Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Δυναμική κατηγοριοποίηση μεγάλων δεδομένων κρίσεων με τη χρήση συνελικτικών νευρωνικών δικτύων

Διπλωματικη Εργασια

του

ΣΤΑΜΑΤΙΟΥ ΑΝΟΥΣΤΗ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

Δυναμική κατηγοριοποίηση μεγάλων δεδομένων κρίσεων με τη χρήση συνελικτικών νευρωνικών δικτύων

Διπλωματικη Εργασια

του

ΣΤΑΜΑΤΙΟΥ ΑΝΟΥΣΤΗ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 11η Ιουλίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Θεοδώρα Βαρβαρίγου	Ε. Βαρβαρίγος	Σ. Παπαβασιλείου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2019

(Υπογραφή)

.....

Σταματιος Ανουστης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Μεγάλα δεδομένα παράγονται κατά την εμφάνιση απροσδόκητων καταστροφικών γεγονότων, και αυτό θέτει νέες προκλήσεις στην χρονικά καθοριστική ανάλυση δεδομένων και στις τεχνικές επιβλεπόμενης μάθησης. Τα κοινωνικά δίκτυα αναγνωρίζονται όλο και συχνότερα ως μέσα αρωγής των δράσεων διάσωσης και αποκατάστασης συνεισφέροντας στην επίγνωση κατάστασης κατά τη διάρκεια φαινομένων μαζικής εκτάκτου ανάγκης. Είναι κοινή πρακτική η χρήση μεθόδων επιβλεπόμενης μάθησης για να ερμηνευτούν μεγάλα σύνολα δεδομένων ως προς τους διάφορους τύπους πληροφοριών που διακινούνται μέσω σύντομων μηνυμάτων υπηρεσιών μικροϊστολογίου τύπου Twitter. Υπάρχει ενδιαφέρον από πλευράς κυβερνητικών υπηρεσιών, ΜΚΟ, οργανισμών δημόσιας υγίειας και ανθρωπιστικών οργανώσεων να ερευνηθεί η δυναμική των κοινωνικών δικτύων στην παροχή ανθρωπιστικής βοήθειας μέσω ενός δεδομενο-κεντρικού τρόπου κατά διαχείριση του πληροφοριακού φόρτου και της καλύτερης επιβεβαίωσης και διαλογής των πληροφοριών βάσει προτεραιοτήτων για τους σκοπούς ενός εκάστου. Στο πλαίσιο της κατηγοριοποίησης σύντομων μηνυμάτων κρίσεων, η εφαρμογή νευρωνικών ταξινομητών είναι καινοτόμα.

Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη αξιόπιστων και αποδότικων, σε ότι αφορά τη μείωση του κόστους και του χρόνου επεξεργασίας, μεθόδων ταξινόμησης μεγάλων δεδομένων κρίσεων, βασιζόμενων σε βαθέα νευρωνικά δίκτυα . Προτείνουμε δύο αρχιτεκτονικές βαθέων νευρωνικών δικτύων, που ανήκουν στη ομάδα των συνελικτικών εμπροσθόδρομων, για το έργο της κατηγοριοποίησης των μικρο-κειμενικών δεδομένων κρίσεων. Απεφεύχθη κάθε προσπάθεια εργο-εξειδικευμένης μηχανικής, τόσο στην προεπεξεργασία των δεδομένων όσο και στη σχεδίαση χαρακτηριστικών, καθώς επίσης αποφεύχθηκε και η χρήση εξωτερικών προ-επιμελημένων βάσεων γνώσης. Αξιοποιώντας τα παραπάνω μοντέλα διεξήχθη πειραματική μελέτη έναντι πραγματικών δεδομένων από παλαιότερες καταστροφές. Ελλείψει ικανού αριθμού επισημειώσεων, συμπεριελήφθη μεταφορά γνώσης από ιστορικά δεδομένα προσαρμοσμένα σε κάθε τύπο καταστροφής ξεχωριστά. Το συμπερασμά αυτής της εργασίας είναι πως τα συνελικτικά νευρωνικά δίκτυα μπορούν να αντιμετωπίσουν αποτελεσματικά το πρόβλημα της ταχείας ανάλυσης μεγάλων δεδομένων κρίσεων και είναι πράγματι μια εφικτή και πολλά υποσχόμενη λύση.

