναλαμβανόμενες εργασίες.

- Προγραμματισμός με HTTP: Περιλαμβάνει την αποστολή HTTP αιτημάτων προς ιστοσελίδες (στατικές ή δυναμικές) με τη χρήση socket programming ή έτοιμων βιβλιοθηκών, επιτρέποντας την αυτοματοποίηση και την ακριβή συλλογή περιεχομένου.
- Λογισμικό Web Scraping: Σύγχρονα εργαλεία και βιβλιοθήκες διευκολύνουν τη δημιουργία προσαρμοσμένων λύσεων. Ενδεικτικά παραδείγματα αποτελούν τα BeautifulSoup, Scrapy, Selenium, τα οποία υποστηρίζονται από περιβάλλον Python, καθώς και εφαρμογές με γραφικό περιβάλλον όπως είναι τα Octoparse και ParseHub.

3.1.4 Στάδια της διαδικασίας Web Scraping

Η διαδικασία Web Scraping περιλαμβάνει μια σειρά από στάδια, τα οποία συνδυάζονται για την αυτόματη συλλογή, ανάλυση και επεξεργασία δεδομένων από ιστοσελίδες. Τα στάδια αυτά περιλαμβάνουν την ανάκτηση του περιεχομένου της ιστοσελίδας, την εξαγωγή των σχετικών δεδομένων και, στη συνέχεια, την περαιτέρω ανάλυση ή αποθήκευσή τους για μελλοντική χρήση. Παρακάτω παρουσιάζονται τα βασικά βήματα της διαδικασίας (Persson, 2019).

1. Fetching Stage (Στάδιο Ανάκτησης)

   Το πρώτο βήμα είναι η αποστολή ενός HTTP GET αιτήματος στη διεύθυνση-στόχο (URL), για την ανάκτηση του HTML περιεχομένου της σελίδας. Χρησιμοποιούνται εργαλεία για την αποστολή αιτημάτων HTTP μέσω γραμμής εντολών ή προγραμματιστικά και για λήψη ολόκληρων ιστοσελίδων ή αρχείων από το Διαδίκτυο.

2. Extraction Stage (Στάδιο Εξαγωγής)

   Ακολουθεί η εξαγωγή των χρήσιμων δεδομένων από το HTML περιεχόμενο, με τεχνικές όπως Κανονικές εκφράσεις (Regular Expressions), βιβλιοθήκες ανάλυσης HTML, όπως BeautifulSoup και ερωτήματα XPath.

Συνοψίζοντας, η διαδικασία Web Scraping στηρίζεται σε μια ακολουθία βημάτων που ξεκινούν με την ανάκτηση του περιεχομένου της ιστοσελίδας και συνεχίζουν με την εξαγωγή και αξιοποίηση των δεδομένων. Η σωστή υλοποίηση κάθε σταδίου εξασφαλίζει την αποτελεσματικότητα της διαδικασίας και την αξιοπιστία των αποτελεσμάτων που προκύπτουν.

3.1.5 Νομικό Πλαίσιο Web Crawling και Χρήση του Robots.txt

Η συλλογή δεδομένων μέσω τεχνικών web crawling και web scraping αποτελεί μία διαδεδομένη πρακτική τόσο στην ακαδημαϊκή έρευνα όσο και στη βιομηχανία. Ωστόσο, η νομιμότητά της εξαρτάται από ένα σύνολο τεχνικών, ηθικών και νομικών κανόνων, οι οποίοι συχνά αφήνουν περιθώρια ασάφειας στην εφαρμογή τους.

Το βασικό σημείο αναφοράς είναι το αρχείο robots.txt, το οποίο εισήχθη το 1994 στο πλαίσιο του Robots Exclusion Protocol (REP). Το αρχείο αυτό τοποθετείται στη ρίζα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

ενός ιστότοπου και καθορίζει, μέσω απλών οδηγιών (Allow / Disallow), ποια τμήματα είναι προσβάσιμα από αυτοματοποιημένα προγράμματα (crawlers) και ποια όχι. Παρά τη χρησιμότητά του, το robots.txt δεν έχει νομικά δεσμευτική ισχύ από μόνο του, λειτουργεί ως ένα πρότυπο που ρυθμίζει τις σχέσεις μεταξύ διαχειριστών ιστοτόπων και αυτοματοποιημένων λογισμικών πρακτόρων (bots). Στη βιβλιογραφία έχει υποστηριχθεί ότι αποτελεί κυρίως ένα ηθικό πλαίσιο συμμόρφωσης (Yang Sun, 2010) και όχι νομικά εκτελεστό μηχανισμό (Yu, 2023).

Σε ευρωπαϊκό επίπεδο, η Οδηγία (ΕΕ) 2019/790 για τα πνευματικά δικαιώματα στην ψηφιακή ενιαία αγορά εισάγει σημαντικές εξαιρέσεις για το Text and Data Mining (TDM)**:**

- Το άρθρο 3 επιτρέπει την εξόρυξη δεδομένων από ερευνητικούς φορείς χωρίς άδεια, για επιστημονικούς σκοπούς.
- Το άρθρο 4 επιτρέπει το TDM σε κάθε φορέα, υπό την προϋπόθεση ότι ο κάτοχος δικαιωμάτων δεν έχει ρητά αποκλείσει τη χρήση μέσω τεχνικών μέσων (όπως το robots.txt).

Επομένως, η συμμόρφωση προς το robots.txt, αλλά και η τήρηση των Όρων Χρήσης κάθε ιστότοπου, αποτελούν προϋποθέσεις για τη νόμιμη άντληση δεδομένων.

Συνολικά, το web scraping αναδεικνύεται ως κρίσιμο εργαλείο στην εποχή της πληροφορίας, επιτρέποντας την αυτοματοποιημένη συλλογή, οργάνωση και προεπεξεργασία δεδομένων μεγάλης κλίμακας. Με μεθόδους όπως η ανάλυση ιστοσελίδων (web parsing), η συστηματική περιήγηση (web crawling) και η αναγνώριση προτύπων (pattern matching), καθίσταται εφικτή η μετατροπή μη δομημένου διαδικτυακού περιεχομένου σε δομημένα σύνολα δεδομένων κατάλληλα για περαιτέρω ανάλυση. Παράλληλα, η τήρηση των κατευθυντήριων γραμμών του robots.txt και των όρων χρήσης περιορίζει νομικούς και ηθικούς κινδύνους, χωρίς να θέτει σε κίνδυνο την πληρότητα του συλλεγόμενου υλικού.

3.2 Αναπαράσταση Κειμένου με Διανύσματα

Στην NLP, η ανάγκη για μετατροπή του κειμένου σε αριθμητική μορφή προέκυψε από τα πρώτα κιόλας βήματα της ML, όταν οι υπολογιστές έπρεπε να κατανοήσουν δεδομένα σε μορφή κειμένου. Οι πρώτες μέθοδοι, όπως το Bag of Words (BoW) και η Συχνότητα Όρων – Αντίστροφη Συχνότητα Εγγράφων (Term Frequency – Inverse Document Frequency, TF-IDF), βασίζονταν στην καταμέτρηση όρων και παρείχαν απλές, αλλά περιορισμένες, αναπαραστάσεις. Με την πρόοδο της έρευνας, αναπτύχθηκαν πιο προηγμένες τεχνικές, που αναπαριστούν τις λέξεις ως διανύσματα σε έναν πολυδιάστατο χώρο. Σε αυτήν τη μορφή, η απόσταση και η κατεύθυνση μεταξύ των διανυσμάτων αντικατοπτρίζουν σημασιολογικές και συντακτικές σχέσεις, επιτρέποντας στα μοντέλα να κατανοούν ομοιότητες και συμφραζόμενα. Αυτή η προσέγγιση έχει βελτιώσει σημαντικά την απόδοση σε πλήθος εφαρμογών, όπως η ταξινόμηση κειμένων, η ανάκτηση και εξόρυξη πληροφοριών, με χαρακτηριστικά παραδείγματα μοντέλων τα Word2Vec (Word to Vector), GloVe (Global Vectors for Word Representation) και fastText.

Στη συνέχεια, θα παρουσιαστούν οι κυριότερες τεχνικές μετατροπής κειμένου σε διανυσματική μορφή, ξεκινώντας από τις πιο κλασικές, όπως το BoW και το TF-IDF,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

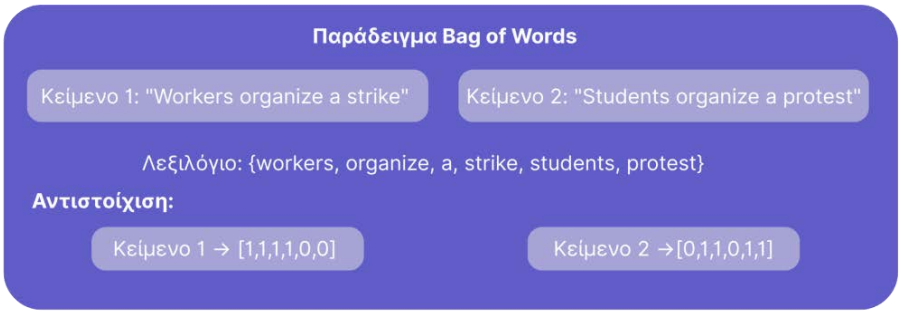
και προχωρώντας σε πιο σύγχρονες προσεγγίσεις, όπως τα Word Embeddings. Κάθε μέθοδος θα αναλυθεί ως προς την αρχή λειτουργίας της, τα πλεονεκτήματα και τους περιορισμούς της, καθώς και τις βασικές εφαρμογές της.

3.2.1 Bag of Words (BoW)

Η μέθοδος BoW αποτελεί μία από τις παλαιότερες και πιο διαδεδομένες τεχνικές αναπαράστασης κειμένων στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Name Entity Recognition, NER). Χρησιμοποιείται ευρέως σε διάφορες εφαρμογές της όπως η ταξινόμηση κειμένων, η ανάλυση συναισθήματος, καθώς και σε άλλους κλάδους όπως η αναγνώριση εικόνων και η ανάλυση βίντεο.

Στην περίπτωση της ταξινόμησης κειμένων, η μέθοδος BoW αναπαριστά κάθε έγγραφο ως ένα διάνυσμα χαρακτηριστικών, όπου κάθε διάσταση αντιστοιχεί σε μία μοναδική λέξη του λεξιλογίου. Η τιμή της διάστασης προκύπτει είτε ως συχνότητα εμφάνισης της λέξης στο έγγραφο είτε ως δυαδική τιμή (0/1) η οποία δηλώνει την παρουσία ή την απουσία της. Σε πρακτικά πειράματα, το μοντέλο BoW χρησιμοποιείται βασικά ως εργαλείο για τη δημιουργία και την ταξινόμηση χαρακτηριστικών. Διάφορα μέτρα κατηγοριοποίησης κειμένου μπορούν να δημιουργηθούν μετασχηματίζοντας το κείμενο χρησιμοποιώντας το BoW. Ο αριθμός των εμφανίσεων μιας λέξης ή ενός όρου σε ένα κείμενο είναι γνωστός ως συχνότητα εμφάνισης όρου, η οποία θα χρησιμοποιηθεί για τον προσδιορισμό της κατηγορίας του κειμένου ή απλά για την ταξινόμηση του κειμένου. Η μέθοδος αγνοεί πλήρως τη σειρά των λέξεων, τη γραμματική δομή και τα συμφραζόμενα, εστιάζοντας αποκλειστικά στη στατιστική καταμέτρηση (Qader, 2019).

Η Εικόνα 7, αποτελεί παράδειγμα μετατροπής κειμένου σε διάνυσμα με την μέθοδο BoW. Χρησιμοποιούνται δύο κείμενα(Κείμενο 1, Κείμενο 2), και η ένωση των λέξεων αποτελεί το λεξιλόγιο. Για κάθε λέξη του λεξιλογίου ελέγχεται αν εμφανίζεται στο κείμενο, αν υπάρχει λαμβάνει την τιμή 1, αλλιώς την τιμή 0. Με αυτόν τον τρόπο, κάθε κείμενο αναπαρίσταται ως διάνυσμα παρουσίας/απουσίας λέξεων.



Εικόνα 7 Παράδειγμα BoW. Κάθε λέξη του λεξιλογίου αντιστοιχεί σε μία διάσταση, και το διάνυσμα κάθε εγγράφου δείχνει την παρουσία ή απουσία των λέξεων σε αυτό.

Πλεονεκτήματα:

- Απλότητα υλοποίησης: Εύκολη στην κατανόηση και την εφαρμογή, ακόμη και σε βασικό επίπεδο.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- Αποτελεσματικότητα: Ιδιαίτερα χρήσιμη σε μικρά σύνολα δεδομένων και σε απλές εφαρμογές ταξινόμησης.

Μειονεκτήματα:

- Απώλεια πληροφορίας: Αγνοεί τη σειρά των λέξεων και τα σημασιολογικά συμφραζόμενα.
- Υψηλή διαστατικότητα: Δημιουργεί μεγάλες και αραιές διανυσματικές αναπαραστάσεις (sparse vectors), ιδιαίτερα όταν το λεξιλόγιο είναι εκτεταμένο.
- Έλλειψη σημασιολογικής κατανόησης: Δεν αναγνωρίζει σχέσεις μεταξύ λέξεων· για παράδειγμα, οι λέξεις "strike" και "protest" αντιμετωπίζονται ως πλήρως άσχετες, παρόλο που ανήκουν στην ίδια εννοιολογική κατηγορία.

3.2.2 Term Frequency – Inverse Document Frequency (TF-IDF)

Η μέθοδος TF-IDF έχει σχεδιαστεί για να επισημαίνει λέξεις που είναι συχνές σε ένα συγκεκριμένο έγγραφο, αλλά σχετικά σπάνιες στο συνολικό σώμα κειμένων. Με αυτόν τον τρόπο αναδεικνύει όρους με υψηλή διακριτική ικανότητα στο έγγραφο. Οι βασικές εφαρμογές της περιλαμβάνουν ανάκτηση πληροφοριών, κατάταξη εγγράφων, σύνοψη κειμένων και εξόρυξη κειμένων.

Αν και πρόκειται για μια παραδοσιακή μέθοδο αναπαράστασης κειμένου που βασίζεται στη συχνότητα, το TF-IDF δεν θεωρείται word embedding, καθώς δεν αποτυπώνει τις σημασιολογικές σχέσεις ή το συμφραζόμενο των λέξεων. Αντίθετα, κάθε λέξη αντιμετωπίζεται ως ανεξάρτητη διάσταση σε έναν αραιό διανυσματικό χώρο, χωρίς να υπάρχει κατανόηση της ομοιότητας μεταξύ των όρων. Οι εφαρμογές του TF-IDF περιλαμβάνουν την ανάκτηση πληροφοριών, την κατάταξη εγγράφων, τη σύνοψη κειμένων και την εξόρυξη κειμένων (Manning, 2008).

Η ομοιότητα TF-IDF για έναν όρο σε ένα έγγραφο υπολογίζεται με βάση τον ακόλουθο τύπο:

Για έναν όρο t σε ένα έγγραφο d μέσα σε ένα σύνολο εγγράφων D:

$$\text{TF IDF}(t,d,D) \quad \text{TF}(t,d) \quad \text{IDF}(t,D)$$

TF(t, d), Term Frequency, συχνότητα εμφάνισης του όρου t στο έγγραφο d

TF (Term Frequency)**:**

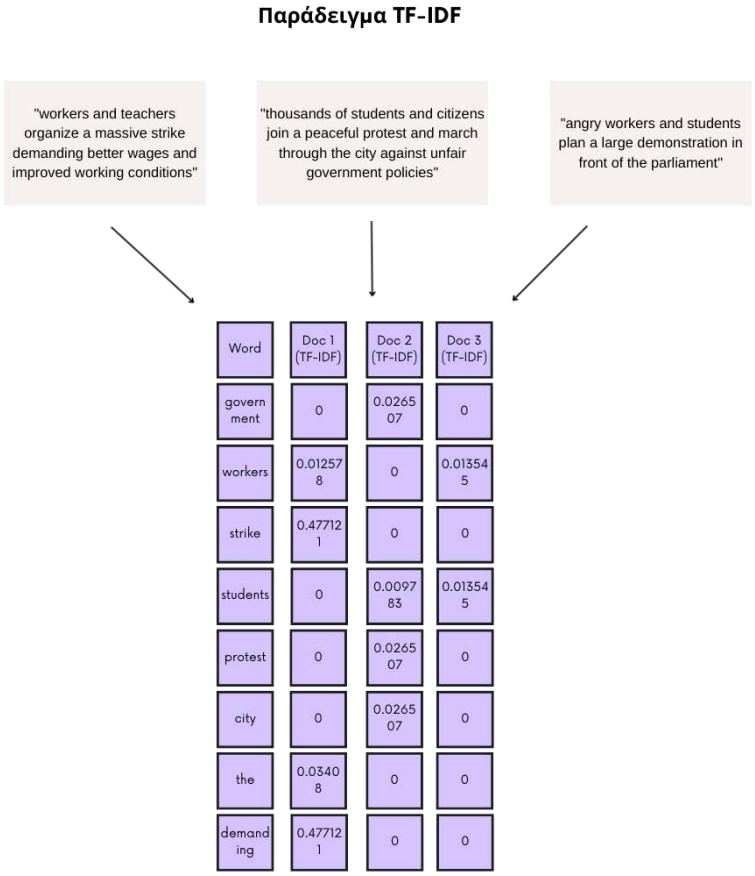
$$\text{TF}(t,d) \quad \frac{\text{Αριθμός εμφανίσεων του όρου t στο έγγραφο d}}{\text{Συνολικός αριθμός λέξεων στο έγγραφο d}}$$

IDF (Inverse Document Frequency)**:**

$$\text{IDF}(t,D) \quad \log \frac{|D|}{1 \quad |\{d \in D : t \in d\}|}$$

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

όπου |D| συνολικός αριθμός εγγράφων, και |{d∈D: t ∈d}| αριθμός εγγράφων που περιέχουν τον όρο t.



Εικόνα 8 Ανάλυση TF-IDF για τρία κείμενα σχετικά με διαδηλώσεις και εργατικά αιτήματα.

Στην Εικόνα 8 παρουσιάζεται η εφαρμογή του υπολογισμού TF-IDF σε τρία διαφορετικά κείμενα, τα οποία αφορούν διαμαρτυρίες και αιτήματα εργαζομένων και φοιτητών. Τα αποτελέσματα δείχνουν τη σχετική σημασία των λέξεων στο κάθε κείμενο, με τους όρους "workers", "strike" και "demanding", να εμφανίζουν υψηλότερες τιμές TF-IDF, γεγονός που υποδηλώνει ότι είναι πιο χαρακτηριστικοί για το περιεχόμενο. Αυτό μπορεί να χρησιμεύσει για την κατανόηση των πιο αντιπροσωπευτικών λέξεων για το κάθε κείμενο, κάτι που είναι χρήσιμο σε εφαρμογές όπως η κατηγοριοποίηση κειμένων ή η ανάλυση θεμάτων.

Πλεονεκτήματα:

- Απλότητα και ευκολία υλοποίησης: Απαιτεί μόνο καταμέτρηση όρων και εφαρμογή λογαρίθμου.
- Αποτελεσματικότητα: Παρουσιάζει καλή απόδοση σε κλασικά προβλήματα ανάκτησης πληροφοριών.
- Υψηλή ταχύτητα: Μπορεί να εφαρμοστεί αποδοτικά ακόμα και σε μεγάλα σώματα κειμένων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- Αναγνώριση λέξεων-κλειδιών: Εντοπίζει αυτόματα σημαντικούς όρους σε κάθε έγγραφο.

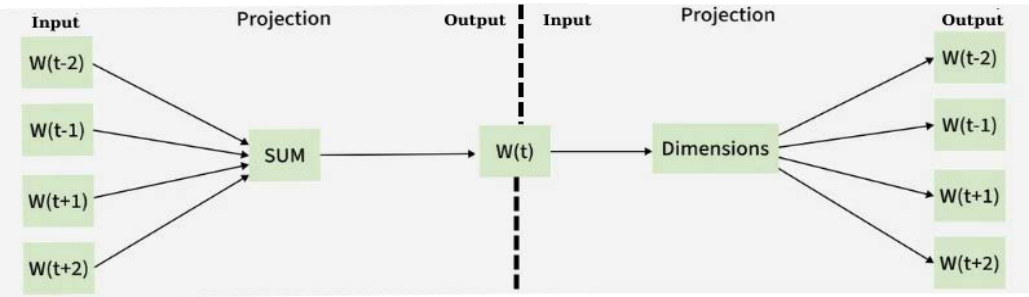
Μειονεκτήματα:

- Έλλειψη σημασιολογικής κατανόησης: Αγνοεί τη σχέση μεταξύ λέξεων και τα συμφραζόμενα.
- Υψηλή διαστατικότητα: Παράγει αραιές διανυσματικές αναπαραστάσεις (sparse vectors) σε μεγάλα λεξιλόγια.
- Αδυναμία διαχείρισης συνωνυμίας και πολυσημίας: Δεν αναγνωρίζει ότι διαφορετικές λέξεις μπορεί να έχουν το ίδιο νόημα ή ότι μία λέξη μπορεί να έχει πολλαπλές έννοιες.
- Απώλεια δομικής πληροφορίας: Δεν λαμβάνει υπόψη τη σειρά των λέξεων και τη συντακτική δομή του κειμένου.

3.2.3 Word to Vector (Word2Vec)

Το Word2Vec είναι μια πρωτοποριακή τεχνική εκμάθησης διανυσματικών αναπαραστάσεων λέξεων (word embeddings), η οποία αναπτύχθηκε το 2013 από ερευνητική ομάδα της Google, υπό την καθοδήγηση του Tomas Mikolov. Στόχος της μεθόδου είναι να μετατρέπει τις λέξεις σε πυκνά διανύσματα χαμηλής διάστασης που αποτυπώνουν τόσο σημασιολογικές όσο και συντακτικές σχέσεις μεταξύ τους. Σε αντίθεση με παραδοσιακές τεχνικές, όπως το BoW ή το TF-IDF, που δημιουργούν αραιές αναπαραστάσεις χωρίς να λαμβάνουν υπόψη τη σημασία και τα συμφραζόμενα των λέξεων, το Word2Vec μαθαίνει εννοιολογικά πλούσιες αναπαραστάσεις που επιτρέπουν τη μέτρηση ομοιοτήτων και τη συσχέτιση όρων με βάση το περιεχόμενο.

3.2.3.1 Αρχιτεκτονικές Word2Vec



Εικόνα 9 Αρχιτεκτονική Word2Vec: Στο πρώτο μισό φαίνεται το CBOW και στο άλλο μισό το Skip-gram.

Η διαδικασία του Word2Vec, όπως απεικονίζεται και στην Εικόνα *9*, αποτελείται από δύο κύρια μοντέλα για τη δημιουργία διανυσματικών αναπαραστάσεων το CBOW και το Skip-gram.

Αρχικά, το CBOW προβλέπει τη λέξη-στόχο χρησιμοποιώντας τις γειτονικές λέξεις μέσα σε ένα προκαθορισμένο παράθυρο κειμένου (context window). Με άλλα λόγια, το μοντέλο μαθαίνει να μαντεύει ποια λέξη ταιριάζει καλύτερα σε ένα συγκεκριμένο συμφραζόμενο.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η είσοδος στο μοντέλο είναι τα διανύσματα των γειτονικών λέξεων. Ο αλγόριθμος υπολογίζει το μέσο όρο αυτών των διανυσμάτων προβλέποντας τη λέξη-στόχο. Η κύρια υπόθεση του μοντέλου είναι ότι η σειρά των λέξεων στο παράθυρο δεν έχει σημασία, αλλά μόνο η παρουσία τους επηρεάζει την πρόβλεψη.

Υπολογίζεται από τον εξής τύπο:

$$P(w_t \mid w_{t\,c},\cdots,w_{t\,1}, w_{t\ 1},\cdots, w_{t\ c})P(w_t,\mid w_{t\,c},\cdots,w_{t\,1},w_{t\ 1},\cdots,w_{t\ c}) = \frac{\exp\left(u_{w_t}^{T} h_t\right)}{\sum_{w\in V}\exp\left(u_{w_t}^{T} h_t\right)}$$

Όπου $h_t = \frac{1}{2c}\sum_{j=-c,j\neq 0}^{c} v_{w_{t+j}}$

Σύμβολα**:**

c: μέγεθος παραθύρου.

V: λεξιλόγιο.

v_w: input embedding της λέξης w.

u_w: output embedding της λέξης w.

Το CBOW είναι υπολογιστικά αποδοτικό και λειτουργεί πολύ καλά σε μεγάλα σύνολα δεδομένων.

Δευτερόν, το Skip-gram είναι η αντίστροφη προσέγγιση του CBOW. Εδώ το μοντέλο λαμβάνει ως είσοδο τη λέξη-στόχο και προσπαθεί να προβλέψει τις γειτονικές λέξεις μέσα στο παράθυρο συμφραζομένων.

Η είσοδος είναι ένα μοναδικό embedding λέξης. Το μοντέλο παράγει πιθανότητες για πολλαπλές λέξεις-πλαίσιο και δημιουργεί πολλά ζεύγη εκπαίδευσης για κάθε λέξη, κάτι που βοηθά στην καλύτερη εκμάθηση σπάνιων λέξεων.

Υπολογίζεται από τον τύπο**:**

$$\frac{1}{2c}\ \sum_{c,\ 0}^{c}\ P(w_t\quad w_t)$$

Εστιάζει στην εκμάθηση καλών αναπαραστάσεων για σπάνιες λέξεις, καθώς δημιουργεί πολλά ζεύγη εκπαίδευσης για κάθε λέξη.

Πλεονεκτήματα:

- Πυκνές διανυσματικές αναπαραστάσεις: Παράγει embeddings χαμηλής διάστασης που συλλαμβάνουν σχέσεις μεταξύ λέξεων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- Σημασιολογικές και συντακτικές συσχετίσεις: Λέξεις με παρόμοια συμφραζόμενα βρίσκονται κοντά στον διανυσματικό χώρο (π.χ. *king - man + woman ≈ queen*) .
- Αποδοτικότητα: Πολύ πιο συμπαγής και αποδοτική αναπαράσταση σε σχέση με τεχνικές όπως το TF-IDF.
- Ευκολία υπολογισμού ομοιότητας: Υποστηρίζει μετρικές όπως το cosine similarity για σύγκριση λέξεων.
- Κλιμακωσιμότητα: Λειτουργεί αποτελεσματικά ακόμα και σε πολύ μεγάλα σώματα κειμένων.

Μειονεκτήματα:

- Μία αναπαράσταση ανά λέξη: Δεν διαφοροποιεί το embedding μιας λέξης ανάλογα με το συμφραζόμενο.
- Αδυναμία σύλληψης σημασιολογικών αποχρώσεων: Δεν αποτυπώνει διαφορετικά νοήματα της ίδιας λέξης (π.χ. bank τράπεζα, όχθη).
- Έλλειψη context-awareness: Προβλήματα που λύνουν πιο σύγχρονες τεχνικές.
- Απαιτεί μεγάλα σύνολα δεδομένων: Η ποιότητα των embeddings εξαρτάται έντονα από το μέγεθος και την ποικιλία του corpus.

3.2.4 Global Vectors for Word Representation (GloVe)

Το Glove είναι ένα μοντέλο εκμάθησης διανυσματικών αναπαραστάσεων λέξεων, το οποίο αξιοποιεί τις παγκόσμιες στατιστικές συν-εμφάνισης λέξεων (global word co-occurrence statistics) μέσα σε ένα σώμα κειμένου. Η μέθοδος εισήχθη το 2014 από τους Jeffrey Pennington, Richard Socher και Christopher D. Manning στο Πανεπιστήμιο Stanford και αποτελεί μία από τις σημαντικότερες εναλλακτικές του Word2Vec (Pennington, 2014).

Σε αντίθεση με το Word2Vec, το οποίο βασίζεται αποκλειστικά στο τοπικό πλαίσιο συμφραζομένων (context window), το GloVe αξιοποιεί ολόκληρη την κατανομή συχνοτήτων των ζευγών λέξεων σε ολόκληρο το corpus. Με τον τρόπο αυτό, το GloVe μαθαίνει αναπαραστάσεις που συλλαμβάνουν σημασιολογικές συσχετίσεις οι οποίες δεν θα μπορούσαν να αναγνωριστούν μόνο με τοπική πληροφόρηση.

Με απλά λόγια, το GloVe χρησιμοποιεί πληροφορίες σχετικά με το πόσο συχνά δύο λέξεις συν-εμφανίζονται σε οποιοδήποτε σημείο ενός μεγάλου συνόλου κειμένων, ώστε να δημιουργήσει ενσωματώσεις λέξεων που αντικατοπτρίζουν τις σημασιολογικές σχέσεις τους. Αυτό του επιτρέπει να συλλαμβάνει πιο περίπλοκες εννοιολογικές συνδέσεις, πέρα από αυτές που καταγράφουν τα μοντέλα που στηρίζονται μόνο σε τοπικά συμφραζόμενα.

Πιθανότητα και λόγος	**k = solid**	**k = gas**	**k = water**	**k = fashion**
P(k|ice)	$1.9 \quad 10^{-4}$	$6.6 \quad 10^{5}$	$3.0 \quad 10^{3}$	$1.7 \quad 10^{5}$
P(k|steam)	$2.2 \quad 10^{5}$	$7.8 \quad 10^{4}$	$2.2 \quad 10^{3}$	$1.8 \quad 10^{5}$
$\frac{P(k\|ice)}{P(k\|steam)}$	8.64	$8.5 \quad 10^{2}$	1.36	0.96

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Πίνακας 1 Πιθανότητες συν-εμφάνισης για τις λέξεις-στόχους ice και steam με επιλεγμένες λέξεις συμφραζομένων από ένα σώμα κειμένων 6 δισεκατομμυρίων tokens. Μόνο στον λόγο (ratio) εξαλείφεται ο θόρυβος από μη διακριτικές λέξεις, όπως water και fashion.

Ο Πίνακας 1 χρησιμοποιεί το μοντέλο Glove, για να δείξει πως η αναλογία πιθανοτήτων συν-εμφάνισης μεταξύ των λέξεων ice, steam με τις λέξει solid, gas, water και fashion αποκαλύπτει σημασιολογικές σχέσεις. Οι λέξεις solid και ice φαίνονται να είναι πολύ κοντά, σε σχέση με την λέξη steam, ενώ η λέξη gas φαίνεται να σχετίζεται με την λέξη steam. Αντίθετα, οι λέξεις water και fashion έχουν τιμή κοντά στην μονάδα, δηλαδή δεν βρίσκονται διανυσματικά κοντά.

Το GloVe χρησιμοποιεί έναν πίνακα συν-εμφάνισης X, όπου κάθε στοιχείο εκφράζει τον αριθμό των φορών που η λέξη j εμφανίζεται στο πλαίσιο της λέξης i. Ο στόχος είναι να εκπαιδευτεί ένα σύνολο διανυσμάτων λέξεων που ελαχιστοποιούν το σφάλμα πρόβλεψης μέσω της βελτιστοποίησης μιας συνάρτησης κόστους (loss function):

$$J \quad \sum_{i,\ 1}^{V} f(X_i\ )\ (w_i^T \tilde{w} \quad b_i\ \ \tilde{b}\ \ \log X_i\ )^2$$

όπου:

w , $\tilde{w}$: διανύσματα λέξεων για τη λέξη i και τη λέξη-πλαίσιο j

, $\tilde{}$: bias terms

f( ): συνάρτηση βαρύτητας για τον περιορισμό της επίδρασης πολύ συχνών/σπάνιων ζευγών

V: μέγεθος λεξιλογίου

Πλεονεκτήματα:

- Συνδυάζει τοπικές και παγκόσμιες πληροφορίες: Αξιοποιεί τις στατιστικές συν-εμφάνισης σε ολόκληρο το corpus.
- Υπολογιστική αποδοτικότητα: Χρησιμοποιεί παραγοντοποίηση πινάκων, η οποία μπορεί να υπολογιστεί offline, μειώνοντας τον χρόνο εκπαίδευσης.
- Καλή απόδοση σε αναλογίες λέξεων: Παράγει ποιοτικά embeddings που αποτυπώνουν σχέσεις όπως: *King - Man + Woman ≈ Queen.*
- Μειωμένη ευαισθησία σε θόρυβο: Η χρήση λόγων πιθανοτήτων ελαχιστοποιεί την επίδραση άσχετων ή μη διακριτικών λέξεων.

Μειονεκτήματα:

- Μία αναπαράσταση ανά λέξη: Δεν μπορεί να χειριστεί την πολυσημία (π.χ. bank τράπεζα / όχθη).
- Υψηλές απαιτήσεις δεδομένων: Απαιτεί μεγάλα corpora (μεγάλα οργανωμένα σύνολα κειμένων) για αξιόπιστη εκτίμηση των στατιστικών συν-εμφάνισης.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- Αυξημένες απαιτήσεις μνήμης: Ο πίνακας συν-εμφάνισης X μπορεί να είναι τεράστιος για μεγάλα λεξιλόγια.
- Περιορισμοί στις λέξεις εκτός λεξιλογίου (OOV): Για την εισαγωγή νέων λέξεων απαιτείται επανεκπαίδευση του μοντέλου.

3.2.5 FastText

Το FastText είναι ένα μοντέλο εκμάθησης διανυσματικών αναπαραστάσεων λέξεων και εργαλείο ταξινόμησης κειμένων, το οποίο αναπτύχθηκε από το Facebook AI Research (FAIR) και κυκλοφόρησε το 2016 ως λογισμικό ανοιχτού κώδικα (T. Yao, 2020). Σχεδιάστηκε για να προσφέρει αποτελεσματική εκπαίδευση σε μεγάλα σώματα κειμένων, αντιμετωπίζοντας το πρόβλημα του αυξημένου χρόνου εκπαίδευσης που παρουσίαζαν προηγούμενα μοντέλα όπως το Word2Vec και το GloVe.

Η αρχιτεκτονική του FastText διαφοροποιείται από τις παραδοσιακές τεχνικές, καθώς ενσωματώνει χαρακτηριστικά n-gram χαρακτήρων, επιτρέποντας στο μοντέλο να συλλαμβάνει πληροφορίες σχετικά με τη μορφολογία των λέξεων και τη σειρά τους μέσα σε μια πρόταση. Αυτό το καθιστά ιδιαίτερα χρήσιμο για γλώσσες με πλούσια μορφολογία και μεγάλα λεξιλόγια.

Σε αντίθεση με το παραδοσιακό μοντέλο BoW, το FastText δεν αναπαριστά απλώς κάθε λέξη ως μοναδικό διάνυσμα, αλλά διασπά τις λέξεις σε n-grams χαρακτήρων. Έτσι, η τελική αναπαράσταση της λέξης προκύπτει από το άθροισμα ή τον μέσο όρο των αναπαραστάσεων όλων των n-grams που την αποτελούν. Για παράδειγμα, για την πρόταση «I like play basketball», η είσοδος του παραδοσιακού μοντέλου σάκου λέξεων είναι «I», «like», «play» και «basketball», ενώ το fastText, στη βάση του μοντέλου σάκου λέξεων, προσθέτει επιπλέον χαρακτηριστικά n-gram. Αν n 2, το επιπρόσθετο χαρακτηριστικό είναι η μέση τιμή των λέξεων που αντιστοιχούν στα «I likeι», «like play» και «play basketball». Με αυτόν τον τρόπο, το FastText λαμβάνει υπόψη τη σειρά των λέξεων σε κάποιο βαθμό, επιτυγχάνοντας πιο ακριβή αναπαράσταση προτάσεων, κάτι που είναι δύσκολο να επιτευχθεί με κλασικά μοντέλα BoW.

Πλεονεκτήματα:

- Διαχείριση λέξεων εκτός λεξιλογίου (OOV): Επειδή το FastText αναλύει τις λέξεις σε n-grams χαρακτήρων, μπορεί να δημιουργήσει embeddings ακόμη και για άγνωστες λέξεις.
- Σύλληψη μορφολογικών χαρακτηριστικών: Τα n-grams βοηθούν στην καλύτερη κατανόηση καταλήξεων, προθεμάτων και παραγώγων λέξεων, βελτιώνοντας την απόδοση σε γλώσσες με πολύπλοκη μορφολογία.
- Αντιμετώπιση παραλλαγών λέξεων: Παράγει παρόμοια embeddings για συναφείς λέξεις, π.χ. protest, protesting, protester.
- Ταχύτερη εκπαίδευση: Η χρήση ιεραρχικής softmax επιτρέπει σημαντική μείωση του υπολογιστικού κόστους σε μεγάλα corpora.
- Υψηλή απόδοση στην ταξινόμηση κειμένων: Έχει καλύτερα αποτελέσματα σε πολλές εργασίες NLP, ειδικά όταν υπάρχουν περιορισμένα δεδομένα.