Λέξεις Κλειδιά

Twitter, μεγάλα δεδομένα, κρίση, διαχείριση καταστροφών, επίγνωση κατάστασης, βαθέα

νευρωνικά δίκτυα, επιβλεπόμενη μάθηση, ταξινόμηση κειμένου.

Abstract

Big data are produced on the onset of unexpected disastrous events which brings challenges in time-critical data analysis and supervised learning techniques. Social media is increasingly acknowledged as a conduit to alleviate the rescue and restore actions and to raise situational awareness during mass emergence events. A common practice is to use supervised learning methods for making sense out of voluminous data-sets for the various types of information disseminated through short messages of micro-blogging platforms such as Twitter. There is an interest from government agencies, NGO's, public health sectors and humanitarian organizations in investigating the potential of social media for humanitarian aid in a data-driven manner that is handling the overload ,better validating and prioritizing the most tactical of the messages for each one's purpose. Supervised predictive models based on deep neural networks have been recently used in various applications achieving remarkable results, among others, in object and speech recognition. Despite these recent advancements, in the context of crisis-related data categorization, the application of neural network based classifiers is novel.

This thesis mainly focuses on developing neural network based classifiers which will enable a reliable and effective, as regards reducing the cost and processing time, classification of big crisis data. We proposed two unified deep neural network architectures, belonging to the class of convolutional feed-forward networks, for the task of short-text crisis data categorization. We intentionally avoid task-specific engineering, in data preprocessing and feature design, as well as the usage of hand-coded external knowledge resources. By utilizing the aforementioned models, an experimental study was conducted against real data from past disasters. Due to the limited amount of labeled data, some form of transfer learning has been involved by utilizing data-sets from past events adapted for each specific type of crisis. The overall conclusion of this thesis is that convolutional neural networks can effectively address the problem of real-time analysis of crisis-related data and it is indeed a viable and a very promising solution.

Keywords

Twitter, big data, crisis, disaster management, situational awareness, deep neural networks, supervised learning, text classification.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια κ. Θεοδώρα Βαρβαρίγου για την στήριξη της προσπάθειας εκπόνησης της διπλωματικής εργασίας.

Επίσης ευχαριστώ ιδιαίτερα τον μεταδιδακτορικό ερευνητή κ. Γεώργιο Κουσιουρή για την καθοδήγηση και την συμβουλευτική του καθόλη τη διάρκεια της εργασίας.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου, που πάντα με στηρίζουν σε κάθε μου βήμα.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Σ αυτή την εισαγωγική ενότητα θα αναπτύξουμε όσα παραθέσαμε στη περίληψη στο βαθμό που ο αναγνώστης να μπορεί να έχει μια ολοκληρωμένη εικόνα του έργου. Θα εξετάσουμε το πρόβλημα που μελετά η παρούσα εργασία τους περιορισμούς ή τις τεχνικές δυσκολίες που ανακύπτουν και τους τρόπους προσέγγισης και επίλυσης που επιχειρεί.

1.1 Αντικείμενο της διπλωματικής

Πολύ έρευνα διεξάγεται γύρω από την αξιοποίηση και την διαχείρηση μεγάλων συνόλων δεδομένων. Ένα μεγάλο μέρος της πληροφορίας που εμπεριέχουν αυτά τα σύνολα μένει αναξιοποίητο χώρις τη κατάλληλη ανάλυση. Μια πηγή τέτοιων συνόλων δεδομένων είναι τα μέσα κοινωνικής δικτύωσης που χρησιμοποιούνται ευρέως.

Σε πεδία που έχουν κοινωνικό αντίκτυπο τα μέσα κοινωνικής δικτύωσης είναι μια φυσική πηγή πληροφόρησης όπως θέματα που άπτονται της ασφάλειας των πολιτών στη διαχείριση έκρυθμων καταστάσεων που μπορεί να προκαλέσει μια φυσική καταστροφή. Το κόστος σε ανθρώπινες ζωές αλλά και σε οικονομικούς όρους που καλείται να αντιμετωπίσει μία κοινωνία που πλήττεται απο μία φυσική καταστροφή έιναι δυσθεώρητο. Η έγκαιρη προειδοποίηση κι ο μετέπειτα καλός συντονισμός της προσπάθειας αντιμετώπισης της καταστροφής προϋποθέτει την ακώλητη και ταχεία διακίνηση των χρήσιμων πληροφοριών.

Τα μέσα κοινωνικής δικτύωσης κάνουν εκτεταμένα τη εμφάνισή τους ως νέες πηγές πληροφόρησης και ταχείας επικοινωνίας και συγκεκριμένα κατα τη διάρκεια φυσικών καταστροφών. Η πλατφόρμα Twitter είναι μια υπηρεσία που επιτρέπει στους χρήστες της να δημοσιεύουν σύντομα κείμενα μέχρι 140 χαρακτήρες σε ένα κοινό από ακολούθους που χρησιμοποιούν τη πλατφόρμα διαδικτυακού ή κινητού λογισμικού. Η καθολική χρήση της δημοφιλούς αυτής υπηρεσίας, η ταχύτητα επικοινωνίας και η διαπλατφορμική προσβασιμότητα που προσφέρει την καθιστούν το πλέον κατάλληλο μέσο για την διάδοση πληροφοριών που αφορούν την καταστροφή.

Η έγκαιρη ανάλυση των μηνυμάτων που δημοσιεύουν οι χρήστες της υπηρεσίας Twitter κατά την εμφάνιση μιας κρίσης προσφέρει χρήσιμα συμπεράσματα για την εξέλιξη της κατάστασης μιας γενικευμένης καταστροφής και μπορεί να βοηθήσει στην λήψη κρίσιμων αποφάσεων.

Ανθρωπιστικές οργανώσεις όπως τα Ηνωμένα Έθνη χρησιμοποιούν αυτή τη πλατφόρμα για να αποκτήσουν γρήγορη πρόσβαση σε μηνύματα που μπορεί να αναφέρουν επείγουσες ανάγκες, αγνοούμενους ή τραυματίες, καταστροφές και φθορές σε κτήρια και υποδομές. Η γρήγορη αντιμετώπιση μιας κρίσης ενέχει σε μεγάλο βαθμό την ικανότητα επεξεργασίας αυτών των μηνυμάτων το συντομότερο ή και σε πραγματικό χρόνο. Ωστόσο, αυτό εξαιτίας της φύσεως των μεγάλων δεδομένων ξεπερνά συχνά τα όρια της ανθρώπινης «επεξεργαστικής ισχύος».

Ένα πρώτο βήμα στην ανάλυση αυτών των δεδομένων είναι η κατηγοριοποίηση τους , σύμφωνα με το συγκεκριμένο τύπο πληροφορίας που μεταφέρουν. Αυτό επιτρέπει στις ανθρωπιστικές οργανώσεις ανάλογα με τις ανάγκες που έχουν να αναζητήσουν καλύτερα τις πληροφορίες που επιζητούν και που τους είναι απαραίτητες για να λάβουν δράση και αποφάσεις για να σώσουν ανθρώπινες ζωες, να περιορίσουν τον πόνο των πληττόμενων πληθυσμών και να αποκαταστήσουν τις φθορές που αφήνει στον απόηχο της μια μεγάλη καταστροφή. Η ανάκτηση των πληροφοριών ενδιαφέροντος βάσει λέξεων κλειδιών πολύ συχνά αποτυγχάνει να διακρίνει τα μηνύματα που είναι πράγματι χρήσιμα και έχουν κατί ουσιαστικό να προσθέσουν στην επίγνωση της κατάστασης που επικρατεί. Ακόμα και για τον άνθρωπο είναι δύσκολο ένα τέτοιο έργο δεδομένου ότι τα μηνύματα είναι πολύ συνοπτικά και με εκφορά έντονα προφορική και ανεπίσημη. Από κει και πέρα η σωστή ερμηνεία τους και κατηγοριοποίηση τους είναι σε κάποιο βαθμό υποκειμενική και συνεπώς ενέχει μια υψηλού επιπέδου γνωσιακή διαδικασία . Η καλύτερη γνωστή λύση είναι η χρήση μεθόδων επιβλεπόμενης μάθησης χρησιμοποιώντας ταξινομημένα παραδείγματα από παλαιότερες καταστροφές.