Μειονεκτήματα:

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- Μοναδική αναπαράσταση ανά λέξη: Όπως και το Word2Vec και το GloVe, το FastText παράγει ένα embedding ανά λέξη, χωρίς να λαμβάνει υπόψη το πλήρες συμφραζόμενο.
- Εξάρτηση από μεγάλα corpora: Αν και πιο αποδοτικό από άλλα μοντέλα, εξακολουθεί να απαιτεί εκτεταμένα σύνολα δεδομένων για ποιοτικά αποτελέσματα.
- Περιορισμοί σε πολύ σπάνιες λέξεις: Παρά τα n-grams, η αναπαράσταση μπορεί να μην είναι ακριβής όταν οι λέξεις εμφανίζονται ελάχιστες φορές.
- Υποδεέστερο από context-aware μοντέλα: Μοντέλα όπως το ELMo και το BERT ξεπερνούν το FastText σε εργασίες που απαιτούν κατανόηση πολυσημίας και προσαρμογή στο συμφραζόμενο.

Συνοψίζοντας, η αναπαράσταση κειμένου σε διανυσματική μορφή αποτελεί τον θεμέλιο λίθο για όλες τις σύγχρονες εφαρμογές NER. Οι παραδοσιακές μέθοδοι, όπως το BoW και το TF-IDF, έθεσαν τις βάσεις παρέχοντας έναν απλό και κατανοητό τρόπο μετατροπής του κειμένου σε αριθμητική μορφή. Ωστόσο, η αδυναμία τους να συλλάβουν τη σημασιολογία και τα συμφραζόμενα οδήγησε στην ανάπτυξη πιο εξελιγμένων τεχνικών, όπως τα Word2Vec και GloVe, που επιτρέπουν στα μοντέλα να κατανοούν σχέσεις, ομοιότητες και αναλογίες μεταξύ λέξεων.

Η μετάβαση από τις αραιές και υψηλής διάστασης αναπαραστάσεις στις πυκνές και σημασιολογικά πλούσιες ενσωματώσεις λέξεων άνοιξε τον δρόμο για μια νέα γενιά εφαρμογών, από την αυτόματη μετάφραση μέχρι την ανάλυση συναισθήματος και τα συστήματα προτάσεων. Παρά τους περιορισμούς τους, όπως η αδυναμία διαχείρισης πολυσημίας, οι τεχνικές αυτές αποτελούν αναπόσπαστο στάδιο σε κάθε pipeline ανάλυσης κειμένου και παραμένουν βασικό αντικείμενο μελέτης και βελτίωσης.

Η εξέλιξη από BoW, TF-IDF σε Word2Vec, GloVe και fastText δείχνει πώς η NLP πέρασε από απλές καταμετρήσεις λέξεων σε σημασιολογικά πλούσιες αναπαραστάσεις. Στο πλαίσιο της παρούσας πτυχιακής, οι τεχνικές αυτές είναι κρίσιμες για την κατανόηση του λεξιλογίου που σχετίζεται με απεργίες και διαδηλώσεις σε όλη την Ευρώπη, καθώς επιτρέπουν την αναγνώριση θεματικών συσχετίσεων και προτύπων που θα ήταν αδύνατο να ανιχνευθούν με κλασικές μεθόδους.

3.3 Βαθιά Μάθηση

Η Βαθιά Μάθηση (Deep Learning, DL) αποτελεί υποκλάδο του ML και επιτρέπει στα υπολογιστικά συστήματα να μαθαίνουν απευθείας από δεδομένα, μέσω νευρωνικών δικτύων με πολλαπλά επίπεδα. Η προσέγγιση αυτή, εμπνευσμένη από αφηρημένα γνωσιακά μοντέλα, μειώνει δραστικά την ανάγκη για ρητό feature engineering, καθώς το δίκτυο μαθαίνει αυτόματα ποια χαρακτηριστικά είναι σημαντικά για την εκάστοτε εργασία, οδηγώντας σε πιο αποδοτική και ακριβή ανάλυση. Το DL έχει συμβάλει καθοριστικά στην πρόοδο της Τεχνητής Νοημοσύνης (Artificial Intelligence – AI), με εφαρμογές στο NLP, την αναγνώριση εικόνων και φωνής, την αυτόνομη οδήγηση, τα συστήματα προτάσεων και τη βιοπληροφορική. Η ικανότητά της να χειρίζεται μεγάλα και πολύπλοκα δεδομένα την καθιστά βασικό εργαλείο για τη μελέτη και την αντιμετώπιση σύγχρονων προβλημάτων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

3.3.1 Αρχιτεκτονικές Βαθιών Νευρωνικών Δικτύων

Κύριες οικογένειες αρχιτεκτονικών είναι τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs), τα επαναλαμβανόμενα δίκτυα (Recurrenr Neural Networks, RNNs/ Long short-term memory, LSTM/Gate Recurrent Unit, GRU), τα γραφικά νευρωνικά δίκτυα (Graph Neural Networks, GNNs) και οι αρχιτεκτονικές προσοχής τύπου Transformer. Σε αντίθεση με τις παραδοσιακές μη νευρωνικές προσεγγίσεις, οι σύγχρονες νευρωνικές αρχιτεκτονικές εξάγουν κατανεμημένες αναπαραστάσεις απευθείας από τα δεδομένα εισόδου. Παρά την επιτυχία τους, τα βαθιά δίκτυα διαθέτουν μεγάλο αριθμό παραμέτρων και, χωρίς επαρκή δεδομένα, είναι επιρρεπή σε υπερπροσαρμογή και σε μειωμένη ικανότητα γενίκευσης. Η πρακτική αντιμετώπιση περιλαμβάνει τεχνικές καταστολής υπερπροσαρμογής (dropout, weight decay, early stopping), ενίσχυση δεδομένων (data augmentation) και —όλο και συχνότερα— αξιοποίηση προεκπαιδευμένων μοντέλων.

3.3.2 Προεκπαιδευμένα Μοντέλα (Pre-Trained Models - PTMs) και Transfer Learning

Τα μεγάλης κλίμακας προεκπαιδευμένα μοντέλα (PTMs), όπως το BERT (Bidirectional Encoder Representations from Transformers) και το GPT (Generative Pre-Trained Transformer), έχουν καθιερωθεί αποτελώντας ορόσημο στον τομέα του AI. Χάρη στους εξελιγμένους στόχους προεκπαίδευσης και τον τεράστιο αριθμό παραμέτρων, τα μεγάλης κλίμακας PTMs μπορούν να αποτυπώσουν αποτελεσματικά τη γνώση από τεράστιες ποσότητες επισημασμένων και μη επισημασμένων δεδομένων. Η γνώση αυτή αποθηκεύεται σε πολυάριθμες παραμέτρους, προσαρμόζοντάς' τες σε συγκεκριμένες εργασίες, η πλούσια γνώση που κωδικοποιείται σε αυτές τις παραμέτρους μπορεί να ωφελήσει μια ποικιλία εργασιών, κάτι που έχει αποδειχθεί εκτενώς μέσω πειραματικής επαλήθευσης και εμπειρικής ανάλυσης. Είναι πλέον κοινή πεποίθηση της κοινότητας της AI να υιοθετούνται τα PTMs ως βάση για εργασίες αντί να εκπαιδεύονται μοντέλα από μηδενική βάση (X. Han, 2021).

Ένα καθοριστικό βήμα για την εκπαίδευση με λίγα δεδομένα αποτελεί το Transfer Learning. Αντί η εκπαίδευση να ξεκινά από το μηδέν, το μοντέλο αξιοποιεί προϋπάρχουσα γνώση ώστε να προσαρμόζεται σε νέες εργασίες με ελάχιστα δείγματα. Το πλαίσιο αυτό αποτελείται από δύο φάσεις: (i) προεκπαίδευση: Το μοντέλο εκπαιδεύεται αρχικά σε μία ή περισσότερες εργασίες-πηγές (ή σε έναν μεγάλο γενικό σύνολο δεδομένων) ώστε να μάθει γενικές αναπαραστάσεις και χαρακτηριστικά του προβλήματος. (ii) Fine-tuning (Λεπτομερής ρύθμιση): Στη συνέχεια, το προεκπαιδευμένο μοντέλο προσαρμόζεται στην εργασία-στόχο, μεταφέροντας και εξειδικεύοντας τη γνώση που απέκτησε στην πρώτη φάση. Χάρη σε αυτή την πλούσια γνώση από το στάδιο της προεκπαίδευσης, το μοντέλο μπορεί να επιτύχει υψηλή απόδοση ακόμη και με πολύ λίγα δείγματα δεδομένων στην εργασία-στόχο.

Με την εισαγωγή του Transformer (Vaswani, 2017), τα προεκπαιδευμένα γλωσσικά μοντέλα πέρασαν σε νέα φάση. Καταστήθηκε δυνατή η εκπαίδευση βαθύτερων και πιο εκφραστικών μοντέλων σε σύγκριση με τα συμβατικά CNNs και RNNs. Σε αντίθεση με τις στατικές αναπαραστάσεις λέξεων (π.χ. word2vec, GloVe) που χρησιμοποιούνταν ως χαρακτηριστικά εισόδου, τα προεκπαιδευμένα μοντέλα τύπου

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Transformer (π.χ. BERT, GPT) μπορούν να λειτουργήσουν ως βασικά μοντέλα για εξειδικευμένες εργασίες. Αφού προεκπαιδευτούν σε μεγάλα γλωσσικά σύνολα δεδομένων, η αρχιτεκτονική και οι εκπαιδευμένες παράμετροί τους αξιοποιούνται ως αρχικοποίηση (initialization) και προσαρμόζονται μέσω fine-tuning σε συγκεκριμένες εφαρμογές της NER, επιτυγχάνοντας ανταγωνιστικά αποτελέσματα ακόμη και με λίγα δεδομένα. Μέχρι σήμερα, τα προεκπαιδευμένα μοντέλα βασισμένα σε Transformer έχουν επιτύχει κορυφαίες επιδόσεις σε πληθώρα εργασιών NLP, πέρα από τα BERT και GPT, χαρακτηριστικά παραδείγματα είναι τα RoBERTa, Albert, BART και T5. Το Transformer είναι μια μη επαναλαμβανόμενη αρχιτεκτονική ακολουθίας προς ακολουθία (Sequence-to-Sequence, Seq2Seq) που αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή.

3.3.2.1 Generative Pre-Trained Transformer (GPT)

Το GPT (X. Han, 2021) βασίζεται στην αρχιτεκτονική του Trasformer και η βασική ιδέα στηρίζεται στην χρήση self attention μηχανισμών που επεξεργάζονται τις λέξεις σχετικά με όλες τις υπόλοιπες λέξεις σε μια πρόταση, σε αντίθεση με αλλές μεθόδους που τις επεξεργάζονται γραμμικά. Η εκπαίδευσή του γίνεται με σκοπό την πρόβλεψη του επόμενου token πάνω σε μεγάλα κειμενικά δεδομένα, και στη συνέχεια μπορεί να προσαρμοστεί διακριτικά σε επιμέρους εργασίες. Εμπειρικά, έχει επιδείξει ισχυρές επιδόσεις σε ευρύ φάσμα εργασιών NLP — από συμπερασματολογία και απάντηση ερωτήσεων έως σημασιολογική σύγκριση και κατηγοριοποίηση.

3.3.2.2 BERT

Η εμφάνιση του BERT (Devlin, 2019) προώθησε σημαντικά τον χώρο των PTMs. Σε αντίθεση με το GPT, το BERT χρησιμοποιεί πολυεπίπεδο αμφίδρομο κωδικοποιητή Transformer και προεκπαιδεύεται με την χρήση μασκών σε ένα μέρος των λέξεων σε κάθε ακολουθία εισόδου και το μοντέλο εκπαιδεύεται με στόχο να προβλέπει τις αρχικές τιμές αυτών των λέξεων βάση, η διαδικασία αυτή ονομάζεται Masked Language Modeling (MLM). Επιπλέον γίνεται πρόβλεψη αν οι παραγώμενες προτάσεις έχουν συνοχή με τις προηγούμενς, αυτή η διαδικασίας ονομάζεται Next Sentence Prediction (NSP). Συνολικά, το MLM βοηθά το BERT να κατανοήσει το πλαίσιο μίας πρότασης, ενώ το NSP βοηθά στο να κατανοήσει την σύνδεση μεταξύ ζευγών προτάσεων.

3.3.2.3 RoBERTa και ALBERT

Μετά το GPT και το BERT προτάθηκαν βελτιώσεις. Το RoBERTa (Robustly Optimized BERT Approach) (Liu Y. O., 2019) αποτελεί παραλλαγή του BERT κατά την οποία καταργεί την διαδικασία του NSP και εκπαιδεύεται για περισσότερα βήματα, με μεγαλύτερα batches και περισσότερα δεδομένα, χρησιμοποιεί εκτενέστερες προτάσεις εκπαίδευσης στις οποίες εφαρμόζονται δυναμική MLM. Οι αλλαγές αυτές, αποδείχθηκαν απλές αλλά αποτελεσματικές, οδηγώντας σε εντυπωσιακά εμπειρικά αποτελέσματα και αναδεικνύοντας ότι η διαδικασία NSP δεν είναι αναγκαία.

Το ALBERT (Lan, 2019) προτάθηκε από ερευνητές του Google Research το 2019. Ο στόχος τους ήταν να βελτιώσουν την εκπαίδευση και τα αποτελέσματα της

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

αρχιτεκτονικής BERT.Διαχώρισαν τους δύο πίνακες ενσωμάτωσης για να μειώσουν τις παραμέτρους, το οποίο πέτυχε μείωση κατά 80% με μικρή πτώση της απόδοσης σε σύγκριση με το BERT. Και εφάρμοσαν κοινή χρήση παραμέτρων μεταξύ διαφορετικών επιπέδων του για τη βελτίωση της αποδοτικότητας και τη μείωση της πλεονασμού.

Στην ενότητα αυτή παρουσιάστηκε το θεωρητικό και τεχνολογικό πλαίσιο της Βαθιάς Μάθησης (Deep Learning, DL): από τις βασικές αρχιτεκτονικές (CNNs, RNNs, GNNs, Transformers) και τις προκλήσεις υπερπροσαρμογής, έως την καθιέρωση των προεκπαιδευμένων μοντέλων και του Transfer Learning ως κυρίαρχης στρατηγικής. Η μετάβαση στην αρχιτεκτονική Transformer επέτρεψε τη δημιουργία θεμελιωδών γλωσσικών μοντέλων (π.χ. BERT, GPT, XLNet, RoBERTa, BART, T5) που μαθαίνουν πλούσιες αναπαραστάσεις από τεράστια σώματα κειμένου και στη συνέχεια εξειδικεύονται αποδοτικά με fine-tuning.

3.4 Επεξεργασία Φυσικής Γλώσσας για την Μετάφραση Κειμένων, Εξαγωγή Τοποθεσίας και Ενοποίηση Δεδομένων

Το NLP αποτελεί έναν από τους πλέον ραγδαία αναπτυσσόμενους κλάδους του AI, με αντικείμενο την ανάπτυξη αλγορίθμων και συστημάτων ικανών να αναλύουν, να κατανοούν, να ερμηνεύουν και να παράγουν ανθρώπινη γλώσσα σε μορφή κειμένου ή ομιλίας. Οι εφαρμογές της είναι ευρύτατες και περιλαμβάνουν συστήματα ερωτήσεων–απαντήσεων, αυτόματη περίληψη κειμένων, ανάκτηση πληροφοριών, ανάλυση συναισθήματος, αναγνώριση ομιλίας, αυτόματη διόρθωση κειμένου, καθώς και εξειδικευμένες εφαρμογές σε τομείς όπως η ιατρική (ανάλυση ιατρικών αρχείων), τα χρηματοοικονομικά (ανάλυση ειδήσεων και αναφορών), η κυβερνοασφάλεια (εντοπισμός απειλών σε κείμενα) και τα κοινωνικά δίκτυα (παρακολούθηση τάσεων και θεμάτων).

Η εξέλιξη του NLP αντικατοπτρίζει την ευρύτερη πορεία του AI. Από τις πρώτες μέρες των punched cards και των αργών αλγορίθμων μαζικής επεξεργασίας, όπου η ανάλυση μίας και μόνο πρότασης μπορούσε να διαρκέσει έως και επτά λεπτά, η τεχνολογία έχει φτάσει στις σημερινές υπερσύγχρονες εφαρμογές, όπως οι μηχανές αναζήτησης της Google, που επεξεργάζονται εκατομμύρια έγγραφα σε χιλιοστά του δευτερολέπτου.

Στις δεκαετίες του 1960–1980, η έμφαση δόθηκε σε κανόνες γραμματικής και λογικά συστήματα. Τη δεκαετία του 1990, η ανάπτυξη της υπολογιστικής στατιστικής και η αξιοποίηση μεγάλων corpora οδήγησαν στην εμφάνιση μεθόδων βασισμένων σε στατιστική μοντελοποίηση και ανάλυση συχνοτήτων. Σήμερα, η πρόοδος της ML και κυρίως της DL έχει επιτρέψει τη δημιουργία συστημάτων που κατανοούν και χειρίζονται τη γλώσσα με επίπεδα ακρίβειας που πλησιάζουν, και σε ορισμένες περιπτώσεις υπερβαίνουν, τις ανθρώπινες επιδόσεις.

Στο πλαίσιο της παρούσας πτυχιακής εργασίας, το NLP αξιοποιείται για την ανάλυση πολυγλωσσικών ειδησεογραφικών δεδομένων, με στόχο την αναγνώριση, κατηγοριοποίηση, περίληψη και ενοποίηση γεγονότων διαμαρτυρίας σε ένα κοινό αγγλόφωνο corpus. Η προσέγγιση αυτή καθιστά εφικτή την αυτοματοποιημένη επεξεργασία, σύγκριση και συνοπτική παρουσίαση άρθρων από διαφορετικές χώρες

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

της Ευρωπαϊκής Ένωσης, δημιουργώντας τις βάσεις για την περαιτέρω ανάλυση και χαρτογράφηση κοινωνικών κινητοποιήσεων.

3.4.1 Μηχανική Μετάφραση

Η Μηχανική Μετάφραση (Machine Translation, MT) αναφέρεται στη χρήση υπολογιστικών συστημάτων για την αυτόματη μετάφραση κειμένων από μία φυσική γλώσσα σε μία άλλη. Το ερευνητικό ενδιαφέρον για τη MT εμφανίστηκε παράλληλα με την ανάπτυξη των ηλεκτρονικών υπολογιστών. Δημοφιλείς αναφορές εντοπίζουν την προέλευσή της σε μια επιστολή που γράφτηκε από τον Warren Weaverin το 1949, μόλις λίγα χρόνια μετά την έναρξη λειτουργίας του ENIAC. Έκτοτε, η MTαποτελεί βασική εφαρμογή στον τομέα της NER, παρουσιάζοντας συνεχή εξέλιξη. Με την πάροδο του χρόνου, οι προσεγγίσεις έχουν μεταβληθεί σημαντικά, ξεκινώντας από την Μηχανική Μετάφραση που βασίζεται σε Κανόνες (Rule-Based Machine Translation, RBMT), προχωρώντας στη Στατιστική Μηχανική Μετάφραση (Statistical Machine Translation, SMT) και φτάνοντας στα σύγχρονα Νευρωνικά Μοντέλα Μετάφρασης (Neural Machine Translation, NMT).

3.4.1.1 Μηχανική Μετάφραση που βασίζεται σε Κανόνες

Μεταξύ 1950 και 1970, η πλειονότητα των συστημάτων MT βασιζόταν σε κανόνες. Το RBMT λειτουργούσε μέσω της εφαρμογής κανόνων και λεξικών που αφορούσαν συγκεκριμένες γλώσσες, προκειμένου να μεταφράζει κείμενα μεταξύ διαφορετικών γλωσσών. Οι κανόνες αυτοί, οι οποίοι συντάσσονταν κυρίως από γλωσσολόγους, επικεντρώνονταν στη γραμματική και τη σημασιολογία, αλλά συχνά δυσκολεύονταν να χειριστούν πολύπλοκες γλωσσικές δομές. Η προσέγγιση αντιμετώπιζε προκλήσεις που σχετίζονταν με την ποιότητα των λεξικών, τις αλληλεπιδράσεις των κανόνων και τις ιδιωματικές εκφράσεις, ενώ επιπλέον δυσκολευόταν να προσαρμοστεί σε νέους θεματικούς τομείς. Ένα χαρακτηριστικό παράδειγμα αποτελεί το SYSTRAN (Wheeler, 1984), το οποίο αναπτύχθηκε στα τέλη της δεκαετίας του 1960 για τη μετάφραση ρωσικών κειμένων στα αγγλικά.

Ενώ το RBMT μπορεί να έχει ορισμένους περιορισμούς, εξακολουθεί να έχει σημαντική αξία σε περιβάλλοντα που απαιτούν εξειδικευμένες γνώσεις ή υψηλό βαθμό ακρίβειας στο μεταφραστικό αποτέλεσμα. Καθιστώντας το ιδιαίτερα χρήσιμο σε τομείς όπου η ακριβής ορολογία είναι κρίσιμη, όπως νομικοί, ιατρικοί ή τεχνικοί τομείς. Επιπλέον, το RBMT μπορεί να ενσωματωθεί με άλλες μεθοδολογίες MT όπως το SMT ή το NMT εντός υβριδικών συστημάτων. Συνδυάζοντας την ακρίβεια που βασίζεται σε κανόνες του RBMT με τις δυνατότητες προσαρμοστικής μάθησης του SMT ή του NMT, τα υβριδικά συστήματα μπορούν να επιτύχουν αξιοπιστία και ευελιξία, καλύπτοντας ποικίλες ανάγκες μετάφρασης με μια ισορροπία δομημένων κανόνων και πληροφοριών που βασίζονται σε δεδομένα.

3.4.1.2 Στατιστική Μηχανική Μετάφραση

Το SMT αναπτύχθηκε στις αρχές της δεκαετίας του 1990, σηματοδοτώντας μια ουσιαστική μετατόπιση από τα παραδοσιακά συστήματα που βασίζονταν σε κανόνες προς προσεγγίσεις που αξιοποιούσαν δεδομένα. Τα πρώιμα μοντέλα, όπως το Candide της IBM, εστίαζαν στη μετάφραση σε επίπεδο λέξης, χρησιμοποιώντας πιθανοτικά μοντέλα που αντλούσαν παραμέτρους από μεγάλα δίγλωσσα σώματα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

κειμένων. Σταδιακά, αναπτύχθηκαν μέθοδοι SMT βασισμένες σε φράσεις, οι οποίες επέτρεπαν τη μετάφραση τμημάτων πολλαπλών λέξεων, βελτιώνοντας σημαντικά την ευχέρεια και την ακρίβεια. Μέχρι τα τέλη της δεκαετίας του 2000, η SMT εξελίχθηκε περαιτέρω με την εισαγωγή συντακτικών μοντέλων, τα οποία αξιοποιούσαν γραμματικές δομές για την καλύτερη διαχείριση σύνθετων γλωσσικών μοτίβων. Αν και η μέθοδος παραμένει χρήσιμη, η αποτελεσματικότητά της εξαρτάται σε μεγάλο βαθμό από τον όγκο, την ποιότητα και τη συνάφεια των διαθέσιμων δεδομένων. Η ανάπτυξη της SMT έθεσε τα θεμέλια για τις σύγχρονες τεχνικές MT, οδηγώντας τελικά στην άνοδο των NMT.

Το Moses (Cho K., 2014) αποτελεί ένα από τα σημαντικότερα εργαλεία ανοιχτού κώδικα για το SMT, προσφέροντας ένα πλήρες σύνολο λειτουργιών που συνέβαλαν καθοριστικά στην πρόοδο του πεδίου. Η υψηλή του απόδοση οφείλεται, σε μεγάλο βαθμό, στην ενσωμάτωση του Pharaoh (Liu Q. Z., 2007), ενός ισχυρού phrase-based αποκωδικοποιητή που αποτέλεσε τη βάση ανάπτυξής του. Το Pharaoh, εισήγαγε βελτιστοποιημένες δομές δεδομένων και τεχνικές διαχείρισης, επιτρέποντας στο Moses να επιτυγχάνει αποτελέσματα συγκρίσιμα με τα πιο εξελιγμένα SMT συστήματα τόσο σε ποιότητα όσο και σε ταχύτητα. Επιπλέον, επεκτείνει τη μετάφραση επιτρέποντας τη διαχείριση ασαφών εισόδων, όπως συμβαίνει στην αναγνώρισης ομιλίας με τη διαδικασία μετάφρασης. Οι αποδοτικές δομές δεδομένων, του επιτρέπουν τη χρήση μεγάλων παράλληλων corpora με περιορισμένους πόρους, καθιστώντας το κατάλληλο για έρευνα και εφαρμογές σε διάφορα ζεύγη γλωσσών. Παρά τα πλεονεκτήματα, το Moses παρουσιάζει κάποιους περιορισμούς. Αρχικά, η διαδικασία δημιουργίας κατάλληλου παράλληλου corpus για κάθε ζεύγος γλωσσών είναι ιδιαίτερα χρονοβόρα και τεχνικά απαιτητική και η ποιότητα των μεταφράσεων εξαρτάται άμεσα από το μέγεθος, την πληρότητα και την ακρίβεια του διαθέσιμου corpus. Τέλος, ένα ακόμη μειονέκτημα αποτελεί η περιορισμένη ενσωμάτωση γλωσσικών πληροφοριών στα SMT συστήματα μπορεί να οδηγήσει σε χαμηλότερη ακρίβεια σε σύγκριση με τα σύγχρονα NMT.

3.4.1.3 Νευρωνική Μηχανική Μετάφραση

Η άνοδος των τεχνικών DL τη δεκαετία του 2010 σηματοδότησε ριζική αλλαγή στο πεδίο του MT, οδηγώντας στην ανάπτυξη των μοντέλων NMT, τα οποία αξιοποιούν τεχνητά νευρωνικά δίκτυα και έχουν ξεπεράσει σε απόδοση τα παραδοσιακά συστήματα SMT. Σημαντικό ορόσημο αποτέλεσε η κυκλοφορία του Google Neural Machine Translation (GNMT) το 2016, το οποίο πέτυχε κορυφαία αποτελέσματα σε πλήθος ζευγών γλωσσών. Η επιτυχία αυτή οφείλεται κυρίως στη χρήση κατανεμημένων αναπαραστάσεων(distributed representations), που επιτρέπουν την εκπαίδευση συστημάτων μετάφρασης από άκρο σε άκρο (end-to-end training), βελτιώνοντας σημαντικά την ακρίβεια και την ευχέρεια των παραγόμενων μεταφράσεων.

Το NMT λειτουργεί μέσω μιας διαδικασίας εκμάθησης από άκρο σε άκρο (end-to-end learning), βασίζεται σε μηχανισμούς προσοχής αντί για παραδοσιακούς επαναλαμβανόμενους μηχανισμούς όπως τα Recurrent Neural Networks (RNNs) και έχει αποδειχθεί ότι αποδίδει καλά σε εργασίες μετάφρασης, (Sukhbaatar, 2015), μετατρέποντας απευθείας τις ακολουθίες πηγής σε ακολουθίες στόχου με τη χρήση τεχνικών DL. Σε αντίθεση με τις προηγούμενες μεθόδους, το NMT εκπαιδεύεται απευθείας πάνω σε μεγάλα παράλληλα corpora, χωρίς να βασίζεται σε χειροκίνητα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

δημιουργημένους κανόνες ή προκαθορισμένα στατιστικά μοντέλα. Η προσέγγιση αυτή απλοποιεί τη διαδικασία δημιουργίας μοντέλων και οδηγεί σε ταχύτερες, πιο ακριβείς και πιο φυσικές μεταφράσεις σε σύγκριση με τις παραδοσιακές στατιστικές τεχνικές.

Τα μοντέλα NMT διαθέτουν αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή οι οποίες ενισχύονται με μηχανισμούς προσοχής (attention mechanisms), επιτρέποντάς τους να αξιοποιούν αποτελεσματικά τα συμφραζόμενα και να παράγουν πιο ακριβείς και φυσικές μεταφράσεις, ακόμα και σε σύνθετες γλώσσες. Σημαντικές εξελίξεις, όπως τα μοντέλα Seq2Seq, τα οποία θα αναλυθούν παρακάτω, και η εισαγωγή της Transformer αρχιτεκτονικής, βελτίωσαν δραστικά την ακρίβεια, την αποδοτικότητα και την κλιμακωσιμότητα των συστημάτων. Η πρόοδος αυτή έχει καθιερώσει το NMT ως το κυρίαρχο παράδειγμα στη MT, σηματοδοτώντας τη μετάβαση σε πιο ισχυρά και ευέλικτα μοντέλα.

3.4.2 Σύγχρονες Προσεγγίσεις Μηχανικής Μετάφρασης

3.4.2.1 Sequence-To-Sequence (Seq2Seq)

Τα μοντέλα Seq2Seq, τα οποία αποτελούν θεμέλιο των NMT, αναπτύχθηκαν από τους Sutskever et al. (2014) (Sutskever, 2014). Βασίζονται σε Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNN) ή Μοντέλα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM) και έχουν σχεδιαστεί για τη μετατροπή ακολουθιών εισόδου μεταβλητού μήκους σε ακολουθίες εξόδου, ακόμη και όταν τα μήκη διαφέρουν σημαντικά. Η αρχιτεκτονική τους αποτελείται από δύο βασικά μέρη τον κωδικοποιητή και τον αποκωδικοποιητή. Ο κωδικοποιητής, επεξεργάζεται την ακολουθία εισόδου και τη μετατρέπει σε μία διανυσματική αναπαράσταση σταθερής διάστασης, η οποία ενσωματώνει τις κρίσιμες γλωσσικές πληροφορίες. Ενώ, ο αποκωδικοποιητής (decoder) παράγει την ακολουθία εξόδου, δημιουργώντας ένα token κάθε φορά.

Παρόλο που τα μοντέλα Seq2Seq είναι αποτελεσματικά, σε ορισμένες περιπτώσεις μόνο η διανυσματική αναπαράσταση δεν είναι αρκετή, κυρίως για μεγάλες ακολουθίες όπου ορισμένα τμήματα της εισόδου είναι πιο σημαντικά από άλλα. Η λύση του παραπάνω προβλήματος επιτεύχθηκε με την εισαγωγή μηχανισμών προσοχής σύντελεσε στην βελτίωση της ικανότητάς του να χειρίζεται μακροπρόθεσμες εξαρτήσεις, ενισχύοντας παράλληλα τη συνολική ποιότητα και ευχέρεια της μετάφρασης.

3.4.2.2 Transformer

Η εισαγωγή του μοντέλου Transformer (Vaswani, 2017) σήμανε σημαντική πρόοδο στο NMT. Οι Transformers χρησιμοποιούν καινοτόμους μηχανισμούς όπως η αυτοπροσοχή και η κωδικοποίηση θέσης, επιτρέποντας στο μοντέλο να λαμβάνει υπόψη όλες τις άλλες λέξεις στην ακολουθία εισόδου ενώ κωδικοποιεί κάθε λέξη. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να καταγράφει πιο αποτελεσματικά τις σύνθετες σχέσεις λέξεων.

Αυτό έφερε επανάσταση στο NLP χρησιμοποιώντας μια νέα αρχιτεκτονική που βασίζεται εξ ολοκλήρου σε μηχανισμούς αυτοπροσοχής (*self-attention*), εξαλείφοντας την ανάγκη για επαναλαμβανόμενα επίπεδα. Αυτός ο σχεδιασμός επιτρέπει στο μοντέλο να επεξεργάζεται τις ακολουθίες εισόδου παράλληλα και όχι διαδοχικά,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

αυξάνοντας σημαντικά την αποτελεσματικότητα της εκπαίδευσης και επιτρέποντας τον χειρισμό εξαρτήσεων μεγάλης εμβέλειας πιο αποτελεσματικά. Ο Transformer αποτελείται από μια δομή κωδικοποιητή-αποκωδικοποιητή, όπου ο κωδικοποιητής αντιστοιχίζει την ακολουθία εισόδου σε μια συνεχή αναπαράσταση και ο αποκωδικοποιητής παράγει την ακολουθία εξόδου. Η ικανότητά του να καταγράφει τις σχέσεις συμφραζομένων επιτυγχάνεται μέσω των multi-head attention, κατά τα οποία κάθε κεφαλή καταγράφει διαφορετικές σχέσεις ή μοτίβα στα δεδομένα. Αυτό εχει ως απόρροια, την βελτιώση σε εργασίες όπως η μηχανική μετάφραση, οδηγώντας στην ευρεία υιοθέτησή τους και χρησιμεύοντας ως βάση για πολλά μοντέλα αιχμής, συμπεριλαμβανομένων των BERT και GPT.

3.4.2.3 Bidirectional and Autoregressive Transformer (BART)

Το BART (Bidirectional and Autoregressive Transformer) (Lewis, 2019) είναι ένα προ-εκπαιδευμένο γλωσσικό μοντέλο που συνδυάζει τη δύναμη της αμφίδρομης και της αυτο-παλίνδρομης μοντελοποίησης. Εκπαιδεύεται χρησιμοποιώντας έναν αυτόματο κωδικοποιητή αποθορυβοποίησης, όπου μερικές προτάσεις καλύπτονται και το μοντέλο μαθαίνει να τις ανακατασκευάζει. Το BART έχει σχεδιαστεί για να χειρίζεται διάφορες εργασίες NLP, όπως η δημιουργία κειμένου, η ταξινόμηση κειμένου και η σύνοψη κειμένου. Επιτυγχάνει απόδοση αιχμής σε πολλά σημεία αναφοράς αξιοποιώντας μια αρχιτεκτονική που βασίζεται σε Transformer και δεδομένα εκπαίδευσης μεγάλης κλίμακας. Χάρη στον συνδυασμό αμφίδρομης κατανόησης και αυτοπαλίνδρομης, το BART αποτελεί ισχυρή βάση για ποικίλες εργασίες NLP.

3.4.2.4 MarianMT

Το MarianMT (Junczys-Dowmunt, 2018) είναι ένα πλαίσιο NMT που εστιάζει στην αποτελεσματικότητα και τη μετάφραση υψηλής απόδοσης. Έχει σχεδιαστεί για πολύγλωσσες και μεγάλης κλίμακας εργασίες μετάφρασης. Το MarianMT χρησιμοποιεί μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή βασισμένη σε μοντέλα Transformer, ενσωματώνοντας προηγμένες τεχνικές όπως ομαλοποίηση στρώσεων, εξομάλυνση ετικετών και αναζήτηση δέσμης. Υποστηρίζει διάφορες εργασίες MT, όπως μετάφραση σε επίπεδο πρότασης, μετάφραση σε επίπεδο εγγράφου και μαζική μετάφραση. Το MarianMT είναι γνωστό για τους γρήγορους χρόνους εκπαίδευσης και εξαγωγής συμπερασμάτων, γεγονός που το καθιστά κατάλληλο για εφαρμογές μετάφρασης σε πραγματικό χρόνο. Έχει επιτύχει ανταγωνιστικά αποτελέσματα σε πολλαπλά benchmarks MT και χρησιμοποιείται ευρέως τόσο για ερευνητικούς όσο και για παραγωγικούς σκοπούς.

3.4.3 Αναγνώριση Ονομασμένων Οντοτήτων

Η Αναγνώριση Ονομασμένων Οντοτήτων (Named Entity Recognition — NER) (Naseer, 2021) αποτελεί ένα βασικό στοιχείο στο NLP, καθώς επιτρέπει τον αυτόματο εντοπισμό και την ταξινόμηση οντοτήτων, όπως πρόσωπα, τοποθεσίες, οργανισμοί, ημερομηνίες και άλλες σημασιολογικές κατηγορίες, μέσα σε αδόμητο κείμενο. Η NER παίζει καίριο ρόλο σε εφαρμογές όπως μηχανική μετάφραση, συστήματα ερωταποκρίσεων, ανάλυση συναισθήματος και εξόρυξη πληροφορίας, καθώς βελτιώνει την κατανόηση του περιεχομένου από τα υπολογιστικά συστήματα. Σήμερα, διακρίνονται τρεις κύριες προσεγγίσεις NER: (α) rule-based, που χρησιμοποιούν χειροκίνητα ορισμένους συντακτικούς και λεξιλογικούς κανόνες, (β)

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

learning-based, που εκπαιδεύονται σε μεγάλα annotated datasets για να εντοπίζουν πρότυπα, και (γ) υβριδικές, που συνδυάζουν την ακρίβεια της ανθρώπινης εμπειρογνωμοσύνης με την προσαρμοστικότητα των αλγορίθμων μάθησης. Η κατάλληλη επιλογή προσέγγισης και εργαλείων εξαρτάται από το γλωσσικό πλαίσιο, το πεδίο εφαρμογής και τις απαιτήσεις ακρίβειας του εκάστοτε έργου, καθιστώντας την αρχική αξιολόγηση τεχνικών και εργαλείων κομβικό στάδιο σε κάθε έργο NLP.

3.4.3.1 Αναγνώριση Ονομασμένων Οντοτήτων βάσει Κανόνων

Οι rule-based προσεγγίσεις για το NER βασίζονται σε χειροκίνητα κατασκευασμένους κανόνες που προέρχονται από συντακτικά και λεξιλογικά μοτίβα, καθώς και από γλωσσολογική γνώση. Οι κανόνες αυτοί δημιουργούνται από ειδικούς και εφαρμόζονται συνήθως σε συγκεκριμένα πεδία, με στόχο την επίτευξη υψηλής ακρίβειας και υπολογιστικής αποδοτικότητας. Ένα χαρακτηριστικό παράδειγμα αυτής της προσέγγισης αποτελεί η μελέτη των Ben Abacha & Zweigenbaum (Ben Abacha, 2011), στην οποία συγκρίνονται σημασιολογικές και στατιστικές μέθοδοι για την αναγνώριση ιατρικών οντοτήτων, δείχνοντας ότι οι rule-based μέθοδοι μπορούν να επιτύχουν καλή απόδοση σε εξειδικευμένα domains. Συνολικά, τα rule-based συστήματα NER μπορούν να προσφέρουν ακριβή αποτελέσματα σε συγκεκριμένους τομείς, αλλά υστερούν σε ευελιξία και κλιμάκωση σε ευρύτερα ή διαφορετικά γλωσσικά περιβάλλοντα.

3.4.3.2 Αναγνώριση Ονομασμένων Οντοτήτων βάσει Μηχανικής Μάθησης

Τα τελευταία χρόνια, η ραγδαία πρόοδος στον τομέα της ML έχει οδηγήσει σε σημαντική βελτίωση της απόδοσης των συστημάτων NER. Σε αντίθεση με τις παραδοσιακές rule-based μεθόδους, που απαιτούν χειροκίνητη δημιουργία κανόνων, οι προσεγγίσεις που βασίζονται στη ML αξιοποιούν μεγάλα annotated datasets και ισχυρά αλγοριθμικά μοντέλα για να ανακαλύπτουν αυτόματα περίπλοκα μοτίβα μέσα στο κείμενο. Αυτή η αυτοματοποιημένη διαδικασία εκμάθησης καθιστά τα μοντέλα πιο προσαρμοστικά, ακριβή και αποδοτικά σε διαφορετικά πεδία εφαρμογής, ειδικά όταν υπάρχουν διαθέσιμα επαρκή δεδομένα εκπαίδευσης. Οι προσεγγίσεις που βασίζονται στη μάθηση (learning-based) στο NER χωρίζονται σε τρεις βασικές κατηγορίες:

- με εποπτευόμενη μάθηση (Supervised Learning)
- με ημι-εποπτευόμενη μάθηση (Semi-Supervised Learning)
- με μη εποπτευόμενη μάθηση (Unsupervised Learning)

Οι τεχνικές εποπτευόμενης μάθησης βασίζονται στην εκπαίδευση μοντέλων μέσω επισημασμένων συνόλων δεδομένων (labeled data), ώστε τα μοντέλα να μπορούν να προβλέπουν σωστά αποτελέσματα για νέα, μη επισημασμένα δεδομένα. Χρησιμοποιούνται αλγόριθμοι όπως τα Hidden Markov Models (HMM), Conditional Random Fields (CRF), Support Vector Machines (SVM) και σύγχρονες νευρωνικές αρχιτεκτονικές, οι οποίες έχουν αποδειχθεί ιδιαίτερα αποδοτικές σε εργασίες NER.

Στις ημι-εποπτευόμενες προσεγγίσεις, το σύστημα εκπαιδεύεται αρχικά σε ένα μικρό σύνολο επισημασμένων δεδομένων, ενώ στη συνέχεια χρησιμοποιεί ένα μεγαλύτερο σύνολο μη επισημασμένων δεδομένων για να βελτιώσει την απόδοσή του, αξιοποιώντας τεχνικές όπως self-training και co-training. Αυτό είναι ιδιαίτερα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

χρήσιμο σε περιπτώσεις όπου η συλλογή annotated δεδομένων είναι δύσκολη ή κοστοβόρα.

Τέλος, οι μη εποπτευόμενες προσεγγίσεις βασίζονται αποκλειστικά σε μη επισημασμένα δεδομένα και ανακαλύπτουν πρότυπα μέσω αλγορίθμων ομαδοποίησης (clustering) ή κατασκευής συσχετίσεων, χωρίς να απαιτείται ανθρώπινη επισημείωση. Ωστόσο, η ακρίβεια αυτών των μεθόδων είναι συνήθως χαμηλότερη συγκριτικά με τις εποπτευόμενες, παρόλο που αποτελούν χρήσιμη επιλογή όταν δεν υπάρχουν διαθέσιμα annotated corpora.

Η κατάλληλη επιλογή χαρακτηριστικών (features) και ετικετών αποτελεί κρίσιμο στάδιο σε τέτοια συστήματα, καθώς οι ετικέτες διαμορφώνουν τη βάση για τη δημιουργία των μοντέλων μάθησης. Τα μοντέλα που προκύπτουν είναι ικανά να αναγνωρίζουν πρότυπα και να ταξινομούν σωστά καινούργια δεδομένα.

3.4.3.3 Υβριδικές Προσεγγίσεις Αναγνώρησης Επώνυμων Οντοτήτων

Οι υβριδικές προσεγγίσεις στο NER στοχεύουν στη βέλτιστη αξιοποίηση των δυνατοτήτων διαφορετικών μεθόδων, συνδυάζοντας την ακρίβεια των τεχνικών βασισμένων σε κανόνες με την προσαρμοστικότητα και τη γενίκευση που προσφέρουν τα μοντέλα ML. Μέσω αυτού του συνδυασμού, τα συστήματα μπορούν να επιτυγχάνουν υψηλότερα ποσοστά επιτυχίας, αντιμετωπίζοντας παράλληλα την πολυπλοκότητα των φυσικών γλωσσών και τη μεταβλητότητα των συμφραζομένων.

3.4.4 Σύγκριση και ανάλυση διαφορετικών εργαλείων και αλγορίθμων για μοντέλα NER

Στο πλαίσιο της ανάπτυξης και εφαρμογής μοντέλων NER, έχουν δημιουργηθεί ποικίλα εργαλεία και βιβλιοθήκες ανοιχτού κώδικα, τα οποία διαφέρουν ως προς την αρχιτεκτονική, τις τεχνικές που αξιοποιούν, την υποστήριξη γλωσσών και την απόδοση σε διαφορετικά σύνολα δεδομένων. Η συγκριτική μελέτη αυτών των εργαλείων, όπως τα SpaCy, StanfordNLP, TensorFlow και Apache OpenNLP είναι κρίσιμη για την επιλογή της κατάλληλης λύσης, λαμβάνοντας υπόψη παραμέτρους όπως η ταχύτητα, η ακρίβεια, η δυνατότητα προσαρμογής και η ευκολία ενσωμάτωσης σε υπάρχουσες εφαρμογές. Στις επόμενες ενότητες παρουσιάζεται αναλυτικά η λειτουργικότητα, τα πλεονεκτήματα και οι περιορισμοί καθενός, με στόχο την τεκμηριωμένη αξιολόγηση και επιλογή του βέλτιστου εργαλείου για τις ανάγκες της παρούσας εργασίας.

3.4.4.1 SpaCy

Η Spacy (Explosion) αποτελεί μία από τις πιο δημοφιλείς βιβλιοθήκες ανοιχτού κώδικα στη γλώσσα Python, η οποία έχει σχεδιαστεί για την υλοποίηση ποικίλων εργασιών στο NLP. Υποστηρίζει λειτουργίες όπως η σήμανση μερών του λόγου (Part-of-Speech Tagging – POS), το NER, η ταξινόμηση κειμένου, η ανάλυση εξαρτήσεων, η μέτρηση ομοιότητας μεταξύ κειμένων, καθώς και η λημματοποίηση. Προσφέρει στατιστικά μοντέλα και pipelines επεξεργασίας για πολλές γλώσσες. Η ανάπτυξη και υποστήριξή της γίνεται από την εταιρεία Explosion AI.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η αρχιτεκτονική της SpaCy βασίζεται σε έναν υβριδικό συνδυασμό διαφορετικών τεχνικών, όπως Κρυφών Μαρκοβιακών Μοντέλων (Hidden Markov Models – HMM), Μοντέλων Μέγιστης Εντροπίας (Maximum Entropy Markov Models – MEMMs) και Δέντρων Απόφασης (Decision Trees), τα οποία ενσωματώνονται σε ένα CNN.

Τα HMM είναι στατιστικά μοντέλα που χρησιμοποιούνται ευρέως στο NLP, βασιζόμενα στην υπόθεση ότι η ακολουθία των παρατηρήσεων (π.χ. λέξεων) εξαρτάται από μια ακολουθία κρυφών καταστάσεων (π.χ. μέρη του λόγου ή τύπους οντοτήτων). Τα MEMMs αποτελούν εξέλιξη των HMM, καθώς ενσωματώνουν χαρακτηριστικά περιβάλλοντος και χρησιμοποιούν τη θεωρία της μέγιστης εντροπίας, επιτρέποντας την καλύτερη μοντελοποίηση πολύπλοκων εξαρτήσεων ανάμεσα στα δεδομένα. Ο συνδυασμός αυτών των τεχνικών με τα CNN παρέχει στη SpaCy τη δυνατότητα να επιτυγχάνει υψηλή ακρίβεια, επεκτασιμότητα και αποτελεσματικότητα, ειδικά όταν επεξεργάζεται μεγάλα σύνολα δεδομένων.

Επιπλέον, η SpaCy παρέχει προεκπαιδευμένα μοντέλα NER που αναγνωρίζουν οντότητες όπως ονόματα προσώπων, οργανισμούς, χρονικές εκφράσεις και τοποθεσίες, διευκολύνοντας την ανάπτυξη εφαρμογών χωρίς την ανάγκη εκπαίδευσης μοντέλων από το μηδέν.

3.4.4.2 StanfordNLP

Η StanfordNLP (Liu Q. Z., 2007) είναι μία ισχυρή βιβλιοθήκη για NLP, η οποία διατίθεται για τη γλώσσα Python και βασίζεται στο σύστημα Stanford CoreNLP. Πρόκειται για ένα ολοκληρωμένο πακέτο εργαλείων που υποστηρίζει πλήθος εργασιών NLP, όπως η σήμανση μερών του λόγου, η εξαγωγή μορφολογικών χαρακτηριστικών, η ανάλυση εξαρτήσεων μεταξύ λέξεων και φράσεων, καθώς και το NER. Η βιβλιοθήκη παρέχει υποστήριξη για περισσότερες από 70 γλώσσες και επιτρέπει την ενσωμάτωση λειτουργιών από το Java-based πακέτο CoreNLP, προσφέροντας έτσι ευελιξία και επεκτασιμότητα.

Στον πυρήνα της, η StanfordNLP αξιοποιεί το Conditional Random Fields (CRF) Sequence Model, το οποίο αποτελεί έναν ισχυρό ταξινομητή για την εκτέλεση εργασιών ακολουθιακής πρόβλεψης, όπως η NER. Εκτός από την NER, υποστηρίζει επιπλέον λειτουργίες όπως η ανάλυση συναισθήματος (sentiment analysis), η τμηματοποίηση κειμένου (tokenization) και η ανάλυση συντακτικών δομών. Ο συνδυασμός της υποστήριξης πολλών γλωσσών, της ακρίβειας των μοντέλων και της δυνατότητας ενοποίησης με το CoreNLP καθιστούν τη StanfordNLP ένα από τα πλέον αξιόπιστα και ευρέως χρησιμοποιούμενα εργαλεία στον χώρο του NLP.

3.4.4.3 TensorFlow

Το Tensorflow (Google) είναι μια ανοιχτού κώδικα βιβλιοθήκη για ML, η οποία αναπτύχθηκε από την Google και υποστηρίζει προγραμματισμό σε Python, C++ και CUDA. Έχει σχεδιαστεί για την ανάπτυξη και εκτέλεση πολύπλοκων μαθηματικών υπολογισμών, με έμφαση στην κατασκευή και εκπαίδευση νευρωνικών δικτύων. Το TensorFlow επεξεργάζεται δεδομένα σε αριθμητική μορφή, υποστηρίζοντας τόσο ακατέργαστες αριθμητικές εισόδους όσο και τεχνικές κωδικοποίησης, όπως το one-

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

hot encoding, για τη μετατροπή κειμενικών δεδομένων σε κατάλληλη μορφή για εκπαίδευση.

Χάρη στην ευελιξία και την υψηλή απόδοση που προσφέρει, η βιβλιοθήκη χρησιμοποιείται σε πλήθος εφαρμογών, όπως η MT(π.χ. Google Translate), η αυτόματη σύνοψη κειμένων, NER, η αναγνώριση ομιλίας, καθώς και η ανάλυση εικόνων και βίντεο. Η αρχιτεκτονική του TensorFlow επιτρέπει την αξιοποίηση επιταχυντών υλικού (GPU και TPU), γεγονός που το καθιστά ιδιαίτερα αποδοτικό για την εκπαίδευση και εξαγωγή προβλέψεων από μεγάλα και σύνθετα μοντέλα.

3.4.4.4 Apache OpenNLP

Το Apache OpenNLP είναι μία βιβλιοθήκη ανοιχτού κώδικα, υλοποιημένη σε Java, η οποία έχει αναπτυχθεί για την εκτέλεση ποικίλων εργασιών στην NLP. Υποστηρίζει λειτουργίες όπως η σήμανση μερών του λόγου, το NER, η τμηματοποίηση κειμένου, η διαίρεση σε προτάσεις και τμήματα κειμένου, καθώς και η ανάλυση εξαρτήσεων.

Η αρχιτεκτονική του OpenNLP βασίζεται, μεταξύ άλλων, σε αλγορίθμους perceptron και MEMMs, προσφέροντας αξιόπιστες λύσεις για βασικές αλλά και πιο εξειδικευμένες εργασίες NLP. Η βιβλιοθήκη παρέχει προεκπαιδευμένα μοντέλα σε πολλές γλώσσες, διευκολύνοντας την άμεση εφαρμογή της σε ποικίλα γλωσσικά περιβάλλοντα. Επιπλέον, προσφέρει χρήσιμες λειτουργίες όπως αναζήτηση συμβολοσειρών, διόρθωση ορθογραφίας και υποβοήθηση στη μετάφραση κειμένων, καθιστώντας την κατάλληλη τόσο για αυτόνομη χρήση όσο και για ενσωμάτωση σε μεγαλύτερα πληροφοριακά συστήματα.

Εργαλείο	Πλεονεκτήματα	Μειονεκτήματα
Spacy	- Πολύ μικρός χρόνος πρόβλεψης των NER μοντέλων - Παρέχει F-score ακρίβεια για κάθε ετικέτα - Εφαρμόζεται άμεσα σε κείμενα ή corpus - Μείωση απώλειας πληροφορίας σε κάθε επανάληψη εκπαίδευσης	- Μεγάλο μέγεθος μοντέλων σε σύγκριση με άλλα εργαλεία
StandfordNLP	- Υποστήριξη πολλών γλωσσών - Βελτιώσεις στο CoreNLP και ευκολία χρήσης - Μεγαλύτερη μνήμη και γρήγορη επεξεργασία - Εξ ολοκλήρου υλοποιημένο σε Python	- Μεγάλο μέγεθος μοντέλων λόγω πολυγλωσσικής υποστήριξης
Tensorflow	-Μικρός χρόνος πρόβλεψης - Εφαρμόζεται σε αριθμητικά δεδομένα - Μικρότερο μέγεθος από άλλα μοντέλα	- Δυσκολία στην αναγνώριση σχέσης μεταξύ οντοτήτων και ακολουθίας - Καλή απόδοση μόνο όταν η σειρά είναι σωστή

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Apache OpenNLP	- Χαμηλός χρόνος πρόβλεψης - Δεν περιλαμβάνει άγνωστα tokens ή ετικέτες - Εφαρμόζεται απευθείας σε corpus	- Πολύ χαμηλή ακρίβεια σε λανθασμένες προβλέψεις - Δεν παρέχει F-measure για κάθε ετικέτα

Στον Πίνακα 2 Μειονεκτήματα και πλεονεκτήματα των NLP

Στον Στον Πίνακα 2, απεικονίζονται τα πλεονεκτήματα και τα μειοντεκτήματα εργαλείων NLP. Το καθένα παρουσιάζει διαφορετικά μειονεκτήματα.

- Το SpaCy έχει τα πιο μεγάλα σε μέγεθος μοντέλα σε σύγκριση με τα άλλα.
- Το StanfordNLP παρουσιάζει υψηλό μέγεθος μοντέλων λόγω υποστήριξης γλωσσών.
- Το TensorFlow ενδέχεται να παρουσιάζει περιορισμούς στην αναγνώριση πολύπλοκων σχέσεων μεταξύ οντοτήτων, ιδιαίτερα αν η σειρά των δεδομένων εισόδου δεν είναι σωστά δομημένη.
- Το Apache OpenNLP δεν παρέχει F1-score για κάθε ετικέτα στο dataset, περιορίζοντας τη δυνατότητα λεπτομερούς αξιολόγησης.

Η ανάλυση που προηγήθηκε δείχνει ότι καμία μεθοδολογία ή βιβλιοθήκη NER δεν είναι απόλυτα ιδανική για κάθε περίπτωση. Οι rule-based προσεγγίσεις προσφέρουν ακρίβεια σε εξειδικευμένα πεδία, αλλά στερούνται στην ευελιξία. Οι learning-based μέθοδοι παρέχουν προσαρμοστικότητα και υψηλή απόδοση, αλλά απαιτούν μεγάλα annotated corpora. Τέλος, οι υβριδικές μέθοδοι συνδυάζουν πλεονεκτήματα, αλλά συνεπάγονται μεγαλύτερη πολυπλοκότητα στην υλοποίηση.

Αντίστοιχα, τα διαθέσιμα εργαλεία (SpaCy, StanfordNLP, TensorFlow, OpenNLP) παρουσιάζουν διαφορετικά πλεονεκτήματα και μειονεκτήματα. Έτσι το καθένα επιλέγεται βάσει των αναγκών της εκάστοτε εφαρμογής, το πλήθος και τη γλώσσα των δεδομένων, αλλά και τους διαθέσιμους υπολογιστικούς πόρους.

3.5 Αυτόματη Περίληψη Κειμένου

Η αυτόματη περίληψη κειμένου (text summarization) αποτελεί μια από τις σημαντικότερες προκλήσεις της NLP, καθώς επιδιώκει τη δημιουργία σύντομων και περιεκτικών περιλήψεων που μεταφέρουν τα ουσιώδη σημεία του αρχικού κειμένου. Από τις πρώτες στατιστικές προσεγγίσεις της δεκαετίας του 1950 (Luhn, 1958) (Baxendale, 1958) έως τα σύγχρονα νευρωνικά δίκτυα, η έρευνα έχει γνωρίσει ραγδαία εξέλιξη.

Οι τεχνικές περίληψης κειμένου έχουν εξελιχθεί σημαντικά τις τελευταίες δεκαετίες, περνώντας από απλές στατιστικές μεθόδους σε προηγμένα μοντέλα βασισμένα σε νευρωνικά δίκτυα.

Η εισαγωγή του ML έφερε πιο σύνθετες προσεγγίσεις, όπως μεθόδους εποπτευόμενης μάθησης με χειροκίνητα σχεδιασμένα χαρακτηριστικά και τεχνικές μοντελοποίησης θεμάτων. Ωστόσο, οι μέθοδοι αυτές συνέχιζαν να παρουσιάζουν περιορισμούς ως προς τη δημιουργία συνεκτικών και νοηματικά κατάλληλων περιλήψεων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η επανάσταση της DL αποτέλεσε σημείο καμπής στην έρευνα της περίληψης κειμένου. Τα μοντέλα Seq2Seq με μηχανισμούς προσοχής κατέστησαν εφικτή την πιο αποτελεσματική abstractive περίληψη, όπου το μοντέλο δημιουργεί νέο κείμενο που αποτυπώνει την ουσία του αρχικού εγγράφου. Τα μοντέλα αυτά παρουσίασαν βελτιωμένη ικανότητα συμπύκνωσης πληροφορίας και παραφράσεων, αλλά συχνά δυσκολεύονταν στη διατήρηση της πραγματολογικής ακρίβειας και στη διαχείριση μακροσκελών εγγράφων.

3.5.1 Εκχυλιστική και Αφαιρετική Περίληψη

Διακρίνονται δύο κύριες κατηγορίες μεθόδων:

- Εκχυλιστική περίληψη (extractive summarization), όπου επιλέγονται προτάσεις απευθείας από το κείμενο με βάση στατιστικά ή σημασιολογικά κριτήρια (π.χ. TF-IDF, Information Extraction). Εφαρμογές όπως το σύστημα RIPTIDES (White et al., 2001) χρησιμοποίησαν τέτοιες τεχνικές, ενώ πιο πρόσφατες μέθοδοι ενσωματώνουν νευρωνικά δίκτυα (Mohsen et al., 2020· Anand & Wagh, 2019).
- Αφαιρετική περίληψη (abstractive summarization), όπου δημιουργούνται νέες προτάσεις μέσω παραφράσεων και συμπύκνωσης πληροφορίας. Οι μέθοδοι αυτές απαιτούν εκτεταμένες τεχνικές NLP, ενώ πιο πρόσφατα έχουν αξιοποιήσει το πλαίσιο encoder–decoder (Xu et al., 2020· Lee et al., 2020).

3.5.2 Προεκπαιδευμένα Μοντέλα για Μηχανική Μετάφραση

Η πρόοδος της DL και η έλευση των Transformers έφεραν επανάσταση στην περίληψη, με την ανάπτυξη προεκπαιδευμένων μοντέλων ειδικά σχεδιασμένων για το abstractive summarization:

Pegasus (Pre-training with Extracted Gap-sentences για Abstractive SUmmarization Sequence-to-sequence model): Ένα μοντέλο που δημιουργήθηκε από την Google και σχεδιάστηκε ειδικά για abstractive summarization. Η καινοτομία του έγκειται στη μέθοδο προεκπαίδευσης, όπου αφαιρούνται προτάσεις-κλειδιά από το κείμενο και το μοντέλο καλείται να τις ανακατασκευάσει. Με αυτόν τον τρόπο, το Pegasus μαθαίνει να εστιάζει στις πιο κρίσιμες πληροφορίες του κειμένου. Έχει αποδειχθεί ιδιαίτερα αποδοτικό σε καθιερωμένα σύνολα δεδομένων, επιτυγχάνοντας υψηλές τιμές στις μετρικές ROUGE και BLEU.

T5 (Text-To-Text Transfer Transformer): Πρόκειται για ένα ενοποιημένο μοντέλο της Google, το οποίο μετατρέπει όλα τα NLP tasks σε μορφή “text-to-text”. Το T5 έχει μεγάλη ευελιξία και μπορεί να εφαρμοστεί σε πλήθος εφαρμογών, όχι μόνο στην περίληψη, αλλά και στη μετάφραση, απάντηση ερωτήσεων, ταξινόμηση κειμένων κ.ά. Ωστόσο, στις εφαρμογές περίληψης κειμένου έχει αποδειχθεί λιγότερο αποδοτικό σε σχέση με πιο εξειδικευμένα μοντέλα, όπως το Pegasus ή το BART.

BART (Bidirectional and Auto-Regressive Transformers): Αναπτύχθηκε από τη Facebook AI και αποτελεί υβριδικό μοντέλο που συνδυάζει τα πλεονεκτήματα του BERT (διπλής κατεύθυνσης κωδικοποίηση) και του GPT (αυτοπαλίνδρομη αποκωδικοποίηση). Εκπαιδεύεται με στόχο denoising autoencoder, δηλαδή να αποκαθιστά κείμενα που έχουν τροποποιηθεί με θόρυβο (masking, αναδιάταξη

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

φράσεων). Αυτό του προσδίδει ισχυρές δυνατότητες στη δημιουργία συνεκτικών και ρεαλιστικών περιλήψεων. Το BART έχει αναδειχθεί ως ένα από τα κορυφαία μοντέλα abstractive summarization, με εξαιρετικές επιδόσεις σε μεγάλα benchmark datasets.

Εν κατακλείδι, το NER αποτελεί θεμελιώδη διαδικασία στο NLP, με κρίσιμη συμβολή στη δομημένη αναπαράσταση μη δομημένου κειμενικού περιεχομένου. Οι προσεγγίσεις που εξετάστηκαν — βασισμένες σε κανόνες, βασισμένες στη μάθηση και υβριδικές — παρουσιάζουν διακριτά πλεονεκτήματα και περιορισμούς. Τα συστήματα βασισμένα σε κανόνες προσφέρουν υψηλή ακρίβεια σε εξειδικευμένα πεδία, αλλά υστερούν σε ευελιξία και δυνατότητα κλιμάκωσης. Οι μέθοδοι ML είναι πιο γενικεύσιμες και μπορούν να επεξεργαστούν μεγάλα και ετερογενή σύνολα δεδομένων, απαιτούν όμως σημαντική ποσότητα και ποιότητα επισημασμένων δεδομένων εκπαίδευσης. Οι υβριδικές προσεγγίσεις, συνδυάζοντας την ανθρώπινη εμπειρογνωμοσύνη με την αυτόματη μάθηση, εμφανίζουν αυξημένη ακρίβεια και προσαρμοστικότητα.

Η συγκριτική ανάλυση εργαλείων όπως τα SpaCy, StanfordNLP, TensorFlow και Apache OpenNLP ανέδειξε ότι κάθε εργαλείο έχει διαφορετικά δυνατά σημεία και περιορισμούς, ανάλογα με τη γλώσσα-στόχο, το μέγεθος του μοντέλου, την ταχύτητα πρόβλεψης και την ακρίβεια (Precision, Recall, F1-score). Για παράδειγμα, το SpaCy διακρίνεται για την ταχύτητα πρόβλεψης και την ευκολία εφαρμογής, ενώ το StanfordNLP υπερτερεί στην πολυγλωσσική υποστήριξη. Το TensorFlow προσφέρει ευελιξία στην προσαρμογή νευρωνικών μοντέλων, ενώ το OpenNLP παρέχει μια σταθερή Java-based λύση με έμφαση στην ενσωμάτωση σε υπάρχουσες εφαρμογές.

Στο πλαίσιο της παρούσας πτυχιακής εργασίας, η θεωρητική κατανόηση των τεχνικών NER και η αξιολόγηση των διαθέσιμων εργαλείων συμβάλλουν στην τεκμηριωμένη επιλογή της κατάλληλης μεθοδολογίας και υλοποίησης. Η επιλογή του κατάλληλου εργαλείου ή συνδυασμού εργαλείων θα καθοριστεί από τις ιδιαιτερότητες του σώματος κειμένων, τις απαιτήσεις ακρίβειας, καθώς και την ανάγκη για προσαρμογή σε συγκεκριμένο θεματικό και γλωσσικό πλαίσιο.

3.6 Σύγκριση και Υλοποίηση Τεχνικών Stemming και Lemmatization για Κανονικοποίηση Κειμένου

Η ανάλυση δεδομένων σε μεγάλη κλίμακα, όπως στην περίπτωση της συλλογής ειδησεογραφικών άρθρων μέσω διαδικασιών web scraping, συνεπάγεται την επεξεργασία ενός τεράστιου όγκου κειμένου με πλούσια γλωσσική ποικιλία. Η ύπαρξη πολλαπλών μορφολογικών παραλλαγών, συνωνυμιών, διαφορετικών κλιτικών μορφών και πτώσεων δημιουργεί σημαντική πολυπλοκότητα στη διαχείριση των δεδομένων και μπορεί να οδηγήσει σε θόρυβο ή επαναλαμβανόμενες καταχωρήσεις δυσχεραίνοντας την ακρίβεια της ανάλυσης.

Για τον λόγο αυτό, καθίσταται αναγκαία η εφαρμογή μεθόδων κανονικοποίησης κειμένου (text normalization), οι οποίες στοχεύουν στη μείωση της μορφολογικής πολυπλοκότητας και στην ενοποίηση των λέξεων με κοινές αναπαραστάσεις. Η διαδικασία αυτή εντάσσεται στο στάδιο της προεπεξεργασίας κειμένου (text preprocessing), το οποίο προηγείται των μεθόδων εξόρυξης πληροφορίας, αναγνώρισης προτύπων και DL.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Δύο από τις βασικότερες τεχνικές που εφαρμόζονται σε αυτό το στάδιο είναι το stemming (αποκοπή καταλήξεων) και το lemmatization (λημματοποίηση). Το stemming αποκόπτει τις λέξεις από τη ρίζα τους, αφαιρώντας καταλήξεις και γραμματικούς τύπους. Το αποτέλεσμα δεν είναι πάντοτε λεξικογραφικά έγκυρο. Ωστόσο, οδηγεί σε μια πιο συμπαγή μορφή του λεξιλογίου, χρήσιμη για εργασίες όπως η αναζήτηση και η θεματική κατηγοριοποίηση. Αντίθετα, το lemmatization επιστρέφει τη βασική μορφή της λέξης, λαμβάνοντας υπόψη τα γραμματικά και συντακτικά συμφραζόμενα. Η διαδικασία βασίζεται σε πιο σύνθετα γλωσσικά μοντέλα, προσφέροντας μεγαλύτερη ακρίβεια και μειώνεται τον κίνδυνο λανθασμένων συγχωνεύσεων όρων. Και οι δύο τεχνικές διαδραματίζουν κρίσιμο ρόλο στην ενίσχυση της συνάφειας και της ακρίβειας κατά την αναζήτηση, την ανάλυση και την ευρετηρίαση δεδομένων. Με τη χρήση stemming ή lemmatization, ο αριθμός των ευρετηρίων μειώνεται σημαντικά, καθώς διαφορετικές μορφές της ίδιας λέξης μπορούν να αναγνωριστούν και να συγχωνευτούν σε μία ενιαία αναπαράσταση.

Για παράδειγμα, η λέξη industrialize μπορεί να συνδεθεί με λέξεις όπως industrious, industry, industrialized, και industrialization, καθώς όλες μοιράζονται κοινή ρίζα. Έτσι, το σύστημα αντιμετωπίζει αυτές τις παραλλαγές ως εκφάνσεις της ίδιας εννοιολογικής οντότητας, μειώνοντας τον θόρυβο και βελτιώνοντας την ακρίβεια της αναζήτησης και της κατηγοριοποίησης.

Αυτό το βήμα είναι καθοριστικό για την απλοποίηση της λεξιλογικής πολυπλοκότητας και για την ενίσχυση της απόδοσης σε εφαρμογές NLP και ανάλυσης κειμένου, όπου η ποιότητα των δεδομένων εισόδου καθορίζει την αποτελεσματικότητα των μοντέλων ML.

3.6.1 Stemming

Η τεχνική του stemming αποτελεί μία από τις βασικότερες μεθόδους κανονικοποίησης κειμένου. Στόχος της είναι η αποκοπή καταλήξεων, ώστε να μειώνονται οι λέξεις από τη ρίζα τους. Με αυτόν τον τρόπο, οι παραλλαγές μιας λέξης ενοποιούνται σε κοινή μορφή, γεγονός που συμβάλλει στην ανάκτηση περισσότερων συναφών εγγράφων και στη μείωση της πολυπλοκότητας του λεξιλογίου. Ωστόσο, το stemming δεν είναι απαλλαγμένο από σφάλματα. Δύο βασικές κατηγορίες προβλημάτων είναι η υπο-αποκοπή (under-stemming) και η υπερ-αποκοπή (over-stemming). Στην υπο-αποκοπή δύο λέξεις που ανήκουν στην ίδια εννοιολογική ομάδα καταλήγουν σε διαφορετικές ρίζες. Για παράδειγμα, μια αναζήτηση για τη λέξη run μπορεί να μη βρει έγγραφα που περιέχουν τις λέξεις running ή ran. Στην υπερ-αποκοπή λέξεις διαφορετικών εννοιολογικών ομάδων καταλήγουν στην ίδια ρίζα. Για παράδειγμα, η αναζήτηση για new μπορεί να επιστρέψει αποτελέσματα που περιέχουν τη λέξη news.

Για την αντιμετώπιση αυτών των ζητημάτων έχουν αναπτυχθεί διάφοροι αλγόριθμοι stemming, με διαφορετικά χαρακτηριστικά και βαθμό ακρίβειας. Ο αλγόριθμος Lancaster Stemmer (Paice/Husk) (A., 1996) αναπτύχθηκε από το Πανεπιστήμιο του Lancaster και αποτελεί έναν από τους πιο διαδεδομένους αλγορίθμους stemming. Το NLTK διαθέτει την κλάση LancasterStemmer, μέσω της οποίας μπορεί να εφαρμοστεί ο αλγόριθμος Lancaster Stemming για την εκτέλεση της διαδικασίας stemming. Λειτουργεί με τη χρήση κανόνων, οι οποίοι διαβάζονται από πίνακα και εφαρμόζονται διαδοχικά μέχρι να προκύψει η τελική ρίζα.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Ένας ακόμη αλγόριθμος είναι ο Lovins (Lovins), είναι ένας από τους παλαιότερους stemmers, σχεδιασμένος για αναζήτηση πληροφοριών. Λειτουργεί με μία μόνο σάρωση της λέξης και αφαιρεί τη μεγαλύτερη κατάληξη που ταιριάζει από μία προκαθορισμένη λίστα περίπου 250 καταλήξεων. Εξασφαλίζει ότι η τελική λέξη έχει τουλάχιστον τρία γράμματα. Χρησιμοποιεί μια λίστα με 250 καταλήξεις και αφαιρεί τη μεγαλύτερη που ταιριάζει, φροντίζοντας η τελική λέξη να έχει τουλάχιστον τρία γράμματα.

Τέλος ο αλγόριθμος Porter (R. Hooper, 2013) είναι ένας από τους πιο γνωστούς αλγόριθμους stemming για την αγγλική γλώσσα. Χρησιμοποιείται κυρίως σε εφαρμογές αναζήτησης και επεξεργασίας κειμένου. Λειτουργεί βήμα-βήμα, ακολουθώντας ένα σύνολο κανόνων για να αφαιρεί καταλήξεις όπως *-ing*, *-ed*, *-ly* κ.ά., ώστε να προκύψει η ρίζα της λέξης. Είναι πιο ακριβής από απλούς αλγόριθμους γιατί λαμβάνει υπόψη του, τη μορφή της λέξης και προσπαθεί να αποφύγει λάθη, όπως η αφαίρεση μέρους της ρίζας. Το NLTK διαθέτει μια κλάση PorterStemmer, η οποία επιτρέπει την χρήση του αλγόριθμου Stemmer Porter. Η εξέλιξη του Porter αποτελεί ο αλγόριθμος Snowball, ο οποίοος είναι ένας εξαιρετικά χρήσιμος αλγόριθμος για τη διαδικασία δημιουργίας ριζών λέξεων σε πολλές γλώσσες. Υποστηρίζει 15 γλώσσες, εκτός από τα Αγγλικά, γεγονός που τον καθιστά ευέλικτο για πολυγλωσσικά δεδομένα.

3.6.2 Lemmatization

Η λημματοποίηση αποτελεί μία πιο εξελιγμένη και ακριβή μέθοδο κανονικοποίησης κειμένου σε σχέση με το stemming. Αντί να εφαρμόζει απλή αποκοπή καταλήξεων, βασίζεται στη χρήση λεξικών πόρων και στη μορφολογική ανάλυση των λέξεων, προκειμένου να αφαιρέσει τις καταλήξεις κλίσης και να επαναφέρει τις λέξεις στη βασική, λεξικογραφική τους μορφή. Η διαδικασία λαμβάνει υπόψη τη συντακτική και γραμματική λειτουργία της λέξης. Δηλαδή, ελέγχει αν χρησιμοποιείται ως ρήμα, ουσιαστικό, ή επίθετο, ώστε να εξασφαλίζεται η σωστή αναγωγή.

Για παράδειγμα:

- running → run (ρήμα)
- better → good (επίθετο, συγκριτικός βαθμός)

Ένα επιπλέον πλεονέκτημα της λημματοποίησης είναι η δυνατότητα αντιστοίχισης συνωνύμων, διευκολύνοντας την ομαδοποίηση διαφορετικών όρων με παρόμοιο εννοιολογικό περιεχόμενο. Για παράδειγμα:

- η λέξη hot (καυτός) μπορεί να συσχετιστεί με το warm (ζεστός),
- η λέξη car (αυτοκίνητο) μπορεί να ανακτηθεί μαζί με τα cars ή automobile (όχημα).

Η τεχνική της λημματοποίησης έχει χρησιμοποιηθεί σε πολλές γλώσσες για εφαρμογές ανάκτησης πληροφορίας και NLP. Ενδεικτικά, οι Ozturkmenoglu και Alpkocak συνέκριναν τρεις διαφορετικούς λεμματοποιητές σε τουρκικό σύνολο δεδομένων και κατέληξαν στο συμπέρασμα ότι η λημματοποίηση βελτιώνει σημαντικά την απόδοση της ανάκτησης, ακόμη και με περιορισμένο αριθμό όρων στο

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

σύστημα. Επιπλέον, διαπίστωσαν ότι η καλύτερη απόδοση επετεύχθη όταν χρησιμοποιήθηκε το μέγιστο δυνατό μήκος λημμάτων.