Η ταξινόμηση των μικρο-κειμενικών αυτών δεδομένων παρουσιάζει σημαντικές προκλήσεις στις μεθόδους και αλγορίθμους μηχανικής μάθησης και ειδικά σε εκείνες που χρησιμοποιούν επιβλεπόμενη μάθηση. Ο μεγάλος όγκος και η ταχύτητα των δεδομένων αυτών απαιτεί απλά μοντέλα που θα μπορούν σε πραγματικό χρόνο και με λίγη «επίβλεψη» να γενικεύσουν εύκολα σε νέα άγνωστα πρότυπα. Ένας άλλος σημαντικός περιορισμός είναι η έλλειψη επισημειωμένων παραδειγμάτων τις πρώτες ώρες έξαρσης της κρίσης. Κατά συνέπεια, είμαστε αναγκασμένοι να προβλέπουμε βάσει μοντέλων που εκπαιδεύτηκαν σε δεδομένα από πολλά και διαφορετικά παλαιότερα γεγονότα. Ο υψηλός θόρυβος λόγω μεγάλης «ποικιλίας» μεταξύ διαφορετικών γεγονότων πρέπει επίσης να ληφθεί υπόψη γιατί κάθε κρίση είναι μοναδική, απρόβλεπτη και εξαρτάται από πολλούς και ανεξάρτητους παράγοντες και μπορεί να οδηγήσει σε υποβάθμιση της αξιοπιστίας των συστημάτων. Χρειαζόμαστε συνεπώς μοντέλα που να είναι απλά και χρονικά αποδοτικά αλλά συνάμα εκφραστικά και σύνθετα.

Τα βαθιά νευρωνικά δίκτυα είναι αναμφισβήτητα αρκετα δημοφιλή και συγκεκριμένα το είδος των συνελικτικών νευρωνικών δικτύων έχει αναγνωριστεί ως πολλά υποσχόμενο και κατάλληλο για εφαρμογές όπου η χρονική αποδοτικότητα είναι κρίσιμη. Η χρήση εξειδικευμένου υλισμικού και της υπολογιστικής νέφους κατέστησε δυνατή πρόσφατα την αποδοτική εκπαίδευση αυτών των αλγορίθμων αξιοποιώντας την υψηλή παραλληλοποιησιμότητα της συνέλιξης με την οποία αυτοί οι αλγόριθμοι εξάγουν τοπικά χαρακτηριστικά από ακολουθίες λέξεων μεταβλητού μήκους τα οποία στη συνέχεια συνθέτουν για να τις ταξινομήσουν κατά το «νοηματικό» τους περιεχόμενο. Συγχρόνως η χρήση κατανεμημένων αναπαραστάσεων ενσωματώσιμων χαρακτηριστικών και ειδικότερα προεκπαιδευμένων αναπαραστάσεων έχουν

επιτύχει state-of-the-art αποτελέσματα τόσο σε εφαρμογές υπολογιστικής όρασης όσο και σε εφαρμογές επεξεργασίας φυσικής γλώσσας λ.χ αναγνώριση συναισθήματος. Η προσέγγιση που προτείνουμε επιτρέπει την καλύτερη αξιοποίηση των επισημειωμένων ιστορικών δεδομένων άλλων καταστροφών λόγω καλής γενίκευσης των νευρωνικών μοντέλων. Τα βαθιά νευρωνικά δίκτυα και η μάθηση αναπαράστσης είναι μια μοντέρνα προσέγγιση στο πρόβλημα αφού οι αλγόριθμοι εξάγουν αυτόματα τα χαρακτηριστικά που τους είναι απαραίτητα για την κατηγοριοποίηση των δεδομένων αυτών χώρις να παρεμβαίνει ο άνθρωπος άμεσα στη σχεδιάση των χαρακτηριστικών.