3.6.3 Σύγκριση Stemming και Lemmatization

Παρόλο που και οι δύο τεχνικές αποσκοπούν στη μείωση της λεξιλογικής πολυπλοκότητας, οι διαφορές τους είναι ουσιώδεις:

- Stemming**:** Βασίζεται σε απλούς κανόνες αποκοπής καταλήξεων και προθημάτων. Είναι γρήγορη τεχνική, αλλά συχνά οδηγεί σε μη λεξικογραφικά έγκυρες ρίζες. Παράδειγμα: studies → studi.
- Lemmatization**:** Χρησιμοποιεί γλωσσολογικούς πόρους και γραμματική ανάλυση για να επιστρέψει τη σωστή λεξικογραφική μορφή της λέξης. Παράδειγμα: studies → study.

Το stemming είναι ταχύτερο αλλά λιγότερο ακριβές**,** ενώ η λημματοποίηση είναι πιο ακριβής αλλά πιο απαιτητική υπολογιστικά. Η επιλογή εξαρτάται από το εκάστοτε πρόβλημα: σε συστήματα μεγάλης κλίμακας, όπου προέχει η ταχύτητα (π.χ. μηχανές αναζήτησης), χρησιμοποιείται συχνά το stemming· αντίθετα, σε εφαρμογές όπου η γλωσσική ακρίβεια είναι κρίσιμη (π.χ. sentiment analysis, machine translation), προτιμάται η λημματοποίηση.

Συνοψίζοντας, οι τεχνικές stemming και lemmatization αποτελούν κρίσιμα εργαλεία κατά την κανονικοποίηση κειμένου και διαδραματίζουν καθοριστικό ρόλο στην NLP. Μέσω της μείωσης της λεξιλογικής πολυπλοκότητας και της ενοποίησης μορφολογικά συγγενικών όρων, βελτιώνουν την ακρίβεια και την αποδοτικότητα σε εφαρμογές όπως η ανάκτηση πληροφορίας, η ταξινόμηση κειμένων, η θεματική ανάλυση και η εξαγωγή γνώσης.

Η επιλογή μεταξύ stemming και lemmatization εξαρτάται από τις απαιτήσεις της εκάστοτε εφαρμογής, καθώς το stemming είναι πιο γρήγορο αλλά λιγότερο ακριβές, ενώ η λημματοποίηση είναι υπολογιστικά πιο απαιτητική, αλλά εξασφαλίζει σημασιολογική συνέπεια

3.7 Ομοιότητα Κειμένων

Η μέτρηση ομοιότητας κειμένου αποτελεί έναν από τους βασικούς άξονες στο NLP, καθώς προσδιορίζει τον βαθμό στον οποίο δύο κείμενα είναι σημασιολογικά συγγενικά ή εκφράζουν παρόμοιο περιεχόμενο. Σύμφωνα με τους (Wang, 2020), η ομοιότητα κειμένων αποτελεί θεμέλιο για πλήθος εφαρμογών όπως η ανάκτηση πληροφοριών, η αυτόματη απάντηση ερωτήσεων, η μηχανική μετάφραση, τα διαλογικά συστήματα και η αντιστοίχιση εγγράφων.

Η ομοιότητα αυτή μπορεί να μετρηθεί με διάφορους τρόπους, οι οποίοι κατηγοριοποιούνται, σε μεθόδους βασισμένες στην απόσταση χαρακτήρων ή λέξεων (π.χ. απόσταση Levenshtein), σε στατιστικές μεθόδους (π.χ. δείκτης Jaccard) και σε διανυσματικές μεθόδους, όπως η συνημιτονική ομοιότητα (cosine similarity).

Αρχικά, είναι σημαντικό να διαχωρίζεται η σημασιολογική ομοιότητα από τη σημασιολογική συγγένεια. Για παράδειγμα, οι λέξεις «king» και «man» έχουν

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

συγγενή σημασία λόγω της σχέσης τους, αλλά δεν είναι σημασιολογικά παρόμοιες. Αντιθέτως, οι λέξεις «king» και «queen» είναι σημασιολογικά παρόμοιες, καθώς ανήκουν στην ίδια κατηγορία εννοιών. Συνεπώς, η σημασιολογική ομοιότητα μπορεί να θεωρηθεί υποκατηγορία της ευρύτερης έννοιας της σημασιολογικής συγγένειας. Αντιστοίχως, στην ανάλυση ειδήσεων σχετικών με διαμαρτυρίες, δύο άρθρα μπορεί να είναι συγγενή επειδή αναφέρονται στην ίδια θεματική (π.χ. «συνεχιζόμενες απεργίες υγείας στην Ιταλία»), χωρίς να είναι παρόμοια σε επίπεδο συγκεκριμένου συμβάντος (ημερομηνία/τοποθεσία/δρώντες). Η διάκριση αυτή είναι κρίσιμη:

- **Διπλότυπο** στοχεύει πολύ υψηλή ομοιότητα, σχεδόν στο ίδιο συμβάν (ίδιος χρόνος και τόπος, ίδιος πυρήνας γεγονότος), συχνά αφορά μικρές λεκτικές παραλλαγές, ενδεικτικά σε περιπτώσεις όπου το ίδιο συμβάν περιγράφεται από διαφορετικό πρακτορείο ειδήσεων.
- **Follow-up** στοχεύει στην εξέλιξη γύρω από ένα θέμα, όπως για παράδειγμα με μια ενδεικτική ακολουθία της μορφής: διαδήλωση → συλλήψεις → δίκη, δηλαδή με συμβάντα που διαθέτουν χαμηλότερη άμεση ομοιότητα κειμένου αλλά ισχυρή χρονική και αιτιακή συνάφεια.

3.7.1 Απόσταση μήκους

Η απόσταση μήκους αποτελεί μία από τις απλούστερες μεθόδους υπολογισμού ομοιότητας μεταξύ κειμένων. Η βασική της ιδέα είναι ότι δύο κείμενα με παρόμοιο αριθμό χαρακτήρων ή λέξεων τείνουν να εμφανίζουν υψηλότερο βαθμό συγγένειας, ενώ οι μεγάλες διαφορές μήκους συνήθως δηλώνουν χαμηλότερη ομοιότητα. Η μέθοδος στηρίζεται σε αριθμητικά χαρακτηριστικά, όπως το πλήθος των λέξεων ή το μέγεθος του διανύσματος αναπαράστασης του κειμένου, και υπολογίζει την απόσταση μεταξύ τους.

3.7.2 Ευκλείδεια απόσταση

Η Ευκλείδεια απόσταση (Euclidean distance, γνωστή και ως L2-norm) (Deza, 2009) αποτελεί το πιο κλασικό μέτρο απόστασης σε έναν Ευκλείδειο χώρο. Ορίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των διαφορών ανάμεσα στις αντίστοιχες συντεταγμένες δύο σημείων. Με άλλα λόγια, αντιστοιχεί στη γραμμική απόσταση που ενώνει δύο σημεία στο χώρο.

$$d(S_a,S_b) \quad \sqrt{\sum_{t\ 1}^{n} (S_a^{(t)},S_b^{(t)})^2}$$

Το μέτρο αυτό μπορεί να επεκταθεί εύκολα σε πολυδιάστατους χώρους, κάτι που το καθιστά ιδιαίτερα χρήσιμο όταν τα κείμενα αναπαρίστανται ως διανύσματα. Παραδείγματα τέτοιων αναπαραστάσεων είναι τα διανύσματα συχνοτήτων λέξεων (BoW) ή τα TF-IDF διανύσματα, όπου κάθε διάσταση αντιστοιχεί σε ένα χαρακτηριστικό του κειμένου. Σε αυτό το πλαίσιο, η ευκλείδεια απόσταση μεταξύ δύο διανυσμάτων ισούται με την απόσταση μεταξύ των αντίστοιχων κειμένων, όσο πιο μικρή απόσταση, τόσο πιο παρόμοια θεωρούνται τα κείμενα.

Η τιμή της ευκλείδειας απόστασης είναι πάντοτε μη αρνητική και κυμαίνεται θεωρητικά από το 0 (ταυτόσημα διανύσματα, άρα απόλυτη ομοιότητα) έως το άπειρο

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

(πλήρως ασυσχέτιστα διανύσματα). Στην πράξη, μικρές αποστάσεις δείχνουν υψηλή ομοιότητα, ενώ μεγάλες τιμές υποδηλώνουν χαμηλή ομοιότητα.

Η μέθοδος αυτή χρησιμοποιείται εκτενώς σε εφαρμογές ομαδοποίησης (clustering), ταξινόμησης (classification) και ανίχνευσης ανωμαλιών (anomaly detection), καθώς προσφέρει έναν απλό αλλά αποτελεσματικό τρόπο μέτρησης της εγγύτητας μεταξύ δεδομένων.

3.7.3 Απόσταση Manhattan

Η απόσταση Manhattan (Deza, 2009), γνωστή και ως L1-norm, είναι μία από τις βασικές μεθόδους μέτρησης της απόστασης μεταξύ δύο διανυσμάτων πραγματικών τιμών. Σε αντίθεση με την ευκλείδεια απόσταση, η οποία λαμβάνει υπόψη τη συντομότερη διαδρομή, η απόσταση Manhattan υπολογίζει το άθροισμα των απόλυτων διαφορών μεταξύ των αντίστοιχων στοιχείων των διανυσμάτων:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Εφαρμόζεται κυρίως σε αραιές διανυσματικές αναπαραστάσεις, όπως το one-hot encoding, όπου κάθε λέξη αντιστοιχίζεται σε διάνυσμα με 1 στη θέση της και 0 στις υπόλοιπες ή τα Term Frequency, όπου κάθε διάσταση δείχνει πόσες φορές εμφανίζεται μια λέξη σε ένα έγγραφο. Σε τέτοιες αναπαραστάσεις, κάθε διάσταση δηλώνει την παρουσία ή τη συχνότητα μιας λέξης σε ένα κείμενο, και η Manhattan απόσταση μετρά το άθροισμα των απόλυτων διαφορών μεταξύ δύο διανυσμάτων, δηλαδή το πόσο διαφέρουν τα κείμενα ως προς την παρουσία ή τη συχνότητα των λέξεων που περιέχουν.

Για παράδειγμα, αν δύο άρθρα έχουν πολύ κοντινές αναπαραστάσεις σε one-hot encoding, τότε η Manhattan απόστασή τους θα είναι μικρή, κάτι που υποδηλώνει μεγάλη ομοιότητα στο λεξιλόγιο που χρησιμοποιούν. Αντίθετα, μια μεγάλη απόσταση σημαίνει ότι τα κείμενα διαφέρουν σημαντικά ως προς τις λέξεις τους. Η μέθοδος αυτή είναι απλή, εύκολη στον υπολογισμό και αρκετά ανθεκτική σε εξαιρέσεις, καθώς μετρά γραμμικά τις διαφορές ανά διάσταση και δεν επηρεάζεται υπερβολικά από μεμονωμένες λέξεις με πολύ υψηλή συχνότητα.

3.7.4 Απόσταση Levenshtein

Η απόσταση Levenshtein (Levenshtein, 1966), είναι ένα μέτρο που ποσοτικοποιεί τη διαφορά μεταξύ δύο συμβολοσειρών χαρακτήρων. Ορίζεται ως ο ελάχιστος αριθμός βασικών λειτουργιών επεξεργασίας που απαιτούνται για να μετατραπεί η μία συμβολοσειρά στην άλλη.

Η απόσταση Levenshtein είναι ιδιαίτερα χρήσιμη όταν συγκρίνονται παρόμοιες αλλά όχι ταυτόσημες συμβολοσειρές, καθώς μετρά το αριθμό των αλλαγών που τις διαφοροποιούν. Συχνά, χρησιμοποιείται σε εφαρμογές όπως η διόρθωση ορθογραφικών λαθών, η αναζήτηση προσεγγιστικών αντιστοιχιών (approximate string matching), αλλά και σε προβλήματα αναγνώρισης κειμένων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

3.7.5 Ομοιότητα Jaccard

Ο δείκτης Jaccard (Jaccard, 1901), γνωστός και ως συντελεστής ομοιότητας Jaccard, μετρά την ομοιότητα μεταξύ δύο συνόλων. Ορίζεται ως ο λόγος του μεγέθους της τομής των συνόλων προς το μέγεθος της ένωσης των συνόλων. Με άλλα λόγια, είναι η αναλογία των κοινών στοιχείων μεταξύ δύο συνόλων. Είναι ιδιαίτερα χρήσιμος όταν η παρουσία ή η απουσία στοιχείων στα σύνολα είναι πιο σημαντική από τη συχνότητα ή τη σειρά τους και μπορεί να χρησιμοποιηθεί για να συγκρίνει την ομοιότητα δύο εγγράφων, λαμβάνοντας υπόψη τα σύνολα λέξεων που εμφανίζονται σε κάθε έγγραφο.

Ο δείκτης Jaccard υπολογίζεται ως εξής:

$$J(A,B) \quad \frac{A \quad B}{A \cup B}$$

όπου A και B είναι σύνολα, και A και B αντιπροσωπεύουν το μέγεθος των συνόλων.

Η τελική τιμή του δείκτη Jaccard κυμαίνεται από 0 έως 1, όπου το 0 υποδηλώνει ότι δεν υπάρχουν κοινά στοιχεία μεταξύ των συνόλων και το 1 υποδηλώνει ότι τα σύνολα είναι πανομοιότυπα. Ο δείκτης Jaccard χρησιμοποιείται ευρέως σε διάφορες εφαρμογές, όπως η ανάκτηση πληροφοριών και είναι ιδιαίτερα χρήσιμος σε αραιά ή υψηλής διάστασης δεδομένα, όπου η παρουσία ή η απουσία χαρακτηριστικών είναι πιο σημαντική από τις πραγματικές τιμές τους.

3.7.6 Απόσταση συνημιτόνου

Η απόσταση συνημιτόνου (cosine distance) (Deza, 2009), γνωστή και ως ομοιότητα συνημιτόνου (cosine similarity) αποτελεί μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους μέτρησης της ομοιότητας μεταξύ διανυσμάτων. Σε αντίθεση με την ευκλείδεια απόσταση, η οποία μετρά το ευθύγραμμο μήκος ανάμεσα σε δύο σημεία και επηρεάζεται από το μέγεθος των διανυσμάτων, η μέθοδος αυτή βασίζεται στη γωνία που σχηματίζουν τα διανύσματα στον διανυσματικό χώρο.

Η ιδέα είναι ότι δύο διανύσματα θεωρούνται παρόμοια αν έχουν παρόμοιο προσανατολισμό, ανεξάρτητα από το μήκος τους. Ο μαθηματικός ορισμός δίνεται από τον τύπο:

$$\text{Sim}(S_a,S_b) \quad \cos\Theta \quad \frac{\overrightarrow{S_a} \quad \overrightarrow{S_b}}{\|S_a\| \, \|S_b\|}$$

Η τιμή του δείκτη κυμαίνεται από −1 έως 1, με το 1 να υποδηλώνει ταυτόσημη κατεύθυνση (τέλεια ομοιότητα), το 0 καμία συσχέτιση και το −1 αντίθετη κατεύθυνση. Στην πράξη, για δεδομένα όπως κείμενα που δεν περιέχουν αρνητικές τιμές, το εύρος περιορίζεται μεταξύ 0 και 1.

Η ομοιότητα συνημιτόνου χρησιμοποιείται εκτενώς στο NLP, ιδίως όταν τα έγγραφα έχουν αναπαρασταθεί με μεθόδους BoW ή TF-IDF, αλλά και σε σύγχρονες αναπαραστάσεις μέσω word embeddings. Η δύναμή της έγκειται στο ότι εστιάζει στο μοτίβο των όρων και όχι στο απόλυτο μήκος του κειμένου. Ένα σύντομο άρθρο και

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

ένα εκτενές κείμενο, μπορεί να χαρακτηριστούν ιδιαίτερα παρόμοια αν χρησιμοποιούν αντίστοιχο λεξιλόγιο και καλύπτουν την ίδια θεματική.

Εν κατακλείδι, η ανάλυση διαφορετικών μεθόδων μέτρησης ομοιότητας κειμένου αναδεικνύει τον καθοριστικό τους ρόλο στο NLP. Από τις κλασικές μεθόδους βασισμένες σε αποστάσεις χαρακτήρων και λέξεων (π.χ. Levenshtein, Hamming, Manhattan) έως τις πιο προηγμένες προσεγγίσεις με ομοιότητα συνημιτόνου, ομοιότητα Jaccard και νευρωνικά μοντέλα με word embeddings, κάθε τεχνική προσφέρει διαφορετικό βαθμό ακρίβειας και υπολογιστικό κόστος.

Η σημασία τους είναι ιδιαίτερα σημαντική στο πλαίσιο της παρούσας έρευνας, όπου γίνεται επεξεργασία μεγάλου όγκου ειδησεογραφικών άρθρων, τα οποία ανανεώνονται τακτικά και εμπλουτίζονται συνεχώς με νέα πληροφορία. Η σωστή επιλογή μεθόδου ομοιότητας επιτρέπει την αναγνώριση διπλότυπων άρθρων, τον εντοπισμό follow-up που συνεχίζουν ή εμπλουτίζουν μία αρχική είδηση και την ομαδοποίηση συναφών άρθρων, βελτιώνοντας την ποιότητα των δεδομένων που τροφοδοτούν τα επόμενα στάδια ανάλυσης, όπως τα embeddings και τα deep learning μοντέλα.

Συνεπώς, η μέτρηση ομοιότητας κειμένων δεν αποτελεί απλώς ένα θεωρητικό εργαλείο, αλλά έναν πρακτικό μηχανισμό που διασφαλίζει την αποτελεσματική διαχείριση και ερμηνεία των δεδομένων. Η εφαρμογή της στο συγκεκριμένο πεδίο ενισχύει την αξιοπιστία της ανάλυσης και συμβάλλει στην πιο ολοκληρωμένη κατανόηση των κοινωνικών κινητοποιήσεων που καταγράφονται στον ευρωπαϊκό χώρο.

3.8 Εισαγωγή στις Βάσεις Δεδομένων

Οι βάσεις δεδομένων (databases) αποτελούν αναπόσπαστο κομμάτι της σύγχρονης πληροφορικής, καθώς επιτρέπουν την οργανωμένη αποθήκευση, τη διαχείριση και την ανάκτηση μεγάλου όγκου πληροφοριών. Μπορούν να εντοπιστούν σε απλές εφαρμογές καθημερινής χρήσης, όπως ηλεκτρονικά καταστήματα και τραπεζικά συστήματα, μέχρι επιστημονική έρευνα και βιοπληροφορική. Όπως γίνεται αντιληπτό, η ανάγκη για αξιόπιστη και αποδοτική διαχείριση δεδομένων είναι πιο έντονη από ποτέ.

Μία βάση δεδομένων μπορεί να οριστεί ως μια συλλογή σχετικών δεδομένων που αποθηκεύονται με τρόπο ώστε να είναι εύκολη η πρόσβαση, η επεξεργασία και η ενημέρωσή τους. Η διαχείρισή τους γίνεται μέσω των Συστημάτων Διαχείρισης Βάσεων Δεδομένων (Database Management System, DBMS), τα οποία προσφέρουν εργαλεία για την οργάνωση των δεδομένων, την ασφάλεια, την ταυτόχρονη πρόσβαση από πολλούς χρήστες και την αποφυγή ασυνεπειών.

Η πιο διαδεδομένη προσέγγιση είναι το σχεσιακό μοντέλο, όπου τα δεδομένα αναπαρίστανται σε πίνακες (relations) και οι σχέσεις μεταξύ τους εκφράζονται μέσω πρωτευόντων και ξένων κλειδιών. Ωστόσο, τα τελευταία χρόνια, η εκρηκτική αύξηση των δεδομένων (Big Data) και την ποικιλία τους, έχουν οδηγήσει στην ανάπτυξη διαφορετικών τύπων Βάσεων Δεδομένων, όπως οι NoSQL, οι οποίες καλύπτουν ανάγκες μεγάλης κλίμακας, ταχύτητας και ευελιξίας.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Συνολικά, οι βάσεις δεδομένων αποτελούν το θεμέλιο για την ανάπτυξη σύγχρονων πληροφοριακών συστημάτων και εφαρμογών, προσφέροντας έναν αξιόπιστο και αποδοτικό τρόπο διαχείρισης της πληροφορίας.

3.8.1 Σχεσιακό μοντέλο δεδομένων

Το σχεσιακό μοντέλο δεδομένων αποτέλεσε για πολλές δεκαετίες την κυρίαρχη επιλογή για την υλοποίηση βάσεων δεδομένων. Η θεμελίωσή του αποδίδεται στον Edgar F. Codd, ο οποίος το 1970 πρότεινε την οργάνωση των δεδομένων βασίζοντας σε μαθηματικές έννοιες από τη θεωρία συνόλων και την κατηγορηματική λογική. Σε αντίθεση με τις προγενέστερες ιεραρχικές ή δικτυακές προσεγγίσεις, το σχεσιακό μοντέλο εισήγαγε μια περισσότερο αφηρημένη και ευέλικτη οπτική, αναπαριστώντας τα δεδομένα σε πίνακες. Κάθε πίνακας αποτελείται από γραμμές (πλειάδες) και στήλες (γνωρίσματα), με τιμές που είναι ατομικές, ώστε να διατηρείται η καθαρότητα και η συνέπεια της βάσης.

Το σχήμα περιγράφει τη λογική δομή μιας σχεσιακής βάσης δεδομένων (πίνακες, γνωρίσματα και σχέσεις μεταξύ τους) και αξιοποιείται από το DBMS, για τη δημιουργία και λειτουργία της. Η αλληλεπίδραση με το σύστημα γίνεται τυποποιημένα μέσω της SQL (Structured Query Language) (Yasin N. Silva, 2016), η οποία επιτρέπει τον ορισμό δομών, τη διαχείριση δεδομένων και την εκτέλεση σύνθετων ερωτημάτων που βασίζονται στη συνδυαστική ισχύ της θεωρίας συνόλων. Τα ώριμα σχεσιακά DBMS (π.χ. MySQL, PostgreSQL, Oracle Database, Microsoft SQL Server) υποστηρίζουν αυτήν τη λειτουργικότητα με αξιοπιστία και συνέπεια.

Η επιτυχία του σχεσιακού μοντέλου βασίζεται στην εγγύηση ακεραιότητας και συνέπειας των δεδομένων, στη σαφή και αυστηρή θεωρητική του βάση, στην τυποποίηση μέσω της SQL, αλλά και στην ωριμότητα των συστημάτων που το υποστηρίζουν. Ωστόσο, οι παραδοσιακές σχεσιακές βάσεις δεδομένων αντιμετωπίζουν σημαντικούς περιορισμούς στη σύγχρονη εποχή της πληροφορίας. Η διαχείριση τεράστιου όγκου δεδομένων, ιδιαίτερα δεδομένων που είναι μη δομημένα ή ημιδομημένα, όπως μεγάλα αρχεία κειμένου, εικόνες, πολυμέσα ή συνεχείς ροές από το διαδίκτυο και τα κοινωνικά μέσα, δεν είναι αποδοτική μέσω του κλασικού σχεσιακού μοντέλου.

Η ανάγκη για αποθήκευση και επεξεργασία δεδομένων μεγάλης κλίμακας, με μεγάλη ποικιλία και υψηλή ταχύτητα, οδήγησε στην ανάπτυξη νέων τεχνολογιών. Η σταδιακή αδυναμία των σχεσιακών βάσεων να ανταποκριθούν στις απαιτήσεις του περιβάλλοντος των Big Data άνοιξε τον δρόμο για την εμφάνιση εναλλακτικών προσεγγίσεων. Σε αυτό το πλαίσιο, εισήχθη ο όρος NoSQL, ο οποίος αρχικά χρησιμοποιήθηκε για να ορίσει συστήματα που δεν βασίζονταν σε σχεσιακές δομές, ενώ αργότερα υιοθετήθηκε με την πιο ευέλικτη έννοια του «Not Only SQL», δηλαδή ως συμπληρωματική και όχι αποκλειστικά ανταγωνιστική λύση σε σχέση με τις σχεσιακές βάσεις δεδομένων.

3.8.2 Μη σχεσιακές βάσεις δεδομένων

Η αδυναμία των παραδοσιακών σχεσιακών βάσεων να διαχειριστούν τον αυξανόμενο όγκο, την ποικιλία και την ταχύτητα των δεδομένων που παράγονται στη σύγχρονη εποχή οδήγησε στην ανάπτυξη εναλλακτικών τεχνολογιών. Για να περιγραφεί αυτή η

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

νέα κατηγορία συστημάτων χρησιμοποιήθηκε ο όρος NoSQL (Strauch, 2024). Ο όρος εμφανίστηκε για πρώτη φορά το 1998 από τον Carlo Strozzi, αναφερόμενος σε βάσεις δεδομένων που δεν χρησιμοποιούσαν SQL. Το 2009 ο Eric Evans επανέφερε τον όρο, δίνοντάς του νέα ώθηση και συσχετίζοντάς τον με την ανάγκη για Βάσεις Δεδομένων ικανές να διαχειρίζονται δεδομένα μεγάλης κλίμακας, σε αντίθεση με τα κλασικά DBMS. Με τον καιρό ο όρος απέκτησε πιο ευέλικτη σημασία, ερμηνευόμενος όχι πλέον ως «No SQL» αλλά ως «Not Only SQL», υποδηλώνοντας ότι οι νέες τεχνολογίες δεν αντικαθιστούν πλήρως το σχεσιακό μοντέλο, αλλά το συμπληρώνουν σε περιβάλλοντα όπου αυτό δεν επαρκεί.

Οι NoSQL βάσεις δεδομένων χαρακτηρίζονται από την ικανότητά τους να αποθηκεύουν και να επεξεργάζονται μη δομημένα και ημιδομημένα δεδομένα, όπως αρχεία κειμένου, πολυμέσα, δεδομένα κοινωνικών μέσων ή logs από ιστοσελίδες. Σε αντίθεση με τις σχεσιακές βάσεις, δεν απαιτούν αυστηρά σχήματα, δεν αποθηκεύουν δεδομένα σε πίνακες και συχνά δεν υποστηρίζουν πλήρως τις κλασικές ιδιότητες ACID (Atomicity, Consistency, Isolation, Durability). Η ευελιξία αυτή επιτρέπει στις NoSQL βάσεις να προσαρμόζονται καλύτερα σε περιβάλλοντα Big Data, όπου ο όγκος και η ετερογένεια των δεδομένων καθιστούν ανεπαρκή τα παραδοσιακά σχήματα.

Ανάλογα με τον τρόπο οργάνωσης των δεδομένων, οι NoSQL βάσεις μπορούν να χωριστούν σε επιμέρους κατηγορίες:

- Document Databases: αποθηκεύουν τα δεδομένα σε μορφή εγγράφων τύπου JSON ή BSON (π.χ. MongoDB, CouchDB).
- Key-Value Stores: οργανώνουν τα δεδομένα ως ζεύγη κλειδιού και τιμής (π.χ. Redis, DynamoDB).
- Column-Oriented Databases: αποθηκεύουν δεδομένα σε στήλες αντί για γραμμές (π.χ. Cassandra, HBase).
- Graph Databases: εστιάζουν σε δεδομένα που περιγράφουν σχέσεις και δίκτυα (π.χ. Neo4).

Ανάμεσα στις παραπάνω κατηγορίες, οι Document Databases έχουν αποκτήσει ιδιαίτερη σημασία, καθώς υποστηρίζουν εγγενώς τη δομή JSON, η οποία αποτελεί τυποποιημένο format για την αποθήκευση και ανταλλαγή δεδομένων στο διαδίκτυο. Στην κατηγορία αυτή ανήκει η MongoDB (Strauch, 2024), η οποία κυκλοφόρησε αρχικά το 2009 και έκτοτε έχει καθιερωθεί ως μία από τις πιο δημοφιλείς NoSQL βάσεις δεδομένων. Η MongoDB είναι υλοποιημένη σε C και βασίζεται στη λογική των συλλογών (collections) και των εγγράφων (documents). Κάθε συλλογή μπορεί να περιέχει πλήθος εγγράφων χωρίς προκαθορισμένο σχήμα, ενώ τα δεδομένα αποθηκεύονται σε μορφή BSON, δηλαδή μια δυαδική επέκταση του JSON που προσφέρει καλύτερη απόδοση και επιπλέον τύπους δεδομένων.

Η διαφορά της MongoDB από τις κλασικές σχεσιακές βάσεις, όπως η MySQL, είναι θεμελιώδης. Ενώ οι σχεσιακές βάσεις χαρακτηρίζονται από στατικά και αυστηρά σχήματα, η MongoDB επιτρέπει τη χρήση δυναμικών σχημάτων, δίνοντας τη δυνατότητα εισαγωγής δεδομένων με διαφορετικά πεδία στην ίδια συλλογή. Επιπλέον, προσφέρει βελτιωμένη απόδοση στις τέσσερις βασικές λειτουργίες βάσεων δεδομένων (Εισαγωγή, Επιλογή, Ενημέρωση, Διαγραφή), γεγονός που την καθιστά

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

κατάλληλη για εφαρμογές όπου οι όγκοι δεδομένων είναι τεράστιοι και οι χρόνοι απόκρισης κρίσιμοι.

Συνοψίζοντας, οι μη σχεσιακές βάσεις δεδομένων αναδείχθηκαν ως απάντηση στις προκλήσεις της εποχής των Big Data, καλύπτοντας ανάγκες που οι παραδοσιακές σχεσιακές βάσεις δεν μπορούσαν να ικανοποιήσουν. Η ευελιξία τους ως προς το σχήμα, η δυνατότητα οριζόντιας κλιμάκωσης, η υποστήριξη για ημιδομημένα και μη δομημένα δεδομένα και η υψηλή απόδοση σε εφαρμογές μεγάλης κλίμακας αποτελούν βασικά πλεονεκτήματα που τις καθιστούν αναγκαίες σε πληθώρα σύγχρονων έργων. Παρόλο που οι σχεσιακές βάσεις εξακολουθούν να έχουν κυρίαρχη θέση σε εφαρμογές όπου απαιτείται αυστηρή ακεραιότητα και συναλλακτική συνέπεια, οι NoSQL βάσεις έχουν εδραιωθεί ως συμπληρωματική τεχνολογία, προσφέροντας νέες δυνατότητες στη διαχείριση δεδομένων. Έτσι, η επιλογή μεταξύ SQL και NoSQL δεν είναι απόλυτη, αλλά εξαρτάται από τη φύση των δεδομένων, τις απαιτήσεις της εφαρμογής και τις προτεραιότητες του εκάστοτε έργου.

3.8.3 Σύνδεση των Άρθρων σε Story Threads

Υπάρχει μεγάλο ενδιαφέρον καθημερινής ενημέρωσης, τόσο από τους πολίτες όσο και από τους αναλυτές ειδήσεων, για τα νέα που αναρτούνται στις ειδήσεις. Όμως σε συνάρτηση του τεράστιου όγκου πληροφοριών που φτάνει καθημερινά το καθιστά δύσκολο και χρονοβόρο. Ως εκ τούτου, υπάρχει αυξανόμενη ανάγκη για αυτόματες τεχνικές οργάνωσης ειδήσεων με τρόπο που θα βοηθά τους χρήστες να τις ερμηνεύουν και να τις αναλύουν πιο γρήγορα. Αυτό το πρόβλημα αντιμετωπίζεται από ένα ερευνητικό πρόγραμμα που ονομάζεται Ανίχνευση και Παρακολούθηση Θεμάτων (Topic Detection and Tracking, TDT) (Allan, 2000), που διεξάγει ετήσιους διαγωνισμούς για την ανάπτυξη τεχνικών οργάνωσης ειδήσεων.

Ένα από τα μειονεκτήματα της τρέχουσας αξιολόγησης του TDT είναι ότι αντιμετωπίζει τα ειδησεογραφικά θέματα ως επίπεδες συλλογές ιστοριών. Για παράδειγμα, το έργο ανίχνευσης του TDT είναι να οργανώνει μια συλλογή ειδησεογραφικών ιστοριών σε συστάδες θεμάτων. Ωστόσο, ένα θέμα στις ειδήσεις είναι κάτι περισσότερο από μια απλή συλλογή ιστοριών και χαρακτηρίζεται από μια συγκεκριμένη δομή αλληλοσχετιζόμενων γεγονότων. Αυτό πράγματι αναγνωρίζεται από το TDT, το οποίο ορίζει το θέμα ως ένα σύνολο ειδήσεων που συνδέονται στενά από κάποιο καθοριστικό πραγματικό γεγονός, όπου γεγονός ορίζεται ως κάτι που συμβαίνει σε συγκεκριμένο χρόνο και τόπο (Allan, 2000). Για παράδειγμα, όταν μια βόμβα εκρήγνυται σε ένα κτήριο, αυτό είναι το καθοριστικό γεγονός που ενεργοποιεί το θέμα. Άλλα γεγονότα στο θέμα μπορεί να περιλαμβάνουν τις προσπάθειες διάσωσης, την αναζήτηση των δραστών, συλλήψεις και δίκες κ.ο.κ. Βλέπουμε ότι υπάρχει ένα μοτίβο εξαρτήσεων μεταξύ ζευγών γεγονότων μέσα στο θέμα. Στο παραπάνω παράδειγμα, το γεγονός των προσπαθειών διάσωσης επηρεάζεται από το γεγονός της έκρηξης της βόμβας και το ίδιο ισχύει για το γεγονός της αναζήτησης των δραστών.

3.8.3.1 Σχετικές εργασίες

Στην εργασία των Nallapati et al. (Nallapati, 2004), όπου εισάγεται η έννοια του event threading για την ανάλυση ειδήσεων, σε αντίθεση με τις παραδοσιακές

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

μεθόδους του TDT, οι συγγραφείς προτείνουν την αναγνώριση επιμέρους γεγονότων μέσα σε ένα θέμα και την αποτύπωση των εξαρτήσεων μεταξύ τους (αιτιότητα ή χρονική ακολουθία). Για την ομαδοποίηση των άρθρων σε γεγονότα εφαρμόζουν τεχνικές συσταδοποίησης βασισμένες σε χαρακτηριστικά όπως η ομοιότητα TF-IDF, οι οντότητες (πρόσωπα, τοποθεσίες) και η χρονική εγγύτητα, ενώ για τη μοντελοποίηση των εξαρτήσεων αξιοποιούν μοντέλα που συνδυάζουν χρονική σειρά και σημασιολογική ομοιότητα. Μάλιστα, προτείνεται ένα απλό μοντέλο simple threshold όπου, αν η συνημιτονοειδής ομοιότητα δύο άρθρων υπερβαίνει ένα κατώφλι, τότε το προγενέστερο άρθρο χαρακτηρίζεται ως γονέας (parent) του μεταγενέστερου. Έτσι δημιουργούνται ζεύγη parent–child ειδήσεων βάσει χρονικής σειράς, αποτυπώνοντας τη ροή ενός γεγονότος στο χρόνο. Τα πειράματά τους σε επισημειωμένα (annotated) σύνολα δεδομένων, δείχνουν ότι η αξιοποίηση της χρονικής διάστασης σε συνδυασμό με την κειμενική ομοιότητα βελτιώνει σημαντικά την ακρίβεια της ομαδοποίησης και καθιστά εφικτή την αναπαράσταση της δομής ενός θέματος με τη μορφή διασυνδεδεμένων γεγονότων.

Οι Liu et al. (Liu B. N., 2017) ανέπτυσσαν ένα ολοκληρωμένο σύστημα που οργανώνει ειδήσεις σε δέντρα ιστοριών (story trees) για την παρακολούθηση της εξέλιξης γεγονότων. Αρχικά, γίνεται ομαδοποίηση ειδησεογραφικών άρθρων σε γεγονότα (clusters) χρησιμοποιώντας πληθώρα χαρακτηριστικών, ανάμεσα στα οποία και τη συνημιτονοειδή ομοιότητα TF-IDF ως μέτρο ομοιότητας περιεχομένου. Στη συνέχεια, τα ανακαλυφθέντα events συνδέονται μεταξύ τους με κατευθυνόμενες ακμές ώστε να σχηματίσουν δέντρα ιστοριών – κάθε σύνδεση μεταξύ δύο γεγονότων υποδηλώνει χρονική συνέχεια ή λογική συσχέτιση (π.χ. εξέλιξη μιας είδησης). Το Story Forest χρησιμοποιεί thresholding και ταξινόμηση ζευγών εγγράφων για να αποφασίσει ποιες ειδήσεις αφορούν το ίδιο γεγονός ή διαδοχικά γεγονότα· για παράδειγμα, υπολογίζει τη συνημιτονοειδή ομοιότητα TF-IDF μεταξύ ενός άρθρου και ενός keyword community και φιλτράρει με ένα κατώφλι δ για να βρει σχετικές ειδήσεις. Τελικά, παράγει μια δενδρική δομή όπου κάθε κόμβος είναι ένα γεγονός και οι ακμές parent-child δείχνουν την εξέλιξη της ιστορίας στο χρόνο.

Επίσης, οι Laban et al. (Laban, 2017) προτείναν το σύστημα newsLens για τη δημιουργία ιστοριών από μεγάλες συλλογές ειδήσεων. Στο πρώτο στάδιο τα άρθρα ομαδοποιούνται σε clusters βάσει χρονικών παραθύρων (π.χ. ημέρα ή εβδομάδα), ενώ τυχόν διπλότυπα αφαιρούνται με έλεγχο συνημιτονοειδούς ομοιότητας, διατηρώντας το πιο πρόσφατο άρθρο. Κάθε cluster αναπαρίσταται με BoW διάνυσμα, και κάθε φορά που προκύπτει ένα νέο cluster, υπολογίζεται η συνημιτονοειδής ομοιότητα με προηγούμενα. Αν η συνημιτονοειδής ομοιότητα υπερβαίνει ένα κατώφλι και τα δύο clusters δεν επικαλύπτονται χρονικά, τότε θεωρούνται μέρη της ίδιας ιστορίας και ενώνονται. Με υψηλό κατώφλι ομοιότητας, εξασφαλίζεται ότι μόνο πολύ συναφή θεματικά άρθρα συνδέονται, αποφεύγοντας λάθος συγχωνεύσεις. Το αποτέλεσμα είναι threads ειδήσεων που απεικονίζουν την εξέλιξη ενός γεγονότος (π.χ. η ιστορία της πτώσης μιας πτήσης και οι μεταγενέστερες εξελίξεις της, όπως ευρήματα σε μεταγενέστερες ημερομηνίες, συγχωνεύονται σε ένα συνεκτικό story).

Η πιο πρόσφατη εργασία, είναι αυτοί των Zhang et al. (Zhang, 2023) και εστιάζει στην εξαγωγή αιτιακών σχέσεων μεταξύ ειδησεογραφικών άρθρων, αξιοποιώντας μια προκαθορισμένη δομή story tree. Το story tree οργανώνει τα άρθρα ενός θέματος σε δομή parent–child με χρονολογική διάταξη, παρόμοια με το event threading. Σε αντίθεση με προσεγγίσεις που βασίζονται αποκλειστικά σε ομοιότητα περιεχομένου,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

οι συγγραφείς ενσωματώνουν τη δομή του story tree σε μοντέλο ακέραιου γραμμικού προγραμματισμού (ILP), εισάγοντας περιορισμούς που αντανακλούν τις χρονικές σχέσεις parent–child. Με αυτόν τον τρόπο, το μοντέλο αξιοποιεί τη ροή των γεγονότων (δηλαδή ποιο άρθρο προηγήθηκε και ποιο ακολούθησε), επιτυγχάνοντας βελτίωση άνω του 5% σε F1-score έναντι των καλύτερων ταξινομητών. Το αποτέλεσμα αναδεικνύει τη σημασία των χρονικών και δομικών σχέσεων για την κατανόηση πιο σύνθετων διασυνδέσεων μεταξύ άρθρων.

Κλείνοντας, τονίζεται ότι η ενσωμάτωση της έννοιας των story threads στην παρούσα εργασία αποτελεί ένα ουσιαστικό βήμα προς την αυτοματοποιημένη οργάνωση, ανάλυση και παρακολούθηση ειδησεογραφικού περιεχομένου σε μεγάλης κλίμακας ροές δεδομένων. Με δεδομένο τον αυξανόμενο όγκο ειδήσεων που δημοσιεύονται καθημερινά, η ανάγκη για την δόμηση και επεξεργασία της πληροφορίας καθίσταται επιτακτική. Η προτεινόμενη μεθοδολογία, μέσω της δημιουργίας αλυσίδων άρθρων που συνδέονται ιεραρχικά βάσει θεματικής συνάφειας, χρονικής ακολουθίας και γεωγραφικής συνοχής, συμβάλλει στη διαμόρφωση μιας ολοκληρωμένης και δυναμικής εικόνας της εξέλιξης ενός γεγονότος.

Σε αντίθεση με τις παραδοσιακές προσεγγίσεις ομαδοποίησης ειδήσεων, που αντιμετωπίζουν τα θέματα ως απλές συλλογές ασύνδετων άρθρων, το προτεινόμενο σύστημα αξιοποιεί ιεραρχικές σχέσεις τύπου parent–child. Μέσω αυτής της δομής, καθίσταται εφικτή η αναπαράσταση της πορείας ενός γεγονότος στο χρόνο, καθώς και η διάκριση μεταξύ νέας πληροφορίας, επικαιροποιήσεων (follow-ups) και αναδημοσιεύσεων (duplicates). Έτσι, οι χρήστες αποκτούν πρόσβαση σε μια συνεκτική, χρονικά και θεματικά ενοποιημένη αφήγηση, διευκολύνοντας τόσο την κατανόηση των γεγονότων όσο και την ανάλυση κοινωνικών, πολιτικών και οικονομικών κινητοποιήσεων.

Παράλληλα, η χρήση τεχνικών όπως η TF-IDF αναπαράσταση κειμένου, η μέτρηση ομοιότητας μέσω cosine similarity, η εξαγωγή τοπωνυμίων και η αξιοποίηση του GeoNames API παρέχουν μια υψηλού επιπέδου υποδομή NLP, ενισχύοντας την ακρίβεια και την αξιοπιστία του συστήματος. Επιπλέον, η αποθήκευση των δεδομένων και των μεταδεδομένων σε βάση MongoDB επιτρέπει την αποδοτική διαχείριση και επεκτασιμότητα του μοντέλου, καθιστώντας το κατάλληλο για εφαρμογή σε μεγάλης κλίμακας ειδησεογραφικά οικοσυστήματα.

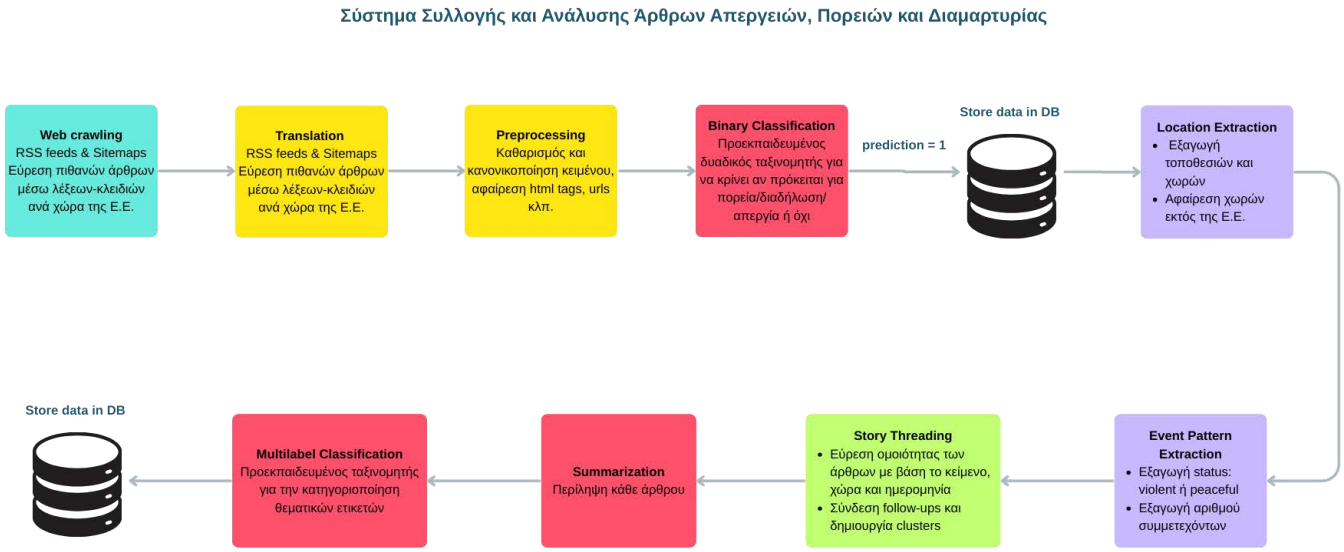
Συνολικά, η εφαρμογή των story threads προσφέρει μια πρωτοποριακή προσέγγιση στην οργάνωση ειδησεογραφικού περιεχομένου, η οποία συμβάλλει ουσιαστικά στην αυτόματη ανάλυση της εξέλιξης γεγονότων και ενισχύει τη δυνατότητα διαχρονικής παρακολούθησης κοινωνικών δράσεων. Μελλοντικά, η παρούσα μεθοδολογία μπορεί να επεκταθεί με την ενσωμάτωση προηγμένων μοντέλων DL για την αναγνώριση αιτιακών σχέσεων μεταξύ άρθρων, τη βελτιστοποίηση της αναπαράστασης γεγονότων και τη δημιουργία ακόμα πιο εξελιγμένων χρονοσειρών πληροφορίας.

Με αυτόν τον τρόπο, η εργασία θέτει τα θεμέλια για ευφυή συστήματα ανάλυσης ειδήσεων, που όχι μόνο οργανώνουν την πληροφορία, αλλά και παρέχουν πολύτιμες γνώσεις για τη δυναμική των κοινωνικών κινητοποιήσεων και των γεγονότων που τις διαμορφώνουν.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

4. Μεθοδολογία και Υλοποίηση της Βάσης Δεδομένων Διαμαρτυριών

4.1 Συνοπτική περιγραφή του συστήματος



Εικόνα 10 Απεικόνιση Ολοκληρωμένου Συστήματος Ανάλυσης Απεργειών, Πορειών και Διαδηλώσεων.

Στην Εικόνα 10 απεικονίζεται το Εικόνα 1σύστημα συλλογής και ανάλυσης ειδησεογραφικών άρθρων απεργιών, πορειών και διαδηλώσεων και έχει σχεδιαστεί με στόχο την αυτόματη επεξεργασία ειδησεογραφικού περιεχομένου που αφορά χώρες της Ευρωπαϊκής Ένωσης, επιτρέποντας την εξαγωγή, ταξινόμηση και ομαδοποίηση πληροφοριών σχετικών με κοινωνικές κινητοποιήσεις. Η διαδικασία ξεκινά με το web crawling, συλλέγοντας άρθρα από RSS feeds και sitemaps ειδησεογραφικών ιστοσελίδων, με βάση λέξεις-κλειδιά που έχουν οριστεί ανά γλώσσα των χωρών της Ε.Ε. Στη συνέχεια, εφαρμόζεται μετάφραση και κανονικοποίηση των κειμένων αποκλειστικά στα άρθρα που περιέχουν προκαθορισμένες λέξεις-κλειδιά, ώστε το περιεχόμενο να επεξεργάζεται σε μία ενιαία γλώσσα, τα Αγγλικά.

Μετά τη μετάφραση, ακολουθεί το βήμα Binary Classificattion, όπου χρησιμοποιείται προεκπαιδευμένος δυαδικός ταξινομητής για να αξιολογήσει αν το άρθρο αφορά πορεία, διαδήλωση ή απεργία. Τα άρθρα που αναγνωρίζονται ως σχετικά αποθηκεύονται στη βάση δεδομένων και πραγματοποιείται εξαγωγή τοποθεσιών και χωρών από το περιεχόμενο του άρθρου, για να διασφαλιστεί ότι η ανάλυση είναι εστιαζόμενη σε ευρωπαϊκά δεδομένα.

Στο επόμενο στάδιο, εφαρμόζεται Event Pattern Extraction, όπου εντοπίζονται κρίσιμα χαρακτηριστικά του γεγονότος, όπως το αν η δράση χαρακτηρίζεται ως ειρηνική ή βίαιη, καθώς και ο εκτιμώμενος αριθμός συμμετεχόντων. Τα αποτελέσματα αυτά συνδυάζονται με τη διαδικασία story threading, η οποία αξιολογεί την ομοιότητα άρθρων με βάση το περιεχόμενο, τη χώρα και την ημερομηνία, επιτρέποντας την ομαδοποίηση σχετικών αναφορών, τον εντοπισμό των διπλότυπων, follow-up άρθρων και τη σύνδεση πολλαπλών άρθρων γύρω από το ίδιο περιστατικό. Παράλληλα, υλοποιείται συνοπτική περίληψη κάθε άρθρου, ώστε να διευκολύνεται η κατανόηση του περιεχομένου χωρίς την ανάγκη πλήρους ανάγνωσης.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Τέλος, πραγματοποιείται πολυκατηγορική ταξινόμηση μέσω προεκπαιδευμένου μοντέλου, το οποίο αποδίδει θεματικές ετικέτες στις δημόσιες κινητοποιήσεις, όπως πολιτική, οικονομία ή εκπαίδευση. Όλες οι νέες επεξεργασμένες πληροφορίες αποθηκεύονται στη βάση δεδομένων, διαμορφώνοντας ένα ολοκληρωμένο, οργανωμένο και αναλυτικό σύνολο δεδομένων, κατάλληλο για περαιτέρω μελέτη και αξιοποίηση στο πλαίσιο της ανάλυσης κοινωνικών κινητοποιήσεων στην Ευρωπαϊκή Ένωση.

4.2 Το πρώτο βήμα: Web Scraping

Στο πλαίσιο της παρούσας πτυχιακής εργασίας, για την τήρηση των Όρων Χρήσης κάθε ιστότοπου πριν το web scraping προηγήθηκε αξιοποιήθηκε η ανάλυση των robots.txt από εφημερίδες 27 κρατών-μελών της ΕΕ.

Ο παρακάτω πίνακας συνοψίζει τα βασικά ευρήματα:

Χώρα	Ειδησιογραφική Πηγή	Περιορισμοί (Disallow)	Απαγορευμένοι User-Agents	Παροχή Sitemap/RSS	Σχόλιο
Malta	Times of Malta	`/archive/`, `/login`, `/register`, `/download`, `/search`	GPTBot (ολική απαγόρευση)	Ναι	Το RSS feed είναι επιτρεπτό για scraping.
Poland	Rzeczpospolita	Search, previews, technical URLs	–	Ναι	Καθαρή χρήση sitemaps για άρθρα.
Finland	Helsingin Sanomat	`/promo/`, `/api/`, `/rest/`, search	GPTBot, ClaudeBot, PerplexityBot, κ.ά.	Ναι	Επιτρέπεται crawling μόνο με ουδέτερο User-Agent.
Croatia	24sata	Accounts, search, previews, comments	–	Ναι	Άρθρα και RSS πλήρως προσβάσιμα.
Denmark	BT.dk	`/api`, `/login`, `/logout`	GPTBot, CCBot	Ναι	Ορίζει crawl-delay 10s.
Estonia	Postimees	`/search*`, `/feed/*`, `/comments`	GPTBot, ClaudeBot, PerplexityBot, κ.ά.	Ναι	Πολύ αυστηρό crawl-delay: 60s/request.
Bulgaria	Standart News	`/search/`, `/cdn-cgi/`	–	Ναι	Απλό robots.txt άρθρα διαθέσιμα.
Belgium	La Libre Belgique	`/preview/`, `/pf/api/*`, `/redirect/`	GPTBot, ClaudeBot, PerplexityBot, Google-Extended κ.ά.	Ναι	Blacklist σε AI/SEO bots.
Netherlands	Volkskrant	`/auth/`, `/search?query=*`,	GPTBot, Claude, CCBot,	Ναι	Ρητή απαγόρευση

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Χώρα	Ειδησιογραφική Πηγή	Περιορισμοί (Disallow)	Απαγορευμένοι User-Agents	Παροχή Sitemap/RSS	Σχόλιο
		APIs	Perplexity, Google-Extended, κ.ά.		scraping για εμπορική.
Lithuania	Veidas	`/admin/`, `/comment/`, `/node/add/`	–	Ναι	Τα άρθρα επιτρέπονται, αλλά admin/σχόλια όχι.
Czech Republic	Blesk	Search, login, captcha, exports	Ahrefs, Semrush, Yandex, Bytespider κ.ά.	Ναι	Στόχος προστασία από SEO crawlers.
Romania	Realitatea	Polls, APIs, search, covid maps	–	Ναι	Τα άρθρα είναι πλήρως προσβάσιμα μέσω sitemap.
Luxembourg	L'Essentiel	`/cmd/`, `.json`, `/comment/`	GPTBot, Claude, Perplexity, Applebot κ.ά.	Ναι	AI bots αποκλείονται από άρθρα/βίντεο.
Germany	Die Welt	`/suche`, `/api/`, previews	GPTBot, Claude, Perplexity, Ahrefs κ.ά.	Ναι	Άρθρα μέσω sitemap SEO bots αποκλείονται.
Austria	Krone	Search, forum, previews, games	–	Ναι	Ρητή νομική απαγόρευση scraping για TDM.
Greece	Τα Νέα	`/wp-admin`, `/wp-includes`	–	Ναι	Ρητή απαγόρευση scraping για εμπορική.
Italy	La Repubblica	Πολλά paths (search, multimedia, παλιό αρχείο)	GPTBot, Claude, Perplexity, Ahrefs, Semrush κ.ά.	Ναι	Αυστηρό: SEO bots πλήρως αποκλεισμένα.
France	Le Monde	Search, previews, APIs, multimedia	GPTBot, Claude, Perplexity, Ahrefs, Semrush, Archive.org κ.ά.	Ναι	Ρητή απαγόρευση scraping για εμπορική.
Portugal	Público	`/pesquisa?*` (search)	GPTBot, Claude, Perplexity, Semrush, Ahrefs, Scrapy	Ναι	Κανονικά άρθρα μέσω sitemap· AI/SEO bots αποκλείονται.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Χώρα	Ειδησιογραφική Πηγή	Περιορισμοί (Disallow)	Απαγορευμένοι User-Agents	Παροχή Sitemap/RSS	Σχόλιο
Spain	El País	`/pruebas/`, `/publicidad/`, previews	Ahrefs, Semrush, Scrapy, GPTBots κ.ά.	Όχι εμφανές	Πολύ μεγάλη blacklist scrapers.
Ireland	The Irish Times	`/search/`, `/poll/`, `/archive/`	GPTBot, ChatGPT-User	Ναι	Χρήση sitemaps· AI bots αποκλείονται.
Hungary	Nepszava	`/digitalisujsag`, `/json/*`	–	Ναι	Τα άρθρα προορίζονται για crawling, όχι η ψηφιακή έκδοση.
Slovakia	SME.sk	`/poll/vote/`	–	Ναι	Από τα πιο χαλαρά robots.txt.
Sweden	Göteborgs-Posten	– (μόνο blacklist bots)	GPTBot, ChatGPT-User, Google-Extended, CCBot, Yandex, Baidu, PetalBot	–	Μπλοκάρονται μόνο AI/SEO bots.
Cyprus	Philenews	Feeds, search, author pages	–	Ναι	Ρητή απαγόρευση scraping για εμπορική.
Latvia	Diena	`/admin`	–	–	Εξαιρετικά χαλαρό robots.txt.

Πίνακας 3 Περιορισμοί πρόσβασης που ορίζονται στα αρχείων robots.txt ευρωπαϊκών εφημερίδων.

Πίνακας 3, αποτελεί την ανάλυση των αρχείων robots.txt και των δηλωμένων πολιτικών πρόσβασης σε ειδησεογραφικούς ιστοτόπους 27 κρατών-μελών καταδεικνύει ότι το δημοσιογραφικό περιεχόμενο επιτρέπεται, κατά κανόνα, να ανιχνεύεται μέσω RSS και sitemaps, ενώ οι σελίδες αναζήτησης, οι διεπαφές σύνδεσης/προεπισκόπησης και τα δημόσια APIs αποκλείονται σχεδόν καθολικά. Παράλληλα, πολλοί εκδοτικοί οργανισμοί εφαρμόζουν καταλόγους αποκλεισμού (blacklists) συγκεκριμένων User-Agents, με έμφαση σε αυτόνομους πράκτορες AI(π.χ. GPTBot, Claude, Perplexity) και εργαλεία SEO (π.χ. Ahrefs, Semrush). Σε ορισμένες χώρες, όπως η Ολλανδία, η Αυστρία, η Γαλλία και η Ιταλία, διαπιστώνεται ρητή νομική απαγόρευση scraping για εμπορικούς σκοπούς, σε συνάρτηση με το κανονιστικό πλαίσιο της Οδηγίας (ΕΕ) 2019/790. Η συλλογή δεδομένων σε αυτή τη μελέτη επικεντρώθηκε αποκλειστικά σε επιτρεπόμενα RSS feeds και sitemaps, με συστηματική παράκαμψη διαδρομών με Disallow και User-Agents που περιλαμβάνονται στις λίστες αποκλεισμού. Με αυτόν τον τρόπο, η

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

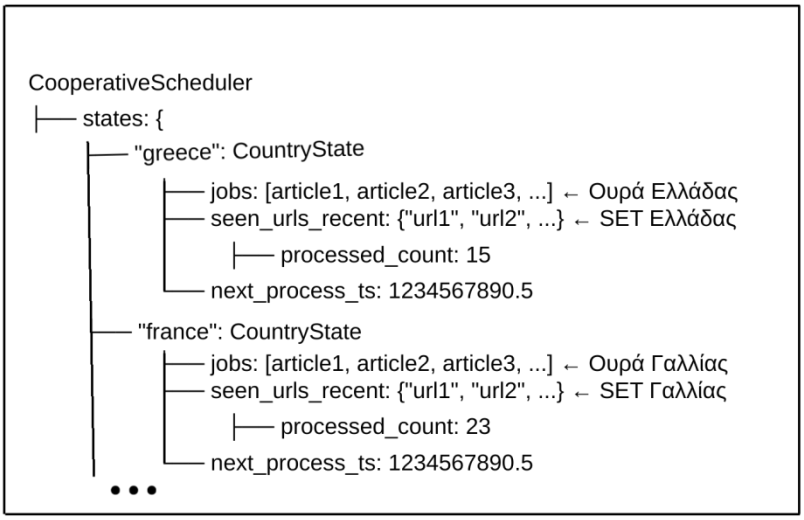
εξαγωγή και συλλογή δεδομένων ήταν πλήρως σύμφωνη με τους τεχνικούς κανόνες και τους όρους χρήσης των ευρωπαϊκών ειδησεογραφικών ιστοσελίδων.

Υλοποιήθηκε ένας μηχανισμός συλλογής ειδησεογραφικού περιεχομένου από ιστότοπους της Ε.Ε. με cooperative (συνεργατικό) χρονοπρογραμματισμό με ένα νήμα εκτέλεσης (thread). Ένας κεντρικός προγραμματιστής, scheduler, εξυπηρετεί πολλαπλές χώρες κυκλικά με την μέθοδο round-robin, προγραμματίζοντας πότε θα γίνει νέο RSS/Sitemap fetch και πότε θα επεξεργαστεί το επόμενο άρθρο για κάθε χώρα. Σε προκαθορισμένα διαστήματα ανά χώρα, ο scheduler ανακτά μαζικά τα πρόσφατα items από το αντίστοιχο RSS ή sitemap και τα εντάσσει στην ουρά της κάθε χώρας. Για κάθε νέο item, ελέγχεται αν το URL υπάρχει ήδη στην ουρά, ώστε να μην εισάγονται διπλότυπα και να αποτρέπεται η επανεπεξεργασία του ίδιου περιεχομένου. Όταν έρθει η σειρά μιας χώρας, ο scheduler ανασύρει το πρώτο στοιχείο από την ουρά της και το υποβάλλει στην αντίστοιχη διαδικασία επεξεργασίας. Μετά την ολοκλήρωση, ενημερώνεται η επόμενη χρονική στιγμή εξυπηρέτησης της χώρας σύμφωνα με τις καθορισμένες παραμέτρους χρονισμού.

Σε κάθε κύκλο, ο scheduler δίνει προτεραιότητα στη χώρα με το μεγαλύτερο crawl-delay, την εξυπηρετεί πρώτη και κατόπιν εισέρχεται σε αναμονή, ενώ παράλληλα επεξεργάζεται τις ουρές των υπολοίπων χωρών. Δεδομένου ότι —με εξαίρεση δύο εκδότες— οι υπόλοιποι εφαρμόζουν παρόμοιο crawl-delay, ως δευτερεύον κριτήριο επιλογής χρησιμοποιείται το μήκος της ουράς, ώστε να μειώνεται ταχύτερα ο συνολικός όγκος των εκκρεμοτήτων και να διατηρείται υψηλή αξιοποίηση πόρων. Ο scheduler ολοκληρώνει όταν κάθε χώρα έχει πραγματοποιήσει τουλάχιστον ένα fetch και όλες οι ουρές έχουν εκκενωθεί, δηλαδή δεν εκκρεμεί κανένα άρθρο προς επεξεργασία.

Τα HTTP αιτήματα υπόκεινται σε throttling: τηρείται ελάχιστο μεσοδιάστημα ανά domain, εφαρμόζονται όρια ταυτόχρονων συνδέσεων, ενώ σε σφάλματα 429/503 εφαρμόζεται εκθετική οπισθοχώρηση (exponential backoff) με ανώτατο αριθμό επαναπροσπαθειών. Το exponential backoff είναι στρατηγική επαναπροσπαθειών όπου ο χρόνος αναμονής διπλασιάζεται μετά από κάθε αποτυχημένο αίτημα, μέχρι ένα ανώτατο όριο. Η χρήση κοινόχρηστου Session επιτρέπει επαναχρησιμοποίηση TCP/TLS συνδέσεων (keep-alive/connection pooling), ενοποιημένη διαχείριση headers/cookies, ομοιόμορφες πολιτικές timeouts/retries και έλεγχο ανοικτών sockets, μειώνοντας handshakes και το μεταβαλλόμενο δικτυακό φορτίο.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 11 Οπτικοποίηση λειτουργίας Cooperative Sheduler

Σε επίπεδο περιεχομένου, κάθε νέο άρθρο κανονικοποιείται (μεταδεδομένα και σώμα), ελέγχεται και εντάσσεται στην ουρά επεξεργασίας. Το pipeline περιλαμβάνει πολυγλωσσική ανίχνευση δεικτών strike/protest (regex), αυτόματη μετάφραση τίτλου, περίληψης και κειμένου στα αγγλικά και σύνοψη, οι οντότητες αποθηκεύονται στις αντίστοιχες συλλογές MongoDB. Η αρχιτεκτονική αυτή ευθυγραμμίζεται με τις αρχές του web crawling και διασφαλίζει δίκαιη κατανομή πόρων, σταθερό ρυθμό λειτουργίας και ελαχιστοποίηση του δικτυακού αποτυπώματος, με περιοδική εκκαθάριση των ουρών και συνεχή, ομαλή λειτουργία.

Το web crawling χρησιμοποιήθηκε για δύο βασικούς σκοπούς. Ο πρώτος ήταν η συλλογή δεδομένων προκειμένου να δημιουργηθεί το corpus που χρησιμοποιήθηκε για την επεξεργασία και ανάλυση γεγονότων, αποτελώντας θεμέλιο στάδιο της συνολικής μεθοδολογίας. Ο δεύτερος σκοπός ήταν για τη συνεχή ενημέρωση της βάσης δεδομένων με νέα άρθρα, επιτρέποντας τη δυναμική και σε πραγματικό χρόνο επικαιροποίηση του συστήματος.

Για τον σκοπό αυτό, αναπτύχθηκε ένας crawler που συγκέντρωνε άρθρα από όλες τις χώρες της Ευρωπαϊκής Ένωσης, με έμφαση σε περιεχόμενο που σχετίζεται με κοινωνικές δράσεις (απεργίες, διαδηλώσεις, πορείες). Η διαδικασία βασίστηκε στη χρήση RSS feeds και Sitemaps ειδησεογραφικών ιστοτόπων, τα οποία εξασφαλίζουν πρόσβαση σε ενημερωμένο και δομημένο περιεχόμενο.

Η αρχιτεκτονική του crawler περιλάμβανε τα ακόλουθα στάδια:

- Εξαγωγή μεταδεδομένων (metadata), όπως τίτλος άρθρου, ημερομηνία δημοσίευσης και τελευταίας τροποποίησης, κατηγορίες, γλώσσα και όνομα της ηλεκτρονικής εφημερίδας, μέσω parsing των στοιχείων `<item>` των RSS feeds και των <url> των sitemaps.
- Ανάκτηση πλήρους κειμένου και περίληψης του άρθρου μέσω HTTP GET requests και parsing της κύριας σελίδας, με χρήση επιλεκτικών CSS selectors.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Protest: 1778	March: 207	Occupation: 61
Demonstrate: 1324	Rally: 164	Stoppage: 7
Strike: 690	Mobilize: 79	Picket: 2
Activism: 31	Riot: 66	Walkout:1
Gather: 246	Boycott:64	N/A

Πίνακας 4 Πλήθος εμφανίσεων ανά λέξη κλειδί.

Για τον εντοπισμό άρθρων που περιέγραφαν γεγονότα κοινωνικής κινητοποίησης εφαρμόστηκε πολυγλωσσικό λεξικό λέξεων-κλειδιών, το οποίο φαίνεται στον Πίνακας 4. Οι λέξεις strike, walkout, picket αναφέρονται σε απεργία και οι υπόλοιπες σε διαδήλωση.

Για κάθε γλώσσα της Ευρωπαϊκής Ένωσης δημιουργήθηκε λίστα αντίστοιχων όρων. Για παράδειγμα, στα Ισπανικά χρησιμοποιήθηκαν οι όροι protesta, huelga, ενώ στα Γαλλικά οι όροι manifestation, grève. Με αυτόν τον τρόπο έγινε εφικτή η ανίχνευση άρθρων σε διαφορετικές ευρωπαϊκές γλώσσες. Στη συνέχεια, τα κείμενα μεταφράστηκαν στα Αγγλικά με σκοπό την ενοποίηση και τη διευκόλυνση της περαιτέρω ανάλυσης.

Το web scraping αποτέλεσε τη βάση για τη συστηματική δημιουργία και δυναμική συντήρηση ενός πολυγλωσσικού και αντιπροσωπευτικού συνόλου ειδησεογραφικού περιεχομένου. Η μεθοδολογία που υιοθετήθηκε συνδυάζει τεχνική αποδοτικότητα και ρυθμιστική συμμόρφωση, αξιοποιώντας εγκεκριμένες πηγές δεδομένων, όπως RSS feeds και sitemaps, και εφαρμόζοντας μετρημένες πολιτικές ρυθμού και δίκαιης κατανομής πόρων μεταξύ χωρών. Επιπλέον, ενσωματώνει ελέγχους κανονικοποίησης, εξασφαλίζοντας τη συνοχή και την ποιότητα των δεδομένων. Παράλληλα, η τήρηση των κατευθυντήριων γραμμών του *robots.txt* και των όρων χρήσης περιορίζει νομικούς και ηθικούς κινδύνους, χωρίς να θέτει σε κίνδυνο την πληρότητα του συλλεγόμενου υλικού. Με αυτόν τον τρόπο, το προτεινόμενο πλαίσιο άντλησης δεδομένων λειτουργεί ως ανθεκτική υποδομή, η οποία υποστηρίζει αξιόπιστα τις επόμενες φάσεις ανάλυσης, όπως ο καθαρισμός, η μετάφραση, ο εντοπισμός γεγονότων και η μοντελοποίηση, διευκολύνοντας την έγκυρη και έγκαιρη μελέτη φαινομένων κοινωνικής κινητοποίησης.

4.3 Μετάφραση Άρθρων στα Αγγλικά

Στο επόμενο στάδιο, κρίθηκε απαραίτητη η χρήση εργαλείων MT λόγω της πολυγλωσσικής φύσης των δεδομένων. Κάθε συλλεγόμενο άρθρο παρουσιάζεται στη γλώσσα της αντίστοιχης χώρας, γεγονός που δυσχέραινε την ενοποιημένη ανάλυσή τους. Για τον σκοπό αυτό, εφαρμόστηκε η μετάφρασή τους στην Αγγλική γλώσσα, η οποία λειτούργησε ως κοινή βάση επεξεργασίας.

Πριν από τη μετάφραση πραγματοποιήθηκε στάδιο προεπεξεργασίας κειμένου, το οποίο περιλάμβανε αφαίρεση διακριτικών χαρακτήρων (Unicode normalization) και αποστρόφων, για την αποφυγή προβλημάτων συμβατότητας και καθαρισμό σημείων στίξης. Η ανάγκη για τέτοια προεπεξεργασία ενισχύθηκε από το γεγονός ότι τα άρθρα αποθηκεύονταν σε CSV αρχεία, τα οποία, αν και απλά και ευρέως υποστηριζόμενα, παρουσιάζουν ορισμένους τεχνικούς περιορισμούς, κυρίως ως προς το character encoding. Η λανθασμένη κωδικοποίηση οδηγεί συχνά σε απώλεια ή παραμόρφωση

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

χαρακτήρων, όπως οι διακριτικοί (π.χ. ñ, č, ž, ő, å) ή τα διαφορετικά είδη αποστρόφου (‘, ’, “, ”, ’, "), με αποτέλεσμα να προκαλείται αποτυχία parsing κατά την ανάγνωση των δεδομένων με εργαλεία όπως το Pandas.

Για τη μετάφραση αξιοποιήθηκαν διαφορετικά εργαλεία, ανάλογα με τη γλώσσα:

- **Googletrans** για τα Κροατικά, καθώς το LibreTranslate δεν υποστηρίζει αυτή τη γλώσσα. Πρόκειται για δωρεάν βιβλιοθήκη Python που υλοποιεί το API του Google Translate, χρησιμοποιώντας τον ίδιο διακομιστή με το translate.google.com. Αν και δεν αποτελεί επίσημο API και παρουσιάζει περιοδικές διακοπές, στην παρούσα εργασία λειτούργησε ικανοποιητικά λόγω περιορισμένων καθημερινών κλήσεων.
- **MarianMT (Helsinki-NLP/opus-mt-sv-en)** για τη μετάφραση από Σουηδικά σε Αγγλικά. Παρά την ευρεία γλωσσική κάλυψη, παρουσίασε αδυναμίες, όπως παραλείψεις προτάσεων (ιδίως σε μεγάλα κείμενα), μειωμένη νοηματική συνοχή και περιορισμένη απόδοση σε ιδιωματισμούς και τεχνικούς όρους. Επίσης, το περιορισμένο μήκος εισόδου απαιτούσε sentence/word tokenization και διαχωρισμό σε μικρότερες προτάσεις, γεγονός που αύξησε τον χρόνο εκτέλεσης.
- **LibreTranslate** για τις υπόλοιπες γλώσσες, το οποίο αποδείχθηκε πιο αξιόπιστο και συνεπές. Παρόλο που έχει μικρότερη γλωσσική κάλυψη από το MarianMT, προσέφερε καλύτερη απόδοση σε δημοσιογραφικά κείμενα, με λιγότερες απώλειες προτάσεων και υψηλότερη νοηματική συνέχεια.

Το LibreTranslate αποτελεί ένα λογισμικό ανοιχτού κώδικα, το οποίο βασίζεται στο εργαλείο Argos Translate. Το Argos Translate χρησιμοποιεί το Stanza για την ανίχνευση προτάσεων και το OpenNMT, το οποίο είναι μια πλατφόρμα MT, για τη μετάφραση κειμένων. Το Argos Translate λειτουργεί σε επίπεδο προτάσεων και επεξεργάζεται τις προτάσεις μετατρέποντάς τις σε tokens, διαδικασία που ονομάζεται tokenization. Τα tokens μπορεί να είναι λέξεις ή μέρη λέξεων. Στη συνέχεια κωδικοποιείται σε μια αριθμητική αναπαράσταση χρησιμοποιώντας ένα νευρωνικό δίκτυο κωδικοποιητή. Η κωδικοποιημένη αναπαράσταση στη συνέχεια διέρχεται μέσω ενός δικτύου αποκωδικοποιητή που δημιουργεί το μεταφρασμένο κείμενο στη γλώσσα-στόχο. Οι προτάσεις μεταφράζονται με τη χρήση του προεκπαιδευμένου μοντέλου του CTranslate2 του OpenNMT .

Αν και το MarianMT προσφέρει ευρεία γλωσσική κάλυψη, η χρήση του στην παρούσα εργασία ανέδειξε ορισμένους περιορισμούς. Παρατηρήθηκε ότι σε αρκετές περιπτώσεις, κυρίως στα μεγαλου όγκου κείμενα, παραλείπονταν ολόκληρες προτάσεις προς το τέλος των κειμένων, ενώ οι παραγόμενες μεταφράσεις παρουσίαζαν συχνά ελλιπή νοηματική συνοχή. Επιπλέον, η απόδοση ήταν μειωμένη σε ιδιωματικές εκφράσεις, τεχνικούς όρους και πολιτισμικές αναφορές, γεγονός που κατέστησε αναγκαία την κριτική ανάγνωση και, όπου απαιτούνταν, τη μετα-επεξεργασία του τελικού κειμένου. Ένας ακόμη περιορισμός αφορά το μέγιστο μήκος εισόδου, το οποίο περιόριζε τη δυνατότητα μετάφρασης μεγάλων άρθρων χωρίς τεχνητό διαχωρισμό.

Αντιθέτως, το LibreTranslate απέδωσε καλύτερα στις περισσότερες περιπτώσεις, προσφέροντας μεταφράσεις με υψηλότερη νοηματική συνέχεια και λιγότερες απώλειες προτάσεων. Αν και δεν διαθέτει την ίδια έκταση γλωσσικής κάλυψης με το

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

MarianMT, αποδείχθηκε καταλληλότερο για περιβάλλοντα όπου η ακρίβεια και η συνοχή του κειμένου αποτελούν βασικές προϋποθέσεις. Οι περιορισμοί του ως προς ιδιωματισμούς και εξειδικευμένη ορολογία μπορούν να αντιμετωπιστούν με προσαρμοσμένες τεχνικές.

Για την υπέρβαση των περιορισμών του MarianMT εφαρμόστηκαν τεχνικές sentence και word tokenization σε συνδυασμό με pattern matching, ώστε να αποφευχθεί η λανθασμένη διάσπαση ειδικών ονομάτων και συμβόλων (π.χ. U.S.A.). Όταν ο αριθμός των tokens υπερέβαινε το επιτρεπτό όριο, τα κείμενα διαχωρίζονταν σε μικρότερες ενότητες, γεγονός που διασφάλισε την πληρότητα της μετάφρασης, με κόστος ωστόσο την αύξηση του χρόνου εκτέλεσης και της υπολογιστικής πολυπλοκότητας.

Συνολικά, η διαδικασία MT στηρίχθηκε κυρίως στο LibreTranslate, με το Googletrans και το MarianMT να λειτουργούν συμπληρωματικά. Η επιλογή αυτή διασφάλισε τη δημιουργία ενός συνεκτικού αγγλόφωνου corpus, επιτρέποντας την ενιαία ανάλυση γεγονότων διαμαρτυρίας σε ευρωπαϊκό επίπεδο.

4.4 Προεπεξεργασία κειμένου και διαμόρφωση εισόδου μοντέλου

Η διαδικασία προεπεξεργασίας των δεδομένων εφαρμόστηκε με στόχο την απομάκρυνση θορύβου και την τυποποίηση του κειμένου, ώστε να είναι κατάλληλο για περαιτέρω ανάλυση. Συγκεκριμένα, αφαιρέθηκαν όλες οι μορφές υπερσυνδέσμων (URLs), HTML tags και entities, καθώς και emo is από τα αντίστοιχα Unicode ranges. Επιπλέον, αφαιρέθηκαν τα stopwords, δηλαδή λέξεις με υψηλή συχνότητα εμφάνισης και χαμηλή σημασιολογική αξία (π.χ. άρθρα, αντωνυμίες, προθέσεις), οι οποίες δεν προσθέτουν ουσιαστική πληροφορία στο κείμενο. Καθώς και αφαίρεση σημείων στίξης, με εξαίρεση τα τερματικά, και μη-ASCII χαρακτήρων, ενώ εφαρμόστηκε κανονικοποίηση Unicode για την ενοποίηση της αναπαράστασης. Στη συνέχεια, το κείμενο υποβλήθηκε σε tokenization και, σε lemmatization με χρήση για τη μείωση των λέξεων στη βασική λεξικογραφική τους μορφή.

Εργαλείο	Accuracy (%)	Perfect Matches	Partial Matches	Wrong Matches	Σύνολο
WordNet	51.3	20	0	19	39
Lancaster	20.5	8	0	31	39
Porter	74.4	17	12	10	39
Snowball	74.4	17	12	10	39
Prefix Match	**100.0**	**39**	**0**	**0**	**39**

Πίνακας 5 Σύγκριση εργαλείων για την ανίχνευση λέξεων σχετικών με γεγονότα διαμαρτυρίας.

Σε αυτό το σημείο, πραγματοποιήθηκε συγκριτική αξιολόγηση πέντε διαφορετικών τεχνικών κανονικοποίησης κειμένου για την αναγνώριση λέξεων σχετικών με γεγονότα διαμαρτυρίας, όπως strike, protest, demonstrate, rally και sit-in. Οι μέθοδοι που συγκρίθηκαν ήταν οι WordNet Lemmatizer, Lancaster Stemmer, Porter Stemmer, Snowball Stemmer και μία προσαρμοσμένη τεχνική prefix matching (Πίνακας 5).

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Η αξιολόγηση βασίστηκε στη σύγκριση των παραγόμενων λημμάτων με τα αναμενόμενα λήμματα-στόχους, μετρώντας τρεις κατηγορίες αποτελεσμάτων: perfect (τέλεια αντιστοίχιση), partial (μερική αντιστοίχιση) και wrong (λανθασμένη αντιστοίχιση). Τα αποτελέσματα έδειξαν ότι τα περισσότερα κλασικά εργαλεία lemmatization είχαν χαμηλή ή μέτρια ακρίβεια, σε λέξεις με ειδική σημασιολογική βαρύτητα, όπως demonstration, rallies, όπου συχνά παρήγαγαν μη έγκυρα αποτελέσματα.

Για την αντιμετώπιση του προβλήματος, αναπτύχθηκε μια προσαρμοσμένη μέθοδος Prefix Matching, η οποία ελέγχει αν η λέξη ξεκινά με το ζητούμενο λήμμα-στόχο και, εφόσον ισχύει, την κανονικοποιεί σε αυτό. Η μέθοδος αυτή απέδωσε 100% ακρίβεια για όλες τις στοχευμένες λέξεις, ξεπερνώντας σημαντικά τα υπόλοιπα εργαλεία. Έτσι, επιλέχθηκε ως η κύρια τεχνική κανονικοποίησης στο τελικό σύστημα, καθώς διασφαλίζει συνεπή και αξιόπιστη αναγνώριση όλων των μορφολογικών παραλλαγών, βελτιώνοντας την ομοιομορφία των δεδομένων και τη συνολική ποιότητα της ανάλυσης.

Επιπλέον, εκτός από τις λέξεις-στόχους, εφαρμόστηκε γενική διαδικασία lemmatization σε ολόκληρο το corpus, ώστε να κανονικοποιηθούν όλες οι μορφολογικές παραλλαγές των λέξεων. Η κανονικοποίηση αυτή ήταν απαραίτητη για να βελτιωθεί η ομοιομορφία του λεξιλογίου και να ενισχυθεί η απόδοση των επόμενων βημάτων, όπως η θεματική κατηγοριοποίηση, η αναζήτηση γεγονότων και η ταξινόμηση άρθρων.

Πέραν του στοχευμένου χειρισμού των όρων που σχετίζονται με απεργίες και διαμαρτυρίες, για το υπόλοιπο λεξιλόγιο (δηλαδή λέξεις που δεν περιέχονται στο corpus με τις λέξεις κλειδιά) εφαρμόστηκε γενικευμένο lemmatization μέσω του WordNet Lemmatizer. Η επιλογή του WordNet τεκμηριώνεται από την ικανότητά του να ενοποιεί συγκενικές λέξεις, διατηρώντας παράλληλα τη σημασιολογική ακεραιότητα των λημμάτων, σε αντίθεση με τους stemmers που συχνά οδηγούν σε υπερ-αποκοπές. Με τον τρόπο αυτό μειώθηκε η λεξιλογική διασπορά, βελτιώθηκε και ενισχύθηκε ομοιομορφία του, γεγονός που βελτίωσε την απόδοση των μεταγενέστερων βημάτων (θεματική κατηγοριοποίηση, εντοπισμός γεγονότων, ταξινόμηση άρθρων), χωρίς απώλεια στη διακριτότητα των εννοιών εκτός του λεξιλογίου στόχου.

4.5 Χρήση προεκπαιδευμένου μοντέλου για την δυαδική κατηγοριοποίηση γεγονότος σε διαμαρτυρίας ή όχι

Δεδομένου ότι το διαθέσιμο σύνολο των επισημασμένων δεδομένων ήταν μικρό, επιλέχθηκε η προσέγγιση Transfer Learning με RoBERTa, ώστε να αξιοποιηθεί προϋπάρχουσα γλωσσική γνώση και να αποφευχθεί η ανάγκη εκπαίδευσης από το μηδέν. Το RoBERTa προσφέρει ισχυρές γενικές αναπαραστάσεις και έχει δείξει καλά αποτελέσματα σε σενάρια περιορισμένων δειγμάτων, γεγονός που το καθιστά κατάλληλο ως βάση για προσαρμογή (fine-tuning) σε εξειδικευμένες εργασίες.

Τα κείμενα προήλθαν από ειδησεογραφικές πηγές που συλλέχθηκαν όπως αναφέρθηκε και παραπάνω με την μέθοδο του web crawling και αφορούν περιγραφές συλλογικής δράσης. Πριν την εκπαίδευση εφαρμόστηκε αφαίρεση διπλοτύπων και άρθρων που αναφέρονταν στο ίδιο γεγονός, και μεταφράστηκαν στα αγγλικά

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

(Ενότητα 4.4) και έλεγχος ισορροπίας κλάσεων, ώστε η κατανομή να είναι κατά το δυνατόν ισομερής. Επιπλέον, επειδή σε μεγάλα κείμενα παρατηρήθηκε ότι η κεντρική πληροφορία για την απεργία ή τη διαδήλωση χάνονταν μέσα στο υπόλοιπο περιεχόμενο, εφαρμόστηκε διαδικασία εστίασης σε συμφραζόμενα: για κάθε ανίχνευση λέξης-κλειδιού σχετικής με συλλογική δράση διατηρήθηκε η ίδια πρόταση, καθώς και οι δύο προηγούμενες και οι δύο επόμενες προτάσεις. Με αυτόν τον τρόπο δημιουργήθηκαν αποσπάσματα που αναδεικνύουν πιο καθαρά το γεγονός και μειώνουν τον θόρυβο από άσχετο περιεχόμενο. Η επισήμανση των δεδομένων έγινε εξ ολοκλήρου χειροκίνητα. Δηλαδή, υλοποιήθηκε δυαδική ταξινόμηση, όπου το μοντέλο εκπαιδεύτηκε για να αναγνωρίζει εάν ένα κείμενο αναφέρεται σε απεργία, διαδήλωση, πορεία ή όχι. Αρχικά, τα πρώτα άρθρα συλλέχθηκαν με βάση λέξεις-κλειδιά σχετικές με κινητοποιήσεις και κοινωνική δράση, το οποίο διασφάλισε την κάλυψη ευρέος φάσματος σχετικών γεγονότων.

Με βάση τα παραπάνω χρησιμοποιήθηκε το προεκπαιδευμένο γλωσσικό μοντέλο RoBERTa. Το σύνολο δεδομένων διαιρέθηκε σε υποσύνολα εκπαίδευσης και επικύρωσης σε αναλογία 80/20. Η είσοδος κωδικοποιήθηκε με μέγιστο μήκος 512 tokens (padding και truncation), επιλογή που ευθυγραμμίζεται με την κατανομή μήκους των κειμένων (mean ≈ 394 tokens) και περιορίζει την αποκοπή μόνο των μακροσκελών εγγράφων. Το fine-tuning γίνεται με optimizer AdamW, μικρό ρυθμό μάθησης, early stopping και επιλογή μοντέλου βάσει macro-F1 στο σύνολο επικύρωσης, ώστε να μην ευνοούνται οι συχνότερες κλάσεις. Το καλύτερο μοντέλο φορτώνεται στο τέλος βάσει του macro-F1 στο validation set.

Η αξιολόγηση υπολογίζει accuracy, precision/recall/F1 (binary, macro, micro), καθώς και ROC-AUC και PR-AUC από τις πιθανότητες της θετικής κλάσης (softmax). Επιπλέον, παράγεται confusion matrix και πλήρες classification report.

Σύμφωνα με τον Vu ović (Vujović, 2021) αναλύονται οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν.

Πίνακας Σύγχυσης (confusion matrix) για δυαδικό ταξινομητή (Πίνακας 3). Οι πραγματικές τιμές σημειώνονται ως Αληθές (1) και Ψευδές (0), ενώ οι προβλέψεις ως Θετικό (1) και Αρνητικό (0). Οι μετρικές της απόδοσης των μοντέλων ταξινόμησης προκύπτουν από τις ποσότητες TP, TN, FP, FN που περιλαμβάνονται στον πίνακα σύγχυσης.

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Πίνακας 6 Πίνακας σύγχυσης για δυαδική κατηγοριοποίηση[Saito, 2017]

TP (True Positive) — Αληθώς Θετικό: Ένα σημείο δεδομένων (δηλαδή μεμονωμένες καταγραφές ή τιμές σε ένα σύνολο δεδομένων) στο confusion matrix είναι Αληθώς Θετικό (TP) όταν προβλέπεται θετικό αποτέλεσμα και το πραγματικό αποτέλεσμα είναι επίσης θετικό.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

FP (False Positive) — Ψευδώς Θετικό: Ένα σημείο δεδομένων στο confusion matrix είναι Ψευδώς Θετικό (FP) όταν προβλέπεται θετικό αποτέλεσμα, αλλά το πραγματικό αποτέλεσμα είναι αρνητικό. Το σενάριο αυτό αντιστοιχεί σε Σφάλμα Τύπου I (Type I Error).

FN (False Negative) — Ψευδώς Αρνητικό: Ένα σημείο δεδομένων στο confusion matrix είναι Ψευδώς Αρνητικό (FN) όταν προβλέπεται αρνητικό αποτέλεσμα, και το πραγματικό αποτέλεσμα είναι θετικό. Το σενάριο αυτό είναι γνωστό ως Σφάλμα Τύπου II (Type II Error) και θεωρείται εξίσου σημαντικό με το Σφάλμα Τύπου I.

TN (True Negative) — Αληθώς Αρνητικό: Ένα σημείο δεδομένων στο confusion matrix είναι Αληθώς Αρνητικό (TN) όταν προβλέπεται αρνητικό αποτέλεσμα και το αποτέλεσμα είναι επίσης αρνητικό.

Accuracy

Το accuracy υπολογίζεται ως το άθροισμα των δύο σωστών προβλέψεων (TP TN) διαιρεμένο με το πλήθος όλων των δειγμάτων (P N).

Η καλύτερη δυνατή τιμή είναι 1.0 και η χειρότερη 0.0.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N + P}$$

TP Rate / Sensitivity / Recall (Αληθώς Θετικό Ποσοστό / Ευαισθησία / Ανάκληση)

TP Rate / Sensitivity / Recall (Αληθώς Θετικό Ποσοστό / Ευαισθησία / Ανάκληση)

Ο True Positive Rate (TPR) ή Sensitivity/Recall είναι ο λόγος των σωστών θετικών προβλέψεων (TP) προς το πλήθος όλων των πραγματικών θετικών (P).

Η μέγιστη τιμή είναι 1.0 και η ελάχιστη 0.0.

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Precision

Η precision είναι ο λόγος των σωστών θετικών (TP) προς όλα όσα προβλέφθηκαν ως θετικά (TP+FP).

Η καλύτερη τιμή είναι 1.0 και η χειρότερη 0.0.

$$Precision = \frac{TP}{TP + FP}$$

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

F-Measure / F-score

Το F-score μετρά την ακρίβεια της δοκιμής συνδυάζοντας precision και recall με τον αρμονικό μέσο:

$$F = \frac{2 * precision * recall}{precision + recall}$$

ROC Area (ROC AUC)

Η καμπύλη ROC απεικονίζει την ανταλλαγή (trade-off) ανάμεσα σε True Positive Rate (TPR) και False Positive Rate (FPR) για όλα τα πιθανά κατώφλια.

Για κάθε κατώφλι υπολογίζονται TPR και FPR και σχεδιάζονται στο γράφημα. Όσο υψηλότερο το TPR και χαμηλότερο το FPR σε κάθε κατώφλι, τόσο καλύτερα.

Το ROC AUC είναι το εμβαδόν κάτω από την καμπύλη ROC και ποσοτικοποιεί το πόσο καλά ταξινομεί το μοντέλο τις θετικές περιπτώσεις πάνω από τις αρνητικές.

PRC Area (PR AUC — Precision-Recall Curve Area)

Το PR AUC συνοψίζει τις επιδόσεις του μοντέλου στην καμπύλη Precision–Recall. Συχνά υπολογίζεται ως Average Precision, δηλαδή μέσος των precision σε διάφορα thresholds ανά τιμή recall στο [0,1].

Η καμπύλη PR προκύπτει συνδυάζοντας την Positive Predictive Value (Precision) και το True Positive Rate (Recall): για κάθε κατώφλι υπολογίζονται τα αντίστοιχα ζεύγη και σχεδιάζονται.

Ένα καλό PRC έχει μεγαλύτερο AUC. Η βιβλιογραφία δείχνει ότι το PRC είναι συχνά πιο πληροφοριακό από το ROC όταν εκτιμώνται δυαδικοί ταξινομητές σε ανισόρροπα σύνολα δεδομένων.

Metric	Τιμή
Macro-F1	0.899
F1 (Θετική κλάση)	0.917
Precision (Θετική)	0.917
Recall (Θετική)	0.917
PR-AUC	0.967
ROC-AUC	0.950
Accuracy και Micro-F1	0.902

Πίνακας 7 Κύρια μετρική η macro-F1 για ισόρροπη αξιολόγηση ανά κλάση.

Τα αποτελέσματα, στον Πίνακας 7, δείχνουν σταθερή και υψηλή επίδοση. Η F1 της θετικής κλάσης (διαμαρτυρία) είναι 0.917, υποδηλώνοντας πολύ καλή ταυτόχρονη

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

ακρίβεια και ανάκληση στην ανίχνευση διαμαρτυριών. Η macro-F1 = 0.899 είναι ελαφρώς χαμηλότερη από τη binary F1, που σημαίνει ότι η επίδοση στη μη θετική κλάση είναι λίγο ασθενέστερη. Η PR-AUC = 0.967 και η ROC-AUC = 0.950 επιβεβαιώνουν ισχυρή διαχωριστική ικανότητα σε εύρος κατωφλίων, με ιδιαίτερα καλή συμπεριφορά όταν προέχει η ακρίβεια στην θετική κλάση. Το accuracy είναι σχεδόν ίσο με το micro-F1 0.902, το οποίο φανερώνει συνολική συνέπεια των προβλέψεων.

4.6 Εξαγωγή τοποθεσιών και χωρών

Αφού διαφοροποιήθηκαν τα άρθρα που αναφέρονται αποκλειστικά σε γεγονότα διαμαρτυρίας, επόμενο βήμα αποτελεί ο διαχωρισμός των άρθρων που αφορούν μόνο τις χώρες της Ε.Ε. αποκλείοντας γεωγραφικά μη συναφή δεδομένα.

Για την επίτευξη αυτού, αξιοποιήθηκε η βιβλιοθήκη spaCy, η οποία, μετά τη διαδικασία μετάφρασης και της κατηγοριοποίησης του μοντέλου, εφαρμοζόταν σε άρθρα που αναφέρονταν σε γεγονότα διαδηλώσεων με σκοπό την αναγνώριση τοποθεσιών και χωρών. Ωστόσο, παρατηρήθηκε ότι σε αρκετές περιπτώσεις η spaCy αναγνώριζε εσφαλμένα και άλλες κατηγορίες λέξεων ως τοπωνύμια. Για να αντιμετωπιστεί αυτός ο περιορισμός, ενσωματώθηκε το GeoNames API, το οποίο παρείχε πιο αξιόπιστη ταυτοποίηση γεωγραφικών οντοτήτων. Παρόλο που τα αποτελέσματά του ήταν αξιόλογα, διαθέτει περιορισμένο αριθμό αιτημάτων. Για αυτόν τον λόγο, κάθε έγκυρη αντιστοίχηση ζεύγους <τοποθεσίας – χώρας> αποθηκευόταν στη βάση δεδομένων, ώστε να αποφεύγεται η υπερβολική χρήση του API.

Με τον τρόπο αυτό, δημιουργήθηκε μια καθαρή και ενοποιημένη βάση άρθρων, η οποία περιλαμβάνει αποκλειστικά διαμαρτυρίες που σχετίζονται με χώρες της Ευρωπαϊκής Ένωσης, εξασφαλίζοντας την ακρίβεια και τη συνέπεια της περαιτέρω ανάλυσης.

4.7 Χρήση pattern extraction για την εξαγωγή αριθμού συμμετεχόντων και του τύπου διαμαρτυρίας (βίαιη ή ειρηνική)

Επιπλέον χαρακτηριστικό την παρούσας πτυχιακής εργασίας είναι η εκτίμηση του αριθμού των ατόμων που συγκεντρώθηκαν στην διαμαρτυρία. Αυτό επιτεύχθηκε με την υλοποίηση pattern extraction.

Το pattern extraction ορίζεται ως μια η διαδικασία αναγνώρισης και εξαγωγής πληροφορίας από κείμενο με βάση προκαθορισμένα πρότυπα (patterns), τα οποία συνήθως υλοποιούνται ως κανόνες κανονικών εκφράσεων (regular expressions), λεξιλόγια ή γλωσσικές δομές. Στόχος είναι ο εντοπισμός συγκεκριμένων μορφών, όπως αριθμοί, ημερομηνίες, ονόματα ή χαρακτηρισμοί (π.χ. «ειρηνική διαμαρτυρία»), χωρίς να απαιτείται πλήρης κατανόηση του συμφραζομένου από το μοντέλο. Με άλλα λόγια, το pattern extraction αξιοποιεί κανόνες επιφανειακής αντιστοίχισης για να ανιχνεύσει και να δομήσει τμήματα πληροφορίας μέσα σε μη δομημένα κείμενα.

Για την εξαγωγή του αριθμού των ατόμων χρησιμοποιήθηκαν δύο τιμές, η μία αντιπροσωπεύει τον ελάχιστο αριθμό ατόμων που συγκεντρώθηκαν και η άλλη χρησιμοποιείται, όταν εντοπίζονται περισσότερα από ένα σύνολα συμμετεχόντων,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

υπολογίζοντας το άθροισμά τους, το οποίο αποτελεί μια επιπλέον εκτίμηση. Στην περίπτωση, που γίνονται δύο διαφορετικές εκτιμήσεις σχετικά με τον αριθμό των ατόμων που συγκεντρώθηκαν, τότε σε αυτήν την περίπτωση αποθηκεύεται ο μικρότερος εκτιμώμενος αριθμός.

Για την εξαγωγή αν η διαμαρτυρία ήταν βίαια ή ειρηνικής, εφαρμόζεται επίσης pattern extraction. Συγκεκριμένα, αναζητούνται λέξεις-κλειδιά και φράσεις που δηλώνουν επεισόδια, συλλήψεις ή τραυματισμούς. Εάν εντοπιστεί κάποιο από αυτά τα στοιχεία, το γεγονός καταγράφεται ως βίαιη διαμαρτυρία· διαφορετικά, χαρακτηρίζεται ως ειρηνική.

Στο Παράρτημα 1, καταγράφονται οι λέξεις-κλειδιά που χρησιμοποιήθηκαν.

4.8 Σύγκριση Ομοιότητας και Ομαδοποίηση Άρθρων

Στο πλαίσιο της παρούσας εργασίας αναπτύχθηκε ένα σύστημα αυτόματης δημιουργίας story threads, βασισμένο στην προσέγγιση των Nallapati et al. (Nallapati, 2004), με στόχο την οργάνωση, σύνδεση και ανάλυση ειδησεογραφικών άρθρων που σχετίζονται με κοινωνικές δράσεις, όπως απεργίες, διαδηλώσεις και άλλες μορφές συλλογικής κινητοποίησης.

Η μεθοδολογία που ακολουθήθηκε βασίζεται στη δημιουργία ιεραρχικών σχέσεων parent–child μεταξύ άρθρων, αξιοποιώντας συνδυαστικά τρεις θεμελιώδεις διαστάσεις:

1. τη θεματική συνάφεια μεταξύ των άρθρων
2. τη χρονική ακολουθία δημοσίευσης, ώστε να αποτυπώνεται η εξέλιξη των γεγονότων, και
3. το γεωγραφικό πλαίσιο στο οποίο αναφέρονται τα άρθρα

Πέρα από το βασικό μοντέλο story threading, το προτεινόμενο σύστημα επεκτάθηκε ώστε να ενσωματώνει και διαδικασίες ανίχνευσης διπλότυπων άρθρων (duplicates), καθώς και κατηγοριοποίησης follow-ups, επιτρέποντας όχι μόνο τη σύνδεση και παρακολούθηση της εξέλιξης ενός γεγονότος, αλλά και την αποφυγή επαναλήψεων και την αποτύπωση της συνέχειας στην κάλυψη. Η ενσωμάτωση αυτών των επιπέδων προσφέρει μια πιο ολοκληρωμένη και αξιόπιστη ανάλυση του πληροφοριακού όγκου.

Επιπλέον, για τη μείωση της υπολογιστικής πολυπλοκότητας υιοθετήθηκε ο εξής περιορισμός: κάθε child άρθρο επιτρέπεται να συνδέεται μόνο με ένα parent άρθρο**.** Παράλληλα, ένα child μπορεί να διαθέτει δικά του children, λειτουργώντας έτσι ταυτόχρονα ως parent για επόμενα άρθρα. Με αυτόν τον τρόπο, η δομή παραμένει απλή και ιεραρχική, ενώ επιτρέπεται η αναπαράσταση της χρονικής εξέλιξης των γεγονότων μέσω αλυσίδων συνδέσεων.

4.8.1 Υπολογισμός θεματικής συνάφειας

Για τον υπολογισμό της θεματικής ομοιότητας μεταξύ των άρθρων εφαρμόστηκε η μέθοδος αναπαράστασης κειμένων TF-IDF, η οποία αποτυπώνει τη σχετική βαρύτητα κάθε λέξης στο κείμενο, λαμβάνοντας υπόψη τόσο τη συχνότητα εμφάνισής της στο ίδιο το άρθρο όσο και τη σπανιότητά της σε ολόκληρο το σύνολο δεδομένων. Η

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

ομοιότητα μετρήθηκε μέσω του μέτρου ομοιότητας συνημιτόνου, ο οποίος αξιολογεί τον συσχετισμό δύο διανυσμάτων αναπαράστασης κειμένου και κυμαίνεται από 0 (καμία ομοιότητα) έως 1 (ταυτόσημα κείμενα). Ένα άρθρο Α και ένα άρθρο Β θεωρούνται υποψήφια προς σύνδεση όταν η τιμή της ομοιότητας ξεπερνά ένα προκαθορισμένο κατώφλι.

4.8.2 Χρονική ακολουθία και ορισμός ιεραρχικών σχέσεων

Για τον καθορισμό πιθανών σχέσεων μεταξύ άρθρων, η σύγκριση πραγματοποιείται με βάση την ημερομηνία δημοσίευσης. Συγκεκριμένα, χρησιμοποιείται ένα εύρος επτά ημερών, ώστε να περιορίζεται το πεδίο αναζήτησης και να μειώνεται η πολυπλοκότητα του αλγορίθμου.

- Η αναζήτηση συνδέσεων με νέα άρθρα περιορίζεται σε δημοσιεύσεις που απέχουν το πολύ επτά ημέρες, εξασφαλίζοντας σημαντική μείωση χρόνου εκτέλεσης.
- Σε κάθε συσχετισμένο ζεύγος, το παλαιότερο άρθρο χαρακτηρίζεται ως parent, καθώς θεωρείται το αρχικό σημείο αναφοράς.
- Το νεότερο άρθρο χαρακτηρίζεται ως child, καθώς είναι πιο πιθανό να ενημερώνει ή να συμπληρώνει το περιεχόμενο του προηγούμενου.

Η διαδικασία αυτή επιτρέπει την κατασκευή αλυσίδων άρθρων που αποτυπώνουν τη χρονική εξέλιξη ενός γεγονότος, διασφαλίζοντας ταυτόχρονα ότι η ανάλυση βασίζεται σε πραγματική χρονική αλληλουχία.

4.8.3 Ενσωμάτωση γεωγραφικού πλαισίου

Για την αποφυγή λανθασμένων συνδέσεων μεταξύ άρθρων που αφορούν διαφορετικά και ασύνδετα μεταξύ τους θεματικά γεγονότα αλλά παρουσιάζουν παρόμοιο λεξιλόγιο, εφαρμόστηκε ένας γεωγραφικός περιορισμός. Η σύνδεση μεταξύ δύο άρθρων επιτρέπεται μόνο όταν αναφέρονται στην ίδια χώρα. Για παράδειγμα, όταν μια διαμαρτυρία κριθεί πολύ σημαντική, είναι πιθανό να αποτελέσει αντικείμενο αναφοράς όχι μόνο στον εγχώριο Τύπο της χώρας όπου εκτελείται το συμβάν, αλλά και σε διεθνή μέσα ενημέρωσης άλλων χωρών. Στις περιπτώσεις αυτές, η πολλαπλή κάλυψη αναδεικνύει τη βαρύτητα του γεγονότος, χωρίς όμως να οδηγεί σε συγχώνευση διαφορετικών γεωγραφικών clusters, αλλά σε ξεχωριστές αναφορές που συνδέονται θεματικά.

Ο γεωγραφικός εντοπισμός πραγματοποιήθηκε μέσω εξαγωγής τοπωνυμίων από το κείμενο με τεχνικές NLP και στη συνέχεια μέσω ταυτοποίησης της αντίστοιχης χώρας με χρήση του GeoNames API. Η προσέγγιση αυτή διασφαλίζει ότι τα story threads αποτυπώνουν γεγονότα που έλαβαν χώρα στο ίδιο γεωγραφικό πλαίσιο, ενισχύοντας έτσι την ακρίβεια και τη θεματική συνοχή των παραγόμενων αλυσίδων.

4.8.4 Κατηγοριοποίηση των σχέσεων άρθρων και αποθήκευση στη Βάση Δεδομένων

Με βάση το ποσοστό ομοιότητας μεταξύ άρθρων, καθώς και την χώρα και την ημερομηνία κατά την οποία εκτυλίσσεται το γεγονός που περιγράφουν, τα άρθρα κατατάσσονται σε τρίεις κατηγορίες:

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

- 0.4 ομοιότητα < 0.7: Τα άρθρα θεωρούνται μέρος ενός Story Thread, δηλαδή συνδέονται ως επιπρόσθετες πληροφορίες ή εξελίξεις για το ίδιο γεγονός/συμβάν.
- 0.7 ομοιότητα < 0.9: Τα άρθρα αναφέρονται στο ίδιο συμβάν με μερικές αλλάγες ή προσθήκες.
- ομοιότητα ≥ 0.9: Τα άρθρα χαρακτηρίζονται ως duplicates, καθώς παρουσιάζουν ουσιαστικά το ίδιο περιεχόμενο.

Η κατηγοριοποίηση αυτή διευκολύνει τη διάκριση ανάμεσα σε νέα πληροφορία, επικαιροποίηση και αναπαραγωγή υφιστάμενου περιεχομένου, παρέχοντας πιο αξιόπιστη ανάλυση της ροής των ειδήσεων.

Οι σχέσεις που προκύπτουν αποθηκεύονται σε διαφορετικές συλλογές της MongoDB:

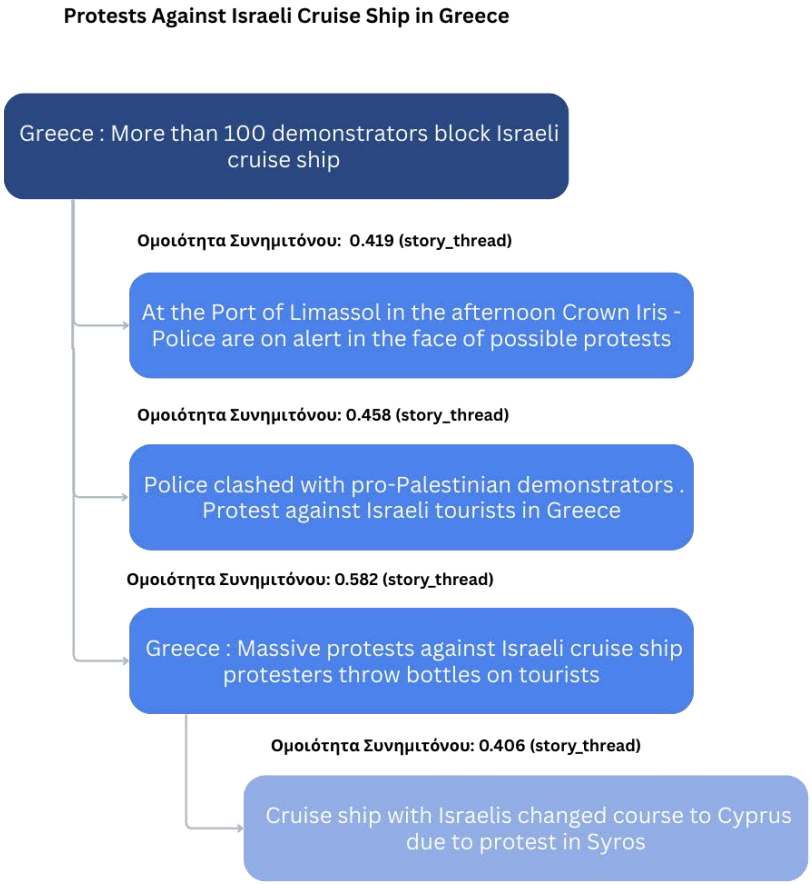
- Η πρώτη χρησιμοποιείται για την ανάλυση των σχέσεων και περιλαμβάνει τα παρακάτω στοιχεία:

  - id της εγγρφής,
  - parent_id,
  - child_id,
  - similarity_score,
  - relationship_type,
  - country

- Στη βασική συλλογή αποθηκεύεται η πληροφορία σχετικά με το αν το άρθρο είναι parent ή child, ώστε να μπορεί να γίνει η σύνδεση με την πρώτη συλλογή.

Η σύνδεση αυτή πραγματοποιείται μέσω του id του άρθρου στη βασική συλλογή, το οποίο αντιστοιχίζεται με το parent id ή child id στην αναλυτική συλλογή. Με αυτόν τον τρόπο η βασική συλλογή παραμένει εστιασμένη στα ίδια τα άρθρα, ενώ η αναλυτική συλλογή αποτυπώνει με λεπτομέρεια τις μεταξύ τους σχέσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 12 Ενδεικτικό Παράδειγμα Ιεραρχικού Δέντρου Parent-Child Σχέσεων.

Στην Εικόνα 12 παρουσιάζεται ένα story thread που αποτυπώνει την εξέλιξη γεγονότων γύρω από την άφιξη ισραηλινού κρουαζιερόπλοιου στην Ελλάδα, γεγονός που προκάλεσε αντιδράσεις και διαμαρτυρίες πολιτών. Το νήμα ξεκινά με το αρχικό συμβάν, όπου περισσότεροι από εκατό διαδηλωτές απέκλεισαν την πρόσβαση του πλοίου, οδηγώντας σε κινητοποίηση. Στη συνέχεια, η ιστορία εξελίσσεται με επεισόδια συγκρούσεων μεταξύ διαδηλωτών και αστυνομίας, καθώς και επιθέσεις εναντίον Ισραηλινών τουριστών. Τελικά, λόγω της έντασης και για λόγους ασφαλείας, το κρουαζιερόπλοιο αναγκάστηκε να αλλάξει πορεία και να κατευθυνθεί προς την Κύπρο. Παράλληλα, οι εξελίξεις επεκτάθηκαν και στην Κύπρο, όπου οι αρχές τέθηκαν σε επιφυλακή για πιθανές διαμαρτυρίες.

4.9 Δημιουργία περίληψης άρθρων διαμαρτυριών

Στη συνέχεια, εφαρμόστηκε διαδικασία αυτόματης παραγωγής περιλήψεων με στόχο την παροχή μιας σύντομης και άμεσης επισκόπησης του περιεχομένου. Η ανάγκη αυτή προκύπτει επειδή πρόκειται να αναπτυχθεί στο μέλλον μια εφαρμογή που θα απευθύνεται τόσο σε επιστήμονες όσο και σε πολίτες· συνεπώς, θεωρείται σημαντικό να υπάρχουν περιλήψεις που να διευκολύνουν την κατανόηση ακόμη και όταν οι αρχικές πηγές δεν παρέχουν έτοιμη σύνοψη. Για τον σκοπό αυτό αξιοποιήθηκε το μοντέλο facebook/bart-large-cnn, ένα προεκπαιδευμένο μοντέλο τύπου Transformer (encoder–decoder) της οικογένειας BART, το οποίο έχει εκπαιδευτεί και

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

προσαρμοστεί ειδικά στο CNN/DailyMail dataset για την εκτέλεση abstractive summarization.

Επιλέχθηκε σε σχέση με άλλα διαθέσιμα μοντέλα (π.χ. extractive μοντέλα ή πιο ελαφριές αρχιτεκτονικές) διότι το BART έχει αποδειχθεί ότι συνδυάζει την κατανόηση περιεχομένου μέσω bidirectional encoding με την παραγωγή φυσικού κειμένου μέσω autoregressive decoding, προσφέροντας έτσι υψηλή απόδοση και συνεκτικότητα στις περιλήψεις. Σε αντίθεση με extractive προσεγγίσεις που απλώς επιλέγουν προτάσεις από το κείμενο, το BART μπορεί να δημιουργήσει νέες φράσεις, εξασφαλίζοντας περιλήψεις πιο σύντομες, ρέουσες και κατανοητές, ιδανικές για ειδησεογραφικά κείμενα.

Εικόνα 13 Στιγμιότυπο διαδικασίας παραγωγής της μετάφρασης και περίληψης της διαμαρτυρίας.

Στην Εικόνα 13 παρουσιάζεται η διαδικασία επεξεργασίας κειμένων που υλοποιήθηκε στην παρούσα εργασία. Αρχικά, τα άρθρα μεταφράζονται στα αγγλικά, στη συνέχεια δημιουργείται αυτόματη σύνοψη και τελικά παράγεται η συνοπτική περίληψη μέσω προεκπαιδευμένου μοντέλου. Με τον τρόπο αυτό δημιουργείται ένα ολοκληρωμένο pipeline που συνδυάζει μετάφραση, ενδιάμεση σύνοψη και τελική περίληψη, ώστε να αποδίδεται με σαφήνεια και συνοχή το περιεχόμενο κάθε άρθρου.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

4.10 Χρήση προεκπαιδευμένου μοντέλου για τη θεματική κατηγοριοποίηση των διαμαρτυριών

Σε επόμενο στάδιο, πάνω από στο ίδιο υπόβαθρο RoBERTa, εκπαιδεύτηκε ταξινομητής multi-label για να κατηγοριοοποιεί τα δεδομένα σε μία ή παραπάνω θεματικές ετικέτες, οι οποίες αναφέρονται στον Πίνακας 8.

Στο μοντέλο εφαρμόστηκε με sigmoid ενεργοποίηση ανά ετικέτα και binary cross-entropy ως συνάρτηση κόστους. Αρχικά, το κατώφλι απόφασης ορίστηκε στο 0.5, αλλά στη συνέχεια έγινε βελτιστοποίηση κατωφλιών ανά ετικέτα στο σύνολο επικύρωσης, προκειμένου να μεγιστοποιηθεί το F1 score και να επιτευχθεί καλύτερη ισορροπία μεταξύ precision και recall. Για τη βελτιστοποίηση αυτή, χρησιμοποιήθηκαν διάφορα κατώφλια για κάθε κατηγορία, προσδιορίζοντας το καλύτερο threshold που βελτιστοποιεί τις επιδόσεις σε κάθε ετικέτα ξεχωριστά.

Political	310	Civil Liberties	50
Labor & Economy	273	Education	46
Foreign Policy & International Relations	208	Security & Law Enforcement	44
Human Rights & Equality	196	Agriculture & Farmers	29
Environment & Climate	124	Safety & Protection	18
Transportation	61	Religion & Ethics	11
Health & Social Care	55		

Πίνακας 8 Πλήθος ετικετών ανά θεματική κατηγορία.

F1 Macro	0.7423	Precision Micro	0.8447
Precision Macro	0.6901	Recall Micro	0.8201
Recall Macro	0.8214	Accuracy	0.8175
F1 Micro	0.8447		

Πίνακας 9 Μετρικές του μοντέλου.

Η αξιολόγηση έγινε με micro-/macro-F1 και weighted-F1, καθώς και με δείκτες ανά ετικέτα για τον εντοπισμό δύσκολων προβλεπόμενων θεματικών. Η χρήση RoBERTa με fine-tuning, σε συνδυασμό με χειροκίνητη κατηγοριοποίηση, επέτρεψε αποτελεσματική μάθηση παρά τον μικρό όγκο δεδομένων και παρήγαγε πλούσια πολυετικετική κατηγοριοποίηση.

Το μοντέλο παρουσιάζει ικανοποιητικά αποτελέσματα με υψηλά επίπεδα F1 Micro και Precision Micro (Πίνακας 9), δείχνοντας ότι οι προβλέψεις του για το σύνολο των

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

δεδομένων είναι γενικά σωστές και καλά ισορροπημένες. Ωστόσο, παρατηρείται μια μικρή ανισότητα στις επιδόσεις μεταξύ Macro και Micro metrics, με το F1 Macro να είναι χαμηλότερο, κάτι που υποδεικνύει ότι το μοντέλο έχει δυσκολίες στην απόδοση σε λιγότερο συχνές κατηγορίες. Η Ακρίβεια (Accuracy) είναι 81.75%, που είναι καλό ποσοστό για multi-label classification, αν και υπάρχει περιθώριο βελτίωσης. Ο Recall Macro είναι υψηλός (0.8214), υποδεικνύοντας ότι το μοντέλο αναγνωρίζει καλά τα θετικά δείγματα, αλλά η Precision Macro (0.6901) μπορεί να βελτιωθεί, καθώς υπάρχουν περισσότερες ψευδώς θετικές προβλέψεις. Η επιλογή του κατωφλίου 0.50 για τις προβλέψεις φαίνεται να αποδίδει καλά, αν και μπορεί να χρειαστεί περαιτέρω βελτιστοποίηση των thresholds ανά κατηγορία για να επιτευχθεί καλύτερη ισορροπία μεταξύ precision και recall.

4.11 Επιλογή Βάσης Δεδομένων

Στην παρούσα εργασία, η επιλογή της MongoDB για τη συλλογή και ανάλυση των άρθρων που σχετίζονται με απεργίες και διαδηλώσεις δεν είναι τυχαία. Το υπό μελέτη dataset προέρχεται από διαδικτυακές εφημερίδες και ειδησεογραφικές πηγές από κάθε χώρα της Ευρωπαϊκής Ένωσης, με τη μορφή JSON. Η MongoDB υποστηρίζει εγγενώς το συγκεκριμένο format, καθιστώντας την ιδιαίτερα ευέλικτη στην αποθήκευση ημιδομημένων εγγραφών που περιέχουν μεταβαλλόμενα πεδία, όπως τίτλο, ημερομηνία, χώρα, θεματική κατηγορία, σύνδεσμο, ακόμα και το πλήρες κείμενο του άρθρου. Σε αντίθεση με τις σχεσιακές βάσεις δεδομένων, δεν απαιτεί αυστηρό ορισμό schema, επιτρέποντας την εύκολη προσαρμογή της βάσης καθώς τα δεδομένα εξελίσσονται και νέες κατηγορίες πληροφορίας καθίστανται διαθέσιμες.

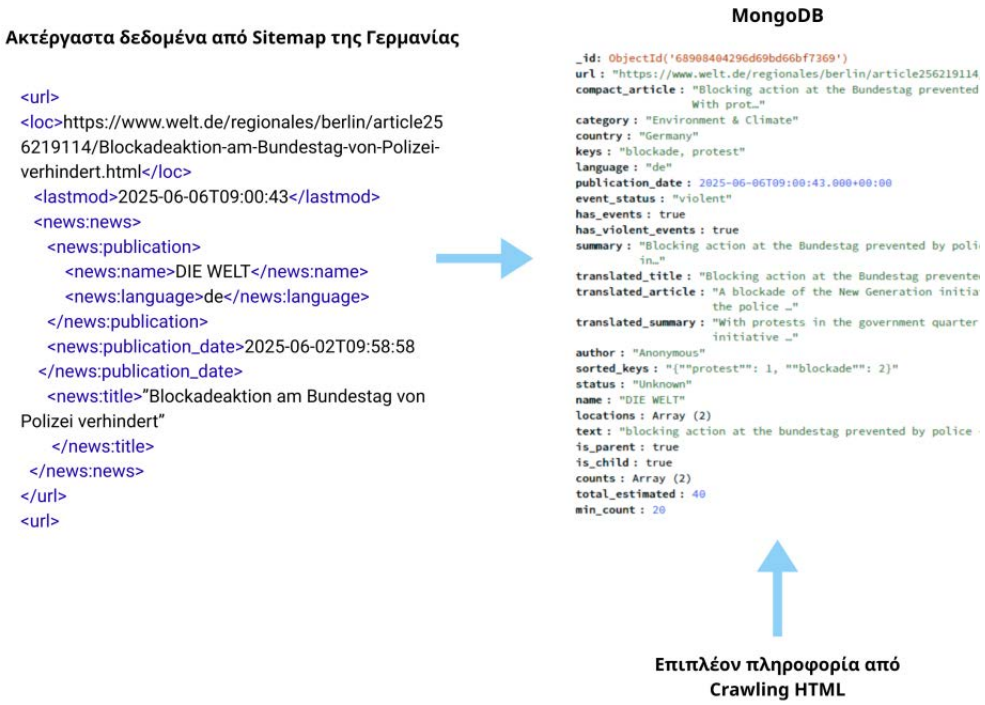
Ένα ακόμη χαρακτηριστικό που ενισχύει την αποτελεσματικότητα της MongoDB είναι η υποστήριξη ευρετηρίων (indexes). Μέσω των ευρετηρίων, η ανάκτηση δεδομένων από μεγάλες συλλογές επιταχύνεται σημαντικά, καθώς το σύστημα δεν χρειάζεται να διατρέχει σειριακά όλα τα έγγραφα για να επιστρέψει τα αποτελέσματα ενός ερωτήματος. Η ύπαρξη κατάλληλων indexes είναι ιδιαίτερα χρήσιμη στο παρόν έργο, όπου απαιτείται ταχύτατη ανάκτηση άρθρων με βάση κριτήρια όπως η ημερομηνία, η χώρα ή το είδος της κινητοποίησης.

Επιπλέον, τα τελευταία χρόνια έχει αναπτυχθεί και η υπηρεσία MongoDB Atlas, η οποία παρέχεται ως πλήρως διαχειριζόμενη πλατφόρμα cloud. Μέσω του Atlas, οι χρήστες έχουν τη δυνατότητα να δημιουργούν, να διαχειρίζονται και να επεκτείνουν τις βάσεις δεδομένων τους χωρίς την ανάγκη τοπικής εγκατάστασης ή πολύπλοκης συντήρησης υποδομών. Η λύση αυτή προσφέρει υψηλή διαθεσιμότητα, δυνατότητες αυτοματοποιημένου scaling και ενσωματωμένους μηχανισμούς ασφαλείας, διευκολύνοντας τη χρήση της MongoDB σε μεγάλα έργα που απαιτούν αποθήκευση και ανάλυση δεδομένων σε πραγματικό χρόνο.

Η MongoDB, επομένως, ανταποκρίνεται βέλτιστα στις ανάγκες του έργου. Συνδυάζει την ταχύτητα και την επεκτασιμότητα με την ευελιξία στη μορφοποίηση των δεδομένων, παρέχοντας ένα περιβάλλον που μπορεί να υποστηρίξει την ανάλυση μεγάλου όγκου ετερογενών δεδομένων, όπως αυτά που προέρχονται από ειδησεογραφικά άρθρα και αναφορές διαμαρτυριών. Η δυνατότητα οριζόντιας κλιμάκωσης, η χρήση ειδικευμένων ευρετηρίων για ταχύτερη αναζήτηση, καθώς και

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

η ύπαρξη της cloud πλατφόρμας MongoDB Atlas, ενισχύουν περαιτέρω την καταλληλότητά της για το αντικείμενο της μελέτης.



Εικόνα 14 Στιγμιότυπο αποθήκευσης στην βάση δεδομένων.

Στην Εικόνα 14, στην αριστερή πλευρά παρουσιάζεται το αρχικό sitemap, το οποίο αποτελεί την πηγή δεδομένων που υποβάλλεται σε διαδικασία crawling για τη συλλογή των άρθρων που θα εισαχθούν στη βάση δεδομένων. Στη δεξιά πλευρά, απεικονίζεται ένα στιγμιότυπο εγγραφής από τη βάση δεδομένων, μετά την ολοκλήρωση όλων των σταδίων επεξεργασίας, αποτυπώνοντας τη δομή και το περιεχόμενο όπως αποθηκεύονται τελικά.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

5. Στατιστική και Εξελικτική Ανάλυση των Δεδομένων Ειδησεογραφίας (Ιούλιο και Αύγουστο 2025)

Στο παρόν κεφάλαιο πραγματοποιείται η ανάλυση των δεδομένων που συλλέχθηκαν μέσω του συστήματος που παρουσιάστηκε αναλυτικά παραπάνω. Η περίοδος αναφοράς εκτείνεται από την 1η Ιουλίου 2025 έως την 31η Αυγούστου 2025 και καλύπτει ειδήσεις που σχετίζονται με κοινωνικές δράσεις, όπως απεργίες και διαδηλώσεις.

Η ανάλυση επικεντρώνεται τόσο στη στατιστική περιγραφή των δεδομένων όσο και στη μελέτη της χρονικής εξέλιξης των γεγονότων. Η μεθοδολογία που ακολουθείται επιτρέπει την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με:

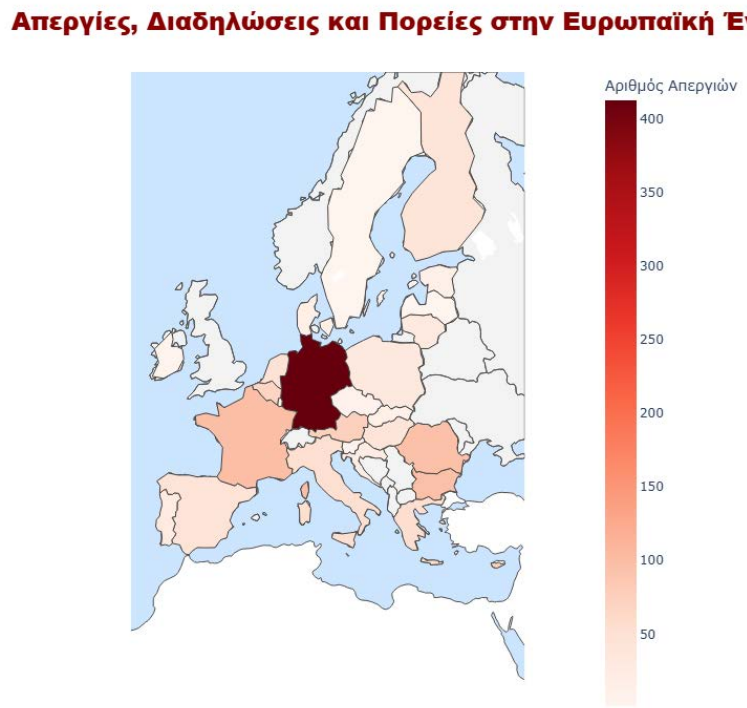
- την ένταση και συχνότητα των κοινωνικών δράσεων,
- τη γεωγραφική κατανομή τους,
- τη διάρκεια και τη δυναμική εξέλιξη των γεγονότων,
- τα μοτίβα follow-ups και αναδημοσιεύσεων,
- καθώς και την εξέλιξη της θεματολογίας μέσα στο χρονικό διάστημα.

Με τον τρόπο αυτό, αναδεικνύεται η χρησιμότητα του συστήματος για τη συστηματική παρακολούθηση της πληροφορίας και την κατανόηση των κοινωνικών κινητοποιήσεων.

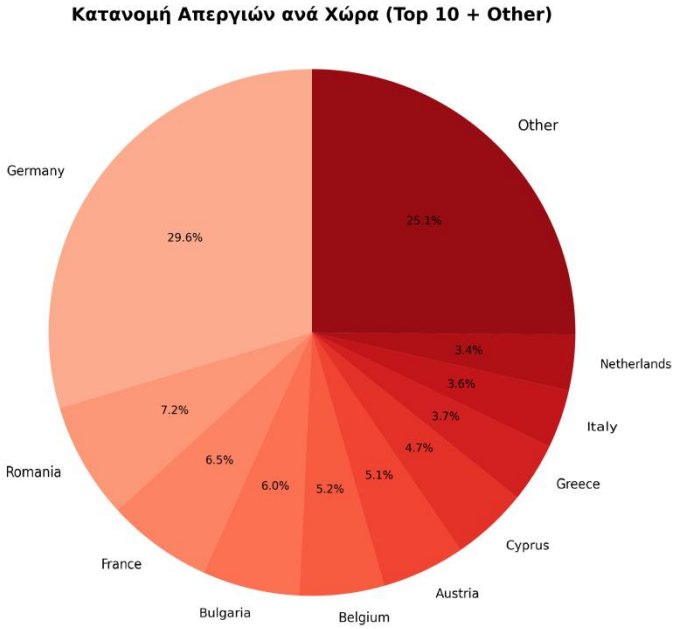
Το σύνολο δεδομένων που συλλέχθηκε και αναλύθηκε αποτελείται από περίπου 1.500 ειδησεογραφικά άρθρα, τα οποία αντλήθηκαν και επεξεργάστηκαν με τη χρήση τεχνικών όπως web crawling, προεκπαιδευμένα μοντέλα NLP και άλλες προηγμένες τεχνολογίες.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

5.1 Χωρική κατανομή κινητοποιήσεων στην Ευρωπαϊκή Ένωση



Εικόνα 15 Η ένταση χρώματος αντιστοιχεί στο πλήθος καταγεγραμμένων κινητοποιήσεων ανά χώρα. Οι συγκρίσεις επηρεάζονται από πληθυσμό και πυκνότητα πηγών.



Εικόνα 16 Κατανομή απεργιών ανά χώρα στην ΕΕ.

Η Εικόνα 15 και η Εικόνα 16 απεικονίζουν την κατανομή και κατάταξη των καταγεγραμμένων κινητοποιήσεων στην Ευρωπαϊκή Ένωση. Στην Εικόνα 15, οι χώρες της Ε.Ε. χρωματίζονται με βάση τον αριθμό των απεργιών που καταγράφηκαν

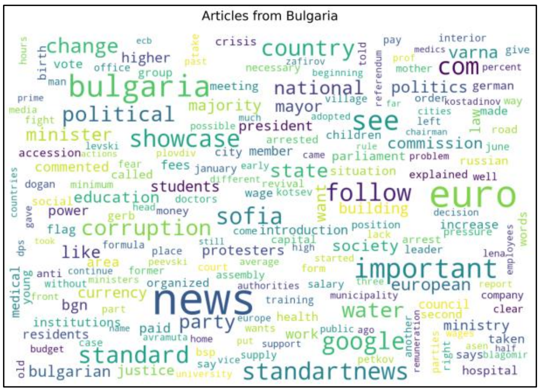
Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

σε κάθε χώρα, η χρήση έντονων χρωμάτων υποδεικνύει υψηλότερο αριθμό απεργιών. Η Εικόνα 16 παρέχει μια ποσοτική κατανομή των διαμαρτυριών ανά χώρα, απεικονίζοντας το ποσοστό συμμετοχής κάθε χώρας στις συνολικές κινητοποιήσεις. Τα δεδομένα επισημαίνουν τις χώρες με τις περισσότερες καταγεγραμμένες απεργίες, δίνοντας έμφαση στις διαφοροποιήσεις μεταξύ των χωρών.

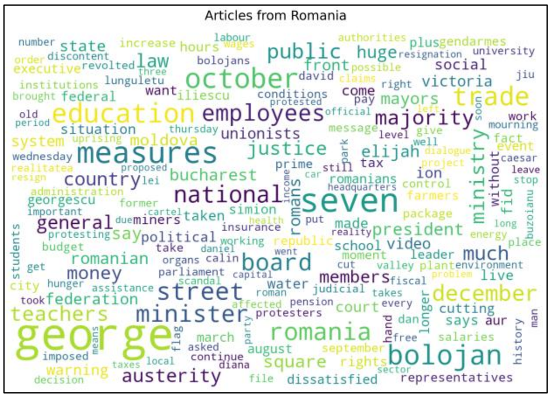
Η Γερμανία καταγράφει το υψηλότερο ποσοστό κινητοποιήσεων στην Ε.Ε., ακολουθούμενη από τη Βουλγαρία και τη Γαλλία, και στη συνέχεια η Ρουμανία, το Βέλγιο και η Αυστρία. Η κλιμάκωση των δεδομένων στις μεγάλες χώρες της Ε.Ε., όπως η Γερμανία, είναι αναμενόμενη, καθώς το πλήθος των κινητοποιήσεων συσχετίζεται με το μέγεθος του πληθυσμού, τη δημοσιογραφική κάλυψη, και την ένταση των κοινωνικών θεμάτων σε αυτές τις χώρες.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Εικόνα 17 Wordcloud απο τα δεδομένα που συλλέχθηκαν για την Γερμανία



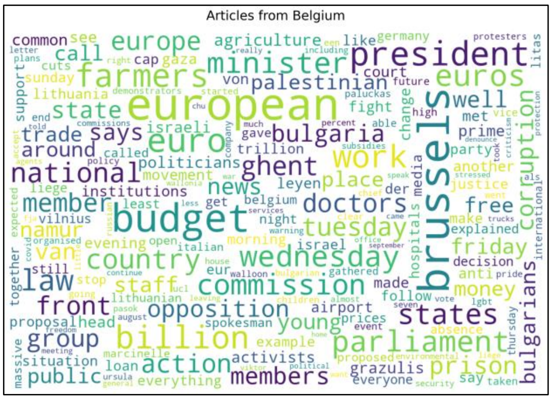
Εικόνα 18 Wordcloud απο τα δεδομένα που συλλέχθηκαν για την Βουλγαρία



Εικόνα 19 Wordcloud από τα δεδομένα συλλέχθηκαν για την Ρουμανία

Εικόνα 20 Wordcloud από τα δεδομένα συλλέχθηκαν για την Γαλλία

Εικόνα 21 Wordcloud από τα δεδομένα συλλέχθηκαν για την Αυστρία



Εικόνα 22 Wordcloud από τα δεδομένα συλλέχθηκαν για το Βέλγιο

Οι Εικόνες Εικόνα 17**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**, REF _Ref209566520 \h Εικόνα 18, Εικόνα 19, Εικόνα 20, Εικόνα 21 και Εικόνα 22, περιλαμβάνουν τα wordclouds των χωρών με τον μεγαλύτερο αριθμό διαμαρτυριών,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

με σκοπό να αναδειχθούν οι αιτίες που τις προκαλεσαν. Το wordcloud για τη Γερμανία, δείχνει έντονη πόλωση γύρω από τον άξονα δικαιωμάτων και αντι-ακροδεξιάς, με λέξεις όπως "Berlin", "CSD", "queer", "participants", "AfD", "anti", "protection", "Bundestag". Αυτή η περίοδος για την Γερμανία, χαρακτηρίζεται απο μαζικές πορείες όπως το Christopher Street Day υποστηρίζοντας τα δικαιώματα των ΛΟΑΤΚΙ και από αντικυβερνητικές και αντι-ακροδεξιές κινητοποιήσεις. Επιπρόσθετα, η παρουσία των όρων όπως "Gaza" και "Israel" υποδηλώνει παράλληλες συναθροίσεις γύρω από τον πόλεμο στη Γάζα, με ταυτόχρονες αντισυγκεντρώσεις.

Σε αντίθεση με την Γερμανία, στη Γαλλία το είδος διαμαρτυριών σχετίζεται κυρίως με εργασικές κινητοποιήσεις. Γεγονός που επαληθεύεται, καθώς στις 3 και 4 Ιουλίου 2025 οι ελεγκτές εναέριας κυκλοφορίας προχώρησαν σε απεργία που προκάλεσε σημαντικές ακυρώσεις και καθυστερήσεις πτήσεων, επηρεάζοντας την ευρωπαϊκή αεροπορική κίνηση. Οι λέξεις που χρησιμοποιούνται στο wordcloud, "france", "social", "budget", "assembly", "public", "march", "teachers", "national", αποτυπώνουν μια ευρύτερη κοινωνικοπολιτική αμφισβήτηση, με δυσφορία για τον προϋπολογισμό και τις δημόσιες δαπάνες, την παιδεία και τα κοινωνικά δικαιώματα, τα οποία αποτελούν λόγους που τροφοδοτούν τις διαμαρτυρίες. Συνολικά, η Γαλλία τον Ιούλιο και τον Αύγουστο 2025 συνδυάζει εργασιακές διεκδικήσεις με κοινωνικές κινητοποιήσεις που ζητούν ενίσχυση του κράτους πρόνοιας.

Το βελγικό wordcloud παρουσιάζει έντονη θεσμική διάσταση, με λέξεις όπως "Brussels", "European", "parliament", "commission", "budget", "corruption", "doctors", "staff", "public", "law". Οι κινητοποιήσεις στο Βέλγιο, βασίζονται λιγότερο σε ευρείες εθνικές γενικές απεργίες και περισσότερο σε τοπικές, αστικές διαμαρτυρίες. Ιδιαίτερη έμφαση δίνεται σε θέματα υγείας, όπως οι απεργίες σε νοσοκομεία με κινητοποιήσεις του ιατρικού προσωπικού και συγκεντρώσεις που εστιάζουν στην ευρωπαϊκή πολιτική, με τις Βρυξέλλες να παραμένουν το κέντρο των θεσμών της Ε.Ε.

Στο αυστριακό wordcloud οι κινητοποιήσεις στο αναφερόμενο διάστημα αντικατοπτρίζεται μια έντονη κοινωνική και πολιτική δυναμική, με έμφαση σε ζητήματα όπως η άνοδος της ακροδεξιάς, η δημόσια υγειονομική ασφάλιση, η αντισημιτική ρητορική και η διεθνής πολιτική κατάσταση, ιδίως σε σχέση με τη σύγκρουση στη Γάζα. Η παρουσία των λέξεων "festival", "activists", "antifascist", "border controls", "security", "doctors", "ÖGK", "Jewish", "antisemitic", "Gaza" στο αυστριακό wordcloud υποδεικνύει την κεντρική θέση αυτών των θεμάτων στις δημόσιες συζητήσεις και τις κινητοποιήσεις της περιόδου.

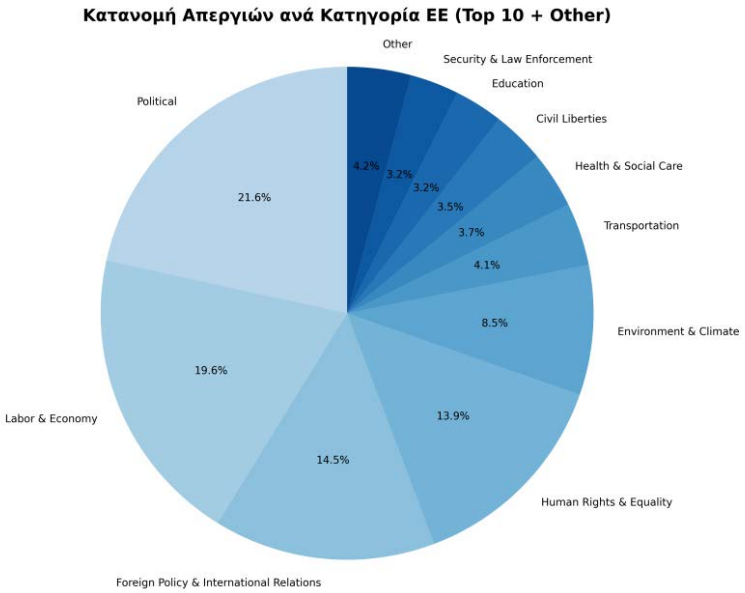
Στο ρουμανικό wordcloud, τονίζονται οι λέξεις "education", "teachers", "measures", "austerity", "public", "employees", "trade", "majority", "law", "national", "square" και "Bucharest". Αποτυπώνει μια οικονομικο-κοινωνική ανησυχία που οδηγεί σε κινητοποιήσεις για την υποστήτιξη της παιδείας, την εργασία και τα μέτρα δημοσιονομικής προσαρμογής. Στο δίμηνο Ιούλιος-Αύγουστος 2025 οι κινητοποιήσεις λαμβάνουν μέρος σε κεντρικές πλατείες του Βουκουρεστίου, ενώ σωματεία εκπαιδευτικών και δημοσίων υπαλλήλων διατηρούν την πίεση μέσω πορειών με αιτήματα για αύξηση των μισθών και χρηματοδοτήσεων.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Στη Βουλγαρία το wordcloud ("Sofia", "corruption", "standard", "euro", "water", "education", "fees", "society", " ustice", "residents", "commission") δείχνει ότι οι διαμαρτυρίες ήταν πολιτικές, κυρίως κατά της διαφθοράς και υπέρ του κράτους δικαίου. Οι διαδηλωτές ζήτησαν την παραίτηση πολιτικών προσώπων που κατηγορούνται για διαφθορά. Κινητοποιήσεις, επίσης υπήρξαν για την απόφαση της κυβέρνησης να υιοθετήσει το ευρώ ("standard", "euro"), καθώς και τοπικά αιτήματα για ζητήματα υποδομών και υπηρεσιών. Τέλος, στα Σόφια και σε άλλες πόλεις πραγματοποιήθηκαν διαδηλώσεις κατά των αυξήσεων στα δίδακτρα και για τη βελτίωση της ποιότητας της εκπαίδευσης. Οι διαδηλωτές ζήτησαν περισσότερους πόρους για τα σχολεία και καλύτερες συνθήκες για τους μαθητές και τους δασκάλους.

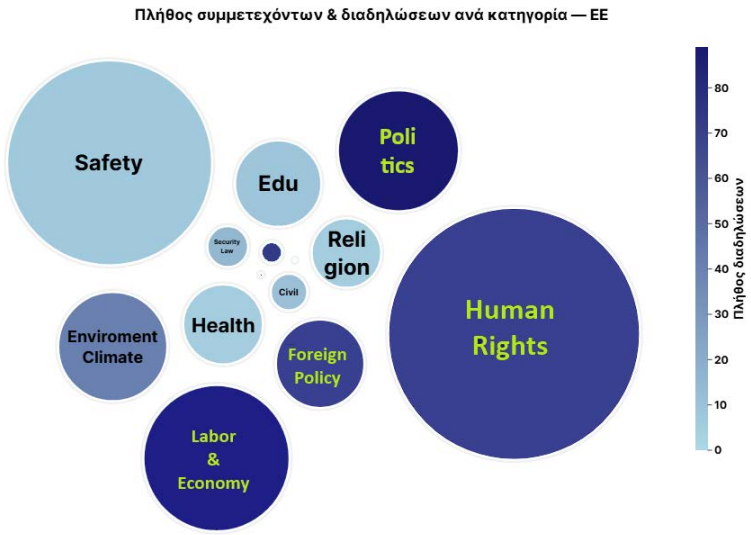
Στο Παράρτημα 2 παρατίθενται τα wordcloud και για τις υπόλοιπες χώρες της Ε.Ε..

5.2 Είδος απεργιών πορειών και διαδηλώσεων



Εικόνα 23 Διάγραμμα πίτας που παρουσιάζει την ποσοστιαία κατανομή των γεγονότων κινητοποιήσεων στην Ε.Ε. ανά θεματική κατηγορία.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250



Εικόνα 24 Γράφημα φυσαλίδων που απεικονίζει την ένταση των θεματικών κινητοποιήσεων στην ΕΕ – το μέγεθος της φυσαλίδας αντιπροσωπεύει τον αριθμό συμμετεχόντων, ενώ η σκίαση (από ανοιχτό σε σκούρο) το πλήθος των γεγονότων.

Οι Εικόνα 23 και Εικόνα 24 παρέχουν μια ολοκληρωμένη απεικόνιση του θεματικού προφίλ των κινητοποιήσεων (απεργίες, πορείες και διαδηλώσεις) σε ολόκληρη την Ευρωπαϊκή Ένωση κατά την εξεταζόμενη περίοδο. Αυτές οι οπτικοποιήσεις επιτρέπουν μια πολυδιάστατη ανάλυση, συνδυάζοντας την ποσοτική κατανομή των θεμάτων με την ένταση της συμμετοχής, και αποτυπώνοντας όχι μόνο τη συχνότητα των γεγονότων αλλά και τη μαζικότητά τους.

Στο διάγραμμα πίτας της Εικόνα 23, το οποίο εστιάζει στην ποσοστιαία κατανομή των γεγονότων ανά θεματική κατηγορία, παρατηρείται μια σαφή ιεράρχηση των προτεραιοτήτων. Οι Πολιτικές κινητοποιήσεις κατέχουν το μεγαλύτερο μερίδιο με 21,6%, αντανακλώντας ευρύτερες πολιτικές εντάσεις, όπως ζητήματα διακυβέρνησης, διαφθοράς ή εκλογικών διαδικασιών. Ακολουθεί η κατηγορία Εργασία & Οικονομία με 19,6%, η οποία περιλαμβάνει διεκδικήσεις σχετικές με μισθούς, εργασιακά δικαιώματα και οικονομικές πολιτικές, συγκεντρώνοντας αθροιστικά με τις Πολιτικές κινητοποιήσεις περίπου το 41% του συνόλου, ένα ποσοστό που υποδηλώνει ότι τα πολιτικά και οικονομικά ζητήματα ήταν βασικοί υπαίτιοι των κοινωνικών αναταραχών στην Ε.Ε. Στη συνέχεια, η Εξωτερική Πολιτική & Διεθνείς Σχέσεις (14,5%) και τα Ανθρώπινα Δικαιώματα & Ισότητα (13,9%) καταγράφουν σημαντικά ποσοστά, αναδεικνύοντας θέματα όπως μετανάστευση, διεθνείς συγκρούσεις και κοινωνική δικαιοσύνη. Χαμηλότερα, αλλά όχι αμελητέα, ποσοστά εμφανίζονται σε κατηγορίες όπως το Περιβάλλον & Κλίμα (8,5%), που σχετίζεται με περιβαλλοντικές πολιτικές και κλιματική αλλαγή, οι Μεταφορές (4,1%) για ζητήματα υποδομών και κινητικότητας, η Υγεία & Κοινωνική Φροντίδα (3,7%) για υγειονομικά και κοινωνικά συστήματα, οι Πολιτικές Ελευθερίες (3,5%) για δικαιώματα έκφρασης και ελευθερίας, η Εκπαίδευση (3,2%) για εκπαιδευτικές μεταρρυθμίσεις, η Ασφάλεια & Επιβολή Νόμου (3,2%) για ζητήματα δημόσιας τάξης, και τέλος οι υπόλοιπες κατηγορίες (4,2%), που καλύπτουν λιγότερο συχνά θέματα όπως πολιτισμός ή τεχνολογία. Αυτή η κατανομή υπογραμμίζει μια συγκέντρωση σε πολιτικο-

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

οικονομικά ζητήματα, ενώ δευτερεύουσες κατηγορίες όπως το Περιβάλλον δείχνουν αυξανόμενο, αλλά ακόμη περιορισμένο, ενδιαφέρον.

Το γράφημα φυσαλίδων στην Εικόνα 24 ενσωματώνει δύο βασικές διαστάσεις. Το μέγεθος κάθε φυσαλίδας αντικατοπτρίζει τον συνολικό αριθμό συμμετεχόντων, ενώ η σκίαση υποδεικνύει τη συχνότητα των γεγονότων. Αυτή η προσέγγιση αποκαλύπτει ασυμμετρίες που δεν είναι άμεσα εμφανείς σε μια απλή ποσοστιαία ανάλυση.

Συγκεκριμένα, η κατηγορία Ανθρώπινα Δικαιώματα & Ισότητα ξεχωρίζει ως η κυρίαρχη, σε συνάρτηση της μαζικότητας, με μια μεγάλη και σχετικά σκουρόχρωμη φυσαλίδα, πράγμα που υποδηλώνει τόσο υψηλή συχνότητα γεγονότων όσο και εξαιρετικά μεγάλη συμμετοχή. Το πιθανότατα οφείλεται σε μαζικά κινήματα για την ισότητα των φύλων, τα δικαιώματα των ΛΟΑΤΚΙ , αντιρατσιστικές διαμαρτυρίες ή υπεράσπισης δικαιωμάτων των μεταναστών.

Η κατηγορία Ασφάλεια & Επιβολή Νόμου παρουσιάζεται με μια μεγάλη αλλά ανοιχτόχρωμη φυσαλίδα, υποδεικνύοντας υψηλή συμμετοχή, αλλά σχετικά λιγότερα γεγονότα συνολικά, όπως σε μεγάλες διαδηλώσεις κατά της αστυνομικής βίας.

Η κατηγορία Περιβάλλον & Κλίμα έχει μια φυσαλίδα με σκίαση πιο σκούρα σε σχέση με το μέγεθός της, κάτι που μεταφράζεται σε περιορισμένο αριθμό διαδηλώσεων, αλλά με σημαντική συγκέντρωση κόσμου ανά εκδήλωση, όπως σε μεγάλες πορείες για την κλιματική αλλαγή.

Οι κατηγορίες Οικονομία & Πολιτική τοποθετούνται σε ενδιάμεσο επίπεδο, με φυσαλίδες μέτριου μεγέθους και σκίασης, αντανακλώντας αξιόλογο αριθμό διοργανώσεων και συμμετοχής, συχνά συνδεδεμένες με οικονομικές κρίσεις ή πολιτικές αλλαγές.

Τέλος, οι μικρότερες και ανοιχτόχρωμες φυσαλίδες, όπως αυτές της Εκπαίδευσης, των Πολιτικών Ελευθεριών και της Ασφάλεια & Επιβολή Νόμου, υποδεικνύουν χαμηλότερη συχνότητα και κλίμακα κινητοποιήσεων, πιθανώς λόγω πιο τοπικού ή εξειδικευμένου χαρακτήρα αυτών των θεμάτων.

Παρά το γεγονός ότι οι Πολιτικές κινητοποιήσεις καταλαμβάνουν το μεγαλύτερο μερίδιο γεγονότων (όπως φαίνεται στο διάγραμμα πίτας), τα Ανθρώπινα Δικαιώματα & Ισότητα εμφανίζουν τη μεγαλύτερη μαζικότητα συμμετοχής (όπως αποτυπώνεται στο γράφημα φυσαλίδων). Αυτό το μοτίβο εξηγείται από τη δομή των κινητοποιήσεων: τα Πολιτικά θέματα διαχέονται σε πολλά, μικρότερου μεγέθους συμβάντα (όπως τοπικές διαμαρτυρίες ή θεματικές συγκεντρώσεις), ενώ τα Ανθρώπινα Δικαιώματα & Ισότητα βασίζονται συχνά σε λιγότερα αλλά πιο μαζικά γεγονότα, όπως δημόσιες συγκεντρώσεις και διαδηλώσεις για την προώθηση κοινωνικών δικαιωμάτων, έχουν χαμηλό ρίσκο συμμετοχής και προσελκύουν ευρύτερα και πιο ετερογενή ακροατήρια.

Συνεπώς, η καταμέτρηση γεγονότων και η καταμέτρηση συμμετεχόντων φωτίζουν διαφορετικές διαστάσεις της κινηματικής έντασης: η πρώτη αναδεικνύει τη συχνότητα και την ποικιλία, ενώ η δεύτερη τη μαζικότητα και την κοινωνική απήχηση.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Συμπερασματικά, η σύγκριση των δύο διαγραμμάτων αποδεικνύει ότι η απλή καταμέτρηση γεγονότων τείνει να υποτιμά κατηγορίες με λίγες αλλά μαζικές κινητοποιήσεις, όπως η Ασφάλεια & Επιβολή Νόμου. Αντίθετα, όταν ενσωματώνεται η συμμετοχή των πολιτών, τα Ανθρώπινα Δικαιώματα & Ισότητα αναδεικνύονται ως η θεματική με τη μεγαλύτερη συνολική μαζικότητα και κοινωνική απήχηση.

Επιπλέον, το Περιβάλλον & Κλίμα φαίνεται να έχει υψηλή συχνότητα γεγονότων, αλλά κατά μέσο όρο μικρότερη κλίμακα συμμετοχής, υποδηλώνοντας ότι, αν και υπάρχει αυξανόμενο ενδιαφέρον, οι κινητοποιήσεις συχνά παραμένουν περιορισμένες σε συμμετοχή σε σύγκριση με άλλα θέματα.

5.3 Ανάλυση του Word Cloud για το Σύνολο των Άρθρων

Εικόνα 25 Wordcloud από το σύνολο δεδομένων που συλλέχθηκαν για την Ε.Ε.

Το wordcloud της Εικόνα 25, αποτελεί αναπαράσταση από το σύνολο των άρθρων της Ε.Ε. και αποτυπώνει τις τάσεις γύρω της. Η έντονη παρουσία των όρων "Berlin", "German", "Germany" επιβεβαιώνει την προηγούμενη παρατήρηση ότι η Γερμανία ξεχωρίζει στο παρόν σύνολο δεδομένων, ως την χώρα με τις περισσότερες διαδηλώσεις και απεργίες ακόμη και όταν εξετάζεται η ευρύτερη ανάλυση όλων των άρθρων.

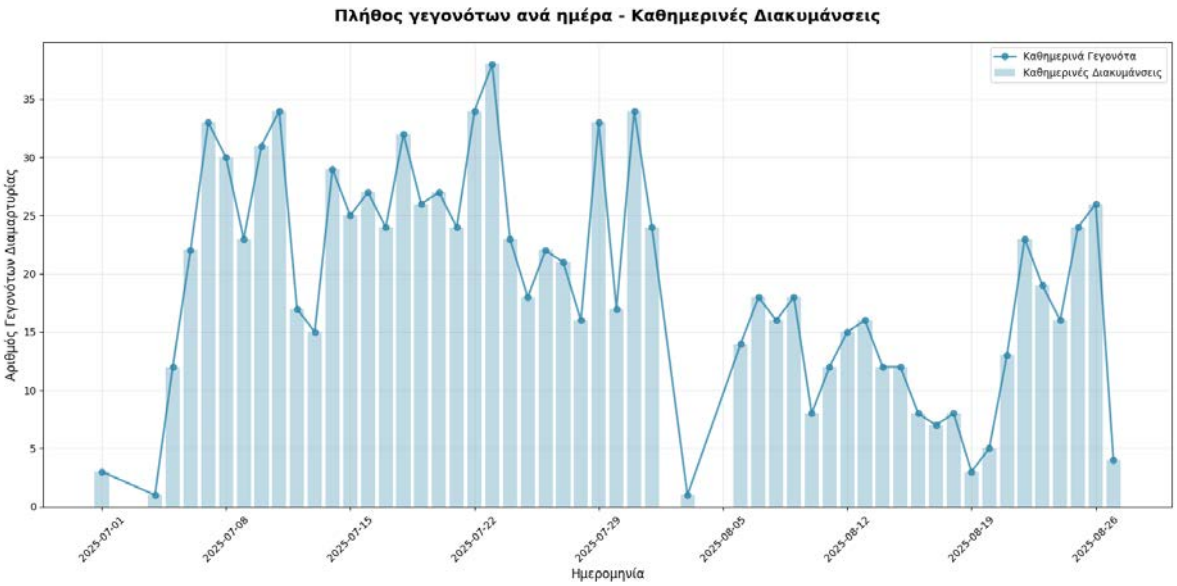
Παράλληλα, η έντονη παρουσία θεσμικών όρων (π.χ. "minister", "president", "party", "court", "law", "state", "national") υπογραμμίζει το πολιτικό και θεσμικό πλαίσιο των κινητοποιήσεων, υποδηλώνοντας συχνές αναφορές σε κυβερνητικούς, κομματικούς και δικαστικούς φορείς, καθώς και σε διαδικασίες λήψης αποφάσεων. Το λεξιλόγιο εργασιακών και οικονομικών θεμάτων (π.χ. "work", "employees", "trade", "labour") είναι εξίσου εμφανές, συνάδοντας με την υψηλή βαρύτητα της κατηγορίας Εργασία & Οικονομία στο διάγραμμα πίτας (19,6%). Τέλος, οι όροι "rights" και " ustice" αντανακλούν την έμφαση στα Ανθρώπινα Δικαιώματα & Ισότητα (13,9%), ενώ η επανάληψη λέξεων όπως "Gaza", "Israel" και "war" αποκαλύπτει τη διείσδυση

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

διεθνών συγκρούσεων στο ευρωπαϊκό πλαίσιο, συνδεόμενη με την κατηγορία Εξωτερική Πολιτική & Διεθνείς Σχέσεις (14,5%).

Συνολικά, το word cloud λειτουργεί ως θεματικός χάρτης, επιβεβαιώνοντας τη κυρίαρχη θέση πολιτικών, εργασιακών και δικαιωματικών αιτημάτων, ενώ φωτίζει το θεσμικό και κοινωνικοοικονομικό υπόβαθρο των γεγονότων. Παράλληλα, αποκαλύπτει πώς διεθνείς συγκρούσεις ενσωματώνονται στον ευρωπαϊκό δημόσιο λόγο, είτε μέσω ενεργειών αλληλεγγύης είτε μέσω διαμαρτυριών για συναφείς πολιτικές, όπως είναι η μετανάστευση.

5.4 Πλήθος γεγονότων διαμαρτυρίας ανά ημέρα



Εικόνα 26 Πλήθος γεγονότων διαμαρτυριών ανά ημέρα

Στην Εικόνα 26 απεικονίζεται το πλήθος των διαμαρτυριών στην Ε.Ε. ανά ημέρα. Οι ράβδοι δείχνουν το πλήθος των απεργιών και η γραμμή την διακύμανσή τους. Το πλήθος των διαμαρτυριών συγκεντρώνεται στον Ιούλιο και προς τα τέλη του αντιστοιχεί η ημέρα με τις περισσότερες διαδηλώσεις, μετά ακριβώς δημιουργείται απότομη αρνητική κλίση.

Αναλυτικότερα:

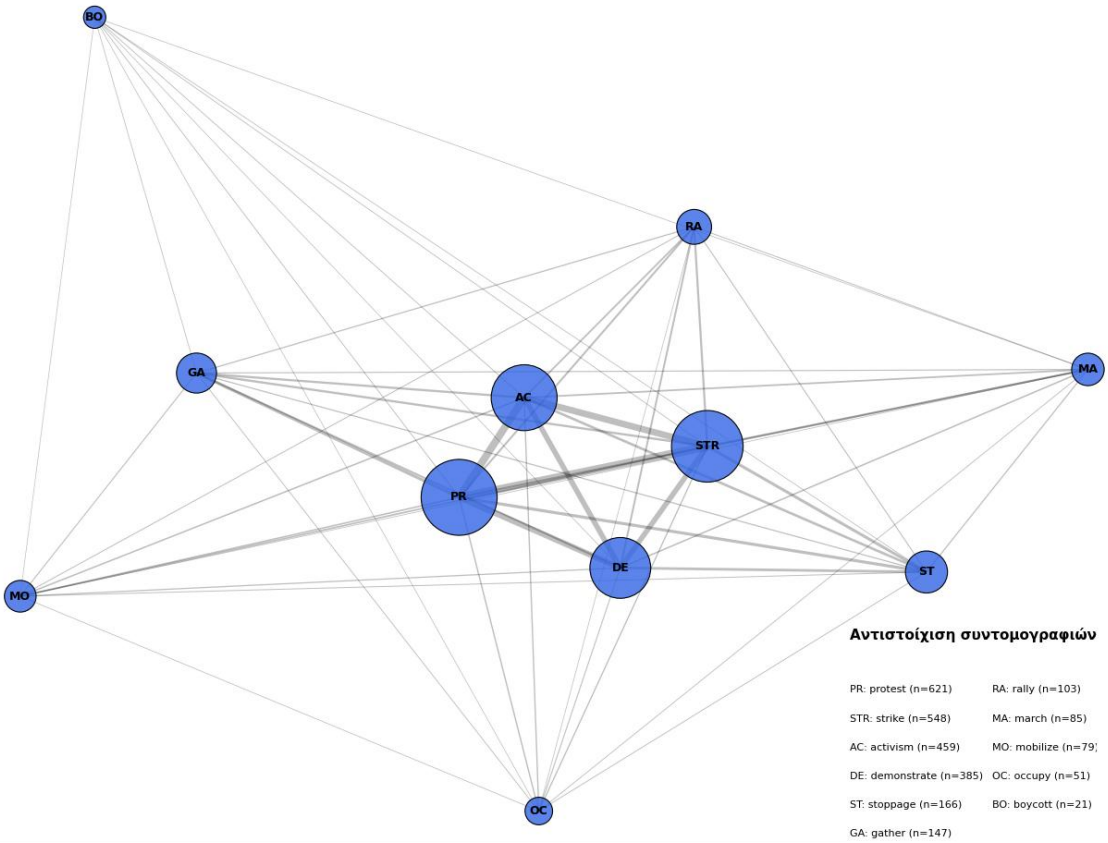
- Η δραστηριότητα διαμαρτυρίας αυξάνεται με ένταση από τις αρχές Ιουλίου και φτάνει σε μια μεγάλη κορύφωση κοντά στα τέλη του μήνα, πιθανόν λόγω γεγονότων που πυροδότησαν ή αύξησαν τη συχνότητα διαμαρτυριών.
- Μετά τη μέγιστη τιμή, υπάρχει απότομη πτώση στον αριθμό των διαμαρτυριών, ενδεχομένως λόγω αποσύνθεσης της κινητοποίησης ή των εξωτερικών συνθηκών που οδήγησαν στη μείωση της δραστηριότητας.

Το μοντέλο αυτό υποδεικνύει ότι οι διαμαρτυρίες στην Ε.Ε. δεν είναι πάντα διαρκείς ή σταθερές, αλλά παρουσιάζουν χαρακτηριστική κυκλική ή αιφνίδια συμπεριφορά,

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

όπου οι κορυφές καταγράφονται κατά την επίδραση συγκεκριμένων κοινωνικών ή πολιτικών γεγονότων.

5.5 Συσχέτιση λέξεων-κλειδιών



Εικόνα 27 Σύνδεση μεταξύ των λέξεων κλειδιών

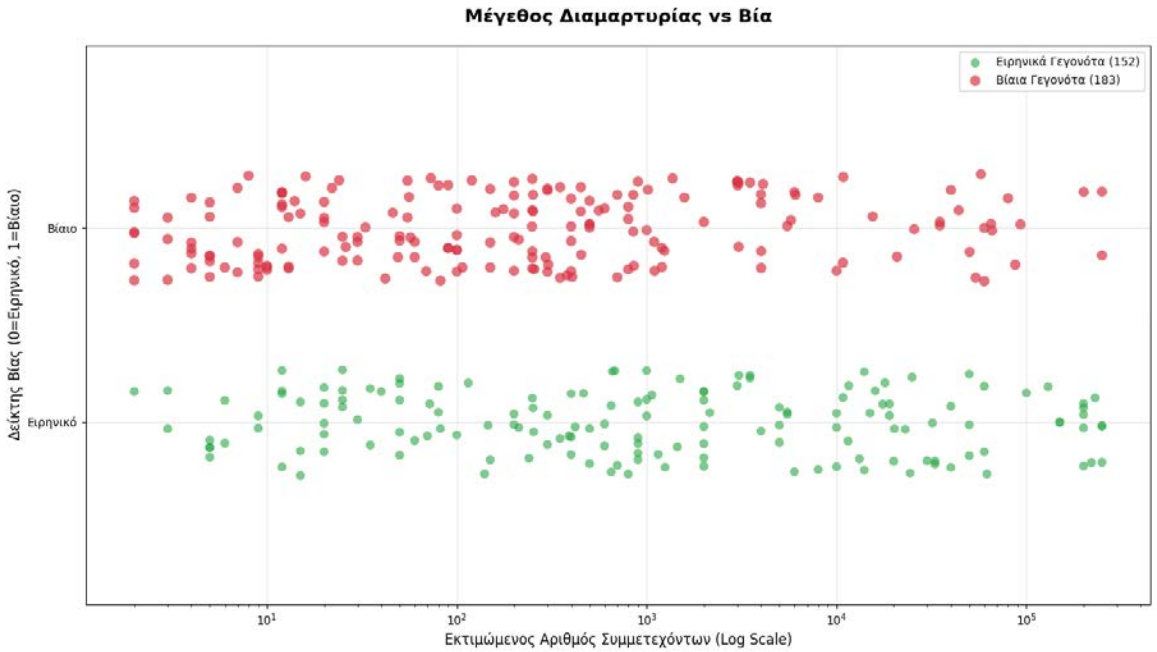
Στην Εικόνα 27, απεικονίζεται ένας γράφος που περιγράφει την σύνδεση και συσχέτιση των λέξεων κλειδιών. Το μέγεθος κάθε κόμβου απεικονίζει τη συχνότητα εμφάνισης του, ενώ οι ακμές απεικονίζουν τη συσχέτιση μεταξύ των λέξεων-κλειδιών Στο κέντρο βρίσκονται οι κόμβοι με τη μεγαλύτερη συχνότητα και τη μεγαλύτερη συσχέτιση σε σχέση με τα ζεύγη λέξεων-κλειδιών. Στο πρώτο επίπεδο, λοιπόν, οι λέξεις "protest", "strike", "activism" και "demonstration" είναι πολύ καλά συνδεδεμένες και εμφανίζονται συχνά. Στο δεύτερο επίπεδο, με μικρότερη συχνότητα εμφάνισης και συσχέτισης, βρίσκονται οι λέξεις "gather", "rally" και "stoppage". Στο τρίτο και τελευταίο επίπεδο ανήκουν οι λέξεις "march", "stoppage", "occupation", "mobilize" και "boycott".

Ο γράφος δείχνει ότι υπάρχουν πολλές συνδέσεις μεταξύ των λέξεων-κλειδιών, κάτι που υποδεικνύει ότι ένα γεγονός διαμαρτυρίας μπορεί να ενσωματώνει πολλές διαστάσεις ή πτυχές. Για παράδειγμα, μια διαμαρτυρία μπορεί να περιλαμβάνει δράσεις όπως "protest", "strike", "activism" και "demonstration", αλλά ταυτόχρονα να συνδέεται και με άλλες μορφές εκφράσεων ή δράσεων όπως "gather", "rally", ή ακόμα και "mobilize". Αυτή η πολυδιάστατη σύνδεση δείχνει την πολυπλοκότητα και

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

την ποικιλία των μορφών διαμαρτυρίας που μπορούν να συνυπάρχουν σε ένα μόνο γεγονός κινητοποίησης.

5.6 Μέγεθος διαμαρτυρίας σε συνάρτηση με την βία



Εικόνα 28 Σύγκριση μεγέθους διαμαρτυρίας και βίας.

Η Εικόνα 28, αποτελεί ένα Scatter plot, στο οποίο απεικονίζεται το μέγεθος της διαμαρτυρίας σε σχέση με την ύπαρξη βίας. Αρχικά, φαίνεται ότι όσο αυξάνεται ο αριθμός των συμμετεχόντων, αυξάνονται οι ειρηνικές διαμαρτυρίες, ενώ για μικρότερες διαμαρτυρίες παρατηρείται το αντίθετο. Ενώ κάποιος θα περίμενε το αντίστροφο, παρατηρείται ότι οι βίαιες διαδηλώσεις συνήθως έχουν λιγότερη συμμετοχή από τις μη βίαιες. Η βία μπορεί να μειώσει τη στήριξη των συμμετεχόντων ή της κοινής γνώμης, υπογραμμίζοντας τη σύνδεση μεταξύ της βίας και της μειωμένης συμμετοχής. Η υποψία εμφάνισης μορφών βίας τρομάζει συνήθως ανθρώπους που διαφορετικά θα συμμετείχαν σε μια ειρηνική διαδήλωση. Όταν η διαδήλωση κλιμακώνεται με βία, η ευρύτερη στήριξη μπορεί να μειωθεί, καθιστώντας την λιγότερο αποτελεσματική.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

6. Συμπεράσματα και προτάσεις για μελλοντική εργασία

6.1 Συμπεράσματα

Η παρούσα εργασία παρουσιάζει ένα καινοτόμο σύστημα αυτοματοποιημένης συλλογής και επεξεργασίας δεδομένων που αφορούν κοινωνικές διαδηλώσεις, όπως απεργίες, πορείες και άλλες μορφές συλλογικής δράσης. Το σύστημα αυτό αποτελεί μια ισχυρή βάση για την ανάπτυξη ενός πλήρως αυτοματοποιημένου εργαλείου, το οποίο όχι μόνο καταγράφει και αναλύει δεδομένα από πολλαπλές πηγές, αλλά και επιτρέπει την παρακολούθηση ακολουθιών γεγονότων σε πραγματικό χρόνο. Για παράδειγμα, μπορεί να εντοπίσει την εξέλιξη μιας απεργίας από την αρχική της κινητοποίηση σε μια συγκεκριμένη περιοχή, λόγω οικονομικών ή πολιτικών αιτιών, μέχρι τις επιπτώσεις της σε ευρύτερο επίπεδο, όπως διαταραχές στην οικονομία ή αλλαγές στην κοινή γνώμη.

Μέσω της χρήσης τεχνολογιών όπως το ML, το NLP και η ανάλυση δεδομένων, το σύστημα επιτυγχάνει υψηλή ακρίβεια στην εξαγωγή κρίσιμων πληροφοριών, όπως ημερομηνίες, τοποθεσίες, αιτίες και συμμετέχοντες. Τα αποτελέσματα δείχνουν ότι η αυτοματοποίηση μειώνει σημαντικά τον χρόνο και το κόστος σε σχέση με χειροκίνητες μεθόδους, ενώ βελτιώνει την αντικειμενικότητα της ανάλυσης. Συνολικά, το έργο αυτό αποδεικνύει τη δυνατότητα του AI να συμβάλει στην κατανόηση κοινωνικών φαινομένων, προσφέροντας εργαλεία για ερευνητές, δημοσιογράφους και πολιτικές αρχές. Παρά τις προκλήσεις, όπως η διαχείριση πολυγλωσσικών δεδομένων και η αντιμετώπιση ψευδών ειδήσεων, το σύστημα θέτει τα θεμέλια για πιο προχωρημένες εφαρμογές στην κοινωνική επιστήμη και την πολιτική ανάλυση.

6.2 Προτάσεις για μελλοντική εργασία

Ενώ το fine-tuning των μοντέλων για την ταξινόμηση και κατηγοριοποίηση των άρθρων διαδραμάτισε έναν καθοριστικό ρόλο στην επιτυχία της εργασίας, καταλήγοντας σε εξαιρετικά ποσοστά ακρίβειας, παραμένει η δυνατότητα για περαιτέρω βελτιστοποίηση και εξέλιξη. Η συνεχιζόμενη βελτίωση της απόδοσης του μοντέλου μπορεί να επιτευχθεί με τη χειροκίνητη επισήμανση (labeling) δεδομένων, η οποία αποτελεί έναν κρίσιμο παράγοντα για την ακριβέστερη εκπαίδευση των μοντέλων. Η χειροκίνητη επισήμανση θα επιτρέψει τη δημιουργία πιο εξειδικευμένων μοντέλων, που θα επικεντρώνονται αποκλειστικά σε γεγονότα διαμαρτυρίας, όπως απεργίες, πορείες και διαδηλώσεις. Ένα ακόμη σημαντικό όφελος από αυτήν την προσέγγιση είναι η δυνατότητα δημιουργίας ενός ελαφρύτερου και πιο αποδοτικού προγράμματος. Με την αποθήκευση μόνο των απαραίτητων και στοχευμένων δεδομένων, οι απαιτήσεις σε αποθηκευτικό χώρο και υπολογιστικούς πόρους μειώνονται σημαντικά.

Επιπλέον, η αποθήκευση στη βάση δεδομένων τόσο των μεταφρασμένων άρθρων όσο και των πρωτότυπων κειμένων τους ανοίγει νέες προοπτικές για τη δημιουργία ενός πολυγλωσσικού μοντέλου. Η συλλογή αυτών των δεδομένων από όλες τις χώρες, περιλαμβάνοντας ποικιλία γλωσσών και πολιτισμικών πλαισίων, παρέχει τη δυνατότητα ανάπτυξης ενός συστήματος που θα μπορεί να επεξεργάζεται δεδομένα

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

σε πολλές γλώσσες ταυτόχρονα. Αυτό το πολυγλωσσικό μοντέλο θα επιτρέπει την άμεση ανάλυση και κατηγοριοποίηση άρθρων χωρίς την ανάγκη για συνεχιζόμενη μετάφραση, η οποία αυτή τη στιγμή είναι μια ιδιαίτερα χρονοβόρα και απαιτητική διαδικασία.

Η αφαίρεση του σταδίου της μετάφρασης θα επιταχύνει σημαντικά τη ροή της ανάλυσης, μειώνοντας το κόστος και τον απαιτούμενο χρόνο για την επεξεργασία των άρθρων. Αυτό θα ενισχύσει την αποδοτικότητα και την ακρίβεια της ανάλυσης των δεδομένων, διευκολύνοντας την ταχύτερη εξαγωγή χρήσιμων πληροφοριών, όπως οι λέξεις-κλειδιά, τα θέματα και τα χαρακτηριστικά των γεγονότων, σε οποιαδήποτε γλώσσα. Ειδικά για περιπτώσεις που εμπλέκονται πολυγλωσσικά δεδομένα, η δυνατότητα του μοντέλου να «κατανοεί» αυτόματα κείμενα σε πολλές γλώσσες θα καθιστά τη διαδικασία πιο εύκολη και αποτελεσματική, προσφέροντας ουσιαστικά εργαλεία για διεθνείς και διαπολιτισμικές αναλύσεις.

Έτσι, η δημιουργία ενός τέτοιου πολυγλωσσικού μοντέλου δε θα διευκολύνει μόνο τη διαχείριση των δεδομένων σε μεγάλες κλίμακες, αλλά θα ανοίξει επίσης τον δρόμο για τη βελτίωση της αλληλεπίδρασης και της συνεργασίας μεταξύ διαφορετικών γλωσσικών και πολιτισμικών ομάδων, κάνοντάς την πολύ πιο αποτελεσματική και άμεση.

Μια σημαντική προοπτική επέκτασης του έργου είναι η ενσωμάτωση λειτουργιών παρακολούθησης σε πραγματικό χρόνο (real-time monitoring), χρησιμοποιώντας APIs από κοινωνικά δίκτυα, όπως Twitter/X και Facebook, για την άμεση καταγραφή των αντιδράσεων και των εμπειριών των πολιτών. Αυτή η προσθήκη θα επιτρέψει την παρακολούθηση των κοινωνικών κινήσεων και των διαδηλωτικών γεγονότων σε πραγματικό χρόνο, προσφέροντας πολύτιμες πληροφορίες για την αντίδραση του κοινού καθώς εξελίσσονται τα γεγονότα. Η ενσωμάτωση αυτών των δεδομένων θα συμβάλλει στην άμεση ανάλυση και στην καλύτερη κατανόηση της δυναμικής των διαμαρτυριών, προσφέροντας έτσι πιο ακριβή και επικαιροποιημένα αποτελέσματα.

Παράλληλα, η εφαρμογή αλγορίθμων πρόβλεψης, όπως η χρονική ανάλυση με LSTM, θα επιτρέψει την πρόβλεψη πιθανών εξελίξεων στα γεγονότα διαμαρτυρίας. Οι αλγόριθμοι αυτοί θα αξιοποιούν ιστορικά δεδομένα και μοτίβα, δίνοντας τη δυνατότητα για προβλέψεις σχετικά με τη διάρκεια, την ένταση ή την εξάπλωση των διαμαρτυριών, προσφέροντας έτσι κρίσιμα εργαλεία για την πρόληψη και διαχείριση κοινωνικών και πολιτικών γεγονότων. Με τον συνδυασμό αυτών των λειτουργιών, το σύστημα θα γίνει πιο δυναμικό και έγκαιρο στην παροχή πληροφορίας, καθιστώντας το χρήσιμο για οργανισμούς που παρακολουθούν κοινωνικά φαινόμενα και πολιτικές εξελίξεις.

Συνοψίζοντας, η εργασία αυτή ανέδειξε τις δυνατότητες αξιοποίησης της επιστήμης των δεδομένων για την κατανόηση και πρόβλεψη κοινωνικών διαδηλώσεων. Οι μελλοντικές επεκτάσεις που προτάθηκαν δεν περιορίζονται μόνο στη βελτίωση της ακρίβειας ή στην ταχύτερη επεξεργασία δεδομένων, αλλά ανοίγουν τον δρόμο για τη δημιουργία ενός πλήρως αυτοματοποιημένου και πολυγλωσσικού συστήματος παρακολούθησης κοινωνικών κινητοποιήσεων. Ένα τέτοιο εργαλείο θα μπορούσε να αποτελέσει ουσιαστικό βοήθημα για ερευνητές, φορείς χάραξης πολιτικής και οργανισμούς της κοινωνίας των πολιτών, ενισχύοντας τη διαφάνεια, την πρόληψη συγκρούσεων και τη βαθύτερη κατανόηση των κοινωνικών διεργασιών.

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Βιβλιογραφία

(Koster), M. K. (2022). 'RFC 9309: Robots Exclusion Protocol'. *IETF Datatracker* .
A., H. D. (1996). "Stemming algorithms: A case study for detailed evaluation". *Journal of the American Society for Information Science* .
ACLED. (2023). *Armed Conflict Location & Event Data Project (ACLED) Dataset*. Ανάκτηση 2025, από https://acleddata.com/
Allan, J. (2000). Topic Detection and Tracking: Event based Information Organization. *Kluwer Academic Publishers* .
Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of research and development, 2(4)* .
Ben Abacha, A. &. (2011). Automatic extraction of semantic relations between medical entities: a rule based approach. . *Journal of biomedical semantics, 2* .
Chang, C. Y. (2025). The Liabilities of Robots.
Cho K., V. M. (2014). On the properties of neural machine translation: Encoder-decoder. *arXiv* .
Consortium, C. C. (n.d.). *Crowd Counting Consortium Dataset*. Ανάκτηση 2025, από https://carnegieendowment.org
Dana R. Fisher, K. T. (2019). The science of contemporary street protest: New efforts in the United States.
Daphi, P. D. (2024). Local protest event analysis: providing a more comprehensive picture? *West European Politics, 48(2)* .
database, G. g. (n.d.). *GeoNames* . Ανάκτηση 2025
Deborah J. Gerner, P. A. (2009). *Conflict and Mediation Event Observations (CAMEO) Codebook.* . Ανάκτηση 2025, από http://eventdata.psu.edu/data.dir/cameo.html
Debouzy, M. (2003). Protest Marches in the United States in the Nineteenth and Twentieth Centuries. *Le Mouvement Social* .
Devlin, J. C. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* .
Deza, M. D. (2009). E. Encyclopedia of distances. In Encyclopedia of Distances. *Springer* .
Explosion . (n.d.). *spaCy*. Ανάκτηση 2025, από https://spacy.io/
Fillieule, O. &. (n.d.). Demonstrations. *Halifax, Can.: Fernwood.*
Google. (n.d.). Ανάκτηση από TensorFlow: https://www.tensorflow.org/
Institute., E. U. (n.d.). *PolDem: Political Conflict and Democracy.* Ανάκτηση 2025
Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* .
Junczys-Dowmunt, M. G. (2018). Marian: Fast neural machine translation in C++. *arXiv* .
Khder, M. A. (2021). "Web scraping or web crawling: State of art, techniques, approaches and application.". *International Journal of Advances in Soft Computing & Its Applications* .
Kriesi, H. J. (2020b). Contention in Times of Crisis: Recession and Political Protest in Thirty European Countries. *Cambridge: Cambridge University Press* .
Kriesi, H. W. (2020). PolDem-Protest Dataset 30 EuropeanCountries.
Laban, P. &. (2017). newsLens: building and visualizing long-ranging news stories. *In Proceedings of the Events and Stories in the News Workshop* .

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Lan, Z. C. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv* .
Leetaru, K. &. (2013). GDELT: Global Data on Events, Location and Tone, 1979–2012 (Version 1.0). *ISA Annual Convention.*
Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*
Lewis, M. L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* .
Liu, B. N. (2017). Growing Story Forest Online from Massive Breaking News.
Liu, B., Han, F. X., Niu, D., Kong, L., Lai, K., & Xu, a. Y. (2020). Story Forest: Extracting Events and Telling Stories from Breaking News. *ACM* .
Liu, Q. Z. (2007). *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models.* (In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2007))
Liu, Y. L. (1997). Tree-to-string alignment template for statistical machine translation. . *International* .
Liu, Y. O. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv* .
Lovins, J. B. (n.d.). “Development of a stemming algorithm”. *Mechanical Translation and Computational Linguistics* .
Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development, 2(2)* .
Manning, C. D. (2008). Introduction to information retrieval. *Syngress Publishing* .
Meyer, P. (2004). The Vanishing Newspaper – Saving Journalism in the Information Age. *Columbia* .
MongoDB. (n.d.). Ανάκτηση από https://www.mongodb.com
Nallapati, R. F. (2004). Event threading within news topics. *In Proceedings of the thirteenth ACM international conference on Information and knowledge management* .
Naseer, S. G. (2021). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance. *Pakistan Journal of Multidisciplinary Research* .
Peace, C. E. (n.d.). *Global Protest Tracker*. Ανάκτηση 2025, από https://carnegieendowment.org/features/global-protest-tracker
Pennington, J. S. (2014). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* .
Persson, E. (2019). "Evaluating tools and techniques for web scraping.".
Qader, W. A. (2019). An overview of bag of words; importance, implementation, applications, and challenges. *In 2019 international engineering conference (IEC)* .
R. Hooper, a. C. (2013). *The Lancaster stemming algorithm*. Ανάκτηση 2025, από http://www.comp.lancs.ac.uk/computing/research/stemming/
Rath, B. P. (2005). Right to Strike: An Analysis. *Indian Journal of Industrial Relations* .
Schrodt, P. A. (2012). Automated coding of political event data. In Handbook of computational approaches to counterterrorism . *NY: Springer New York.*
Sirisuriya, D. S. (2015). "A comparative study on web scraping.".
Stanford AI Lab. (n.d.). *Stanford Natural Language Processing Group*. Ανάκτηση από https://nlp.stanford.edu/
Strauch, C. S. (2024). NoSQL databases. *Stuttgart Media University* .

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Sukhbaatar, S. W. (2015). End-to-end memory networks. Advances in neural information processing systems. *Advances in neural information processing systems 28* .
Sutskever, I. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems* .
T. Saito, M. R. (2017). *„Bas c evaluat on measures from the confus on matr x.”*. Ανάκτηση από WordPress.
T. Yao, Z. Z. (2020). "Text Classification Model Based on fastText". *IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)* .
Vaswani, A. S. (2017). Attention is all you need. *Advances in neural information processing systems* .
Vu ović, Ž. (2021). *Classification model evaluation metrics.* Ανάκτηση από International Journal of Advanced Computer Science and Applications.
W. A. Qader, M. M. (2019). "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges". *International Engineering Conference (IEC)* .
Wang, J. &. (2020). Measurement of Text Similarity: A Survey. *Information* .
Wheeler, P. J. (1984). Changes and improvements to the European Commisssion’s Systran system 1976/84. *In Proceedings of the International Conference on Methodology and Techniques of Machine Translation: Processing from words to language.*
X. Han, Z. Z. (2021). “Pre-trained models: Past, present and future”. *AI Open* .
Yadav, D. S. (2010). *Users search trends on WWW and their analysis*. *Intelligent Interactive Technologies and Multimedia* .
Yang Sun, I. G. (2010). ‘The Ethicality of Web Crawlers’. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (IEEE 2010).*
Yasin N. Silva, I. A. (2016). SQL: From Traditional Databases to Big Data. *In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE '16). Association for Computing Machinery, New York, NY, USA* .
Yu, D. I. (2023). ‘Donottrain: A Metadata Standard for Indicating Consent for Machine Learning’. *Proceedings of the 40th International Conference on Machine Learning* .
Zhang, C. L. (2023). A storytree-based model for inter-document causal relation extraction from news articles. *Knowl Inf Syst 65* .
Zhao, B. (2017). "Web scraping.". *Encyclopedia of big data. Springer* .
Zweigenbaum, A. B. (2011). Medical Entity Recognition: A Comparaison of Semantic and Statistical Methods. *Association for Computational Linguistics.*

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Παράρτημα 1

violence	violent	attack	attacked	fighting	fights	clash	clashes
riot	riots	confrontation	confrontations	arrest	arrested	arrests	detain
detained	detention	injured	injury	injuries	hurt	wounded	casualty
casualties	tear gas	pepper spray	rubber bullets	water cannon			

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Παράρτημα 2

Articles from Croatia

Articles from Cyprus

Articles from Czech Republic

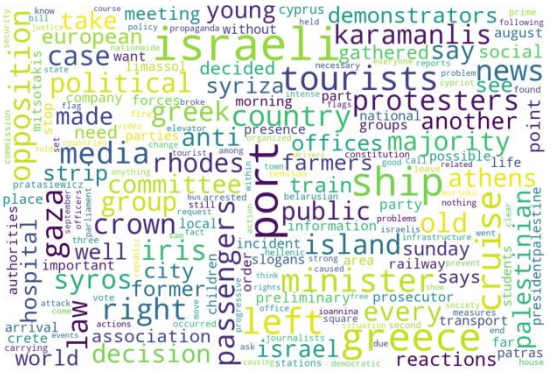
Articles from Denmark

Articles from Estonia

Articles from Finland

Articles from Finland

Articles from Greece

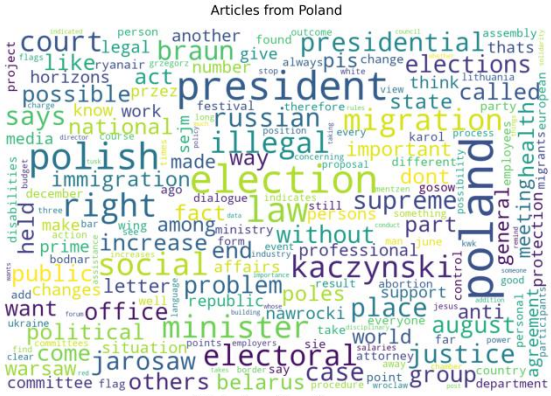


Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250

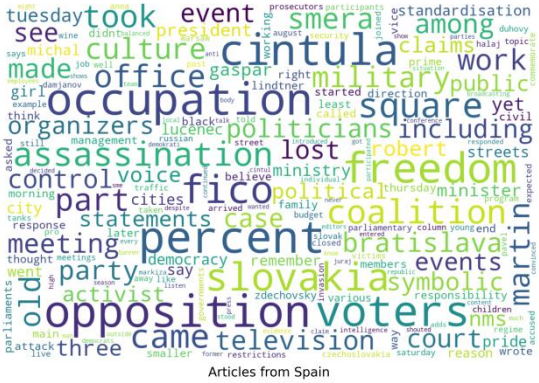
Articles from Netherlands

Articles from Poland



Articles from Portugal

Articles from Slovakia



Articles from Slovenia

Articles from Spain

Articles from Sweden

Institutional Repository - Library & Information Centre - University of Thessaly
06/09/2026 04:45:16 EEST - 216.73.216.250