Στη παρούσα μελέτη θα επιχειρήσουμε να εφαρμόσουμε, να προσαρμόσουμε και να συγκρίνουμε δύο δοκιμασμένες οικογένειες αλγορίθμων που χρησιμοποιούν συνελικτικά νευρωνικά δίκτυα , που έχουν εφαρμοστεί με σημαντική επιτυχία σε αναγνώριση συναισθήματος σε μικρο-κειμενικά δεδομένα κοινωνικών δικτύων, επιλύοντας όμως το πρόβλημα της κατηγοριοποίησης και ανάκτησης των μηνυμάτων που περιέχουν πληροφορίες που μπορούν να καθορίσουν δράσεις ή αποφάσεις και να προσφέρουν έτσι πολύτιμη ανθρωπιστική βοήθεια . Πρόκειται για τα συνελικτικά νευρωνικά δίκτυα που έχουν χρησιμοποιήσει οι Yoon Kim et al. [13] και Kalchbrener et al. [11] με μεγιστοποιητικό συνδυασμό χαρακτηριστικών στατικό και δυναμικό αντίστοιχα. Αυτά τα δίκτυα εξάγουν χαρακτηριστικά μεγαλύτερου εύρους και υψηλότερου επιπέδου από τις μεθόδους σακιδίου λέξεων που παραδοσιακά χρησιμοποιούνται σε εφαρμογές κατηγοριοποίησης κειμένου. Πειραματιστήκαμε με διάφορες μεθόδους μείωσης του φαινομένου της υπερπροσαρμογής που συχνά αντιμετωπίζουν αυτοί οι αλγόριθμοι όπως η μέθοδος πολλαπλών καναλιών αναπαραστάσης και η μέθοδος παράλειψης χαρακτηριστικών dropout. Τέλος χρησιμοποιήσαμε μεθόδους προσαρμογής, ανά κατηγορία καταστροφής, για να μειώσουμε την «ποικιλία» του πληθυσμού των δεδομένων εκπαίδευσης.

1.1.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

1. Συγκεντρώσαμε και επιμεληθήκαμε δεδομένα από έτοιμες σύλλογές επισημειωμένων μηνύματων από μεγάλες ανθρωπιστικές κρίσεις και φυσικές καταστροφές [9][2][23] για τους σκοπούς της εκπαίδευσης των αλγορίθμων.

2. Αποτιμήσαμε την αξιοπιστία του συστήματος κατηγοριοποίησης των μικρο-κειμενικών δεδομένων κρίσεων με βάση συνελικτικά νευρωνικά δίκτυα [11][13] και συγκρίναμε τις επιδόδεις τους με έναν κλασσικό αλγόριθμο κατηγοριοποίησης κειμένου βασισμένο στον αλγόριθμο SVM [32].

3. Πειραματιστήκαμε με διάφορα σύνολα προεκπαιδευμένων ενσωματώσιμων χαρακτηριστικών [18][26], και την αξιοποίηση πολλαπλών καναλιών αναπαράστασης της εισόδου.

4. Εξετάσαμε την επίδραση της ασσυμετρίας των τάξεων στην ικανότητα γενίκευσης των αλγορίθμων.

5. Μελετήσαμε τα πλεονεκτήματα των τεχνικών προσαρμογής και μεταφοράς γνώσης ενσωματώνοντας δεδομένα από παλαιότερες φυσικές καταστροφές.

1.2 Οργάνωση του τόμου

Εργασίες σχετικές με το αντικείμενο της διπλωματικής παρουσιάζονται στο Κεφάλαιο 2. Το Κεφάλαιο 3 αναφέρεται στο θεωρητικό υπόβαθρο. Στο Κεφάλαιο 4 αναπτύσσουμε το πρόβλημα της διαχείρισης ανθρωπιστικών κρίσεων με την ανάλυση δεδομένων κοινωνικών δικτύων. Στο Κεφάλαιο 5 συζητείται η λύση που προτείνει η παρούσα εργασία. Η πειραματική διαδικασία και τα συμπεράσματα της περιγράφονται στο Κεφάλαιο 6. Τεχνικές λεπτομέριες της υλοποίησης του συστήματος αναφέρονται στο Κεφάλαιο 7. Τα συμπεράσματα της εργασίας και μελλοντικά βήματα που μπορούν να ακολουθηθούν συζητούνται στο Κεφάλαιο 8.

Κεφάλαιο 2

Συγγενικές εργασίες

2.1 Εισαγωγή

Σ αυτή την ενότητα θα περιγράψουμε συνοπτικά τις δύο ευρύτερες περιοχές έρευνας από τις οποίες αντλήσαμε αρχές και μεθοδολογίες καθώς επίσης δεδομένα και αλγορίθμους. Η πρώτη θεματική, διαχείριση κρίσεων και καταστροφών (emergency management) αφορά τον τρόπο οργάνωσης και διαχείρησης των πόρων και των αρμοδιοτήτων για την αντιμετώπιση όλων των πτυχών μίας έκτακτης κατάστασης (emergency) που μπορεί να εξελιχθεί σε καταστροφή (disaster). Η άλλη θεματική αφορά τα βαθιά νευρωνικά δίκτυα και τη μάθηση αναπαράστασης (Deep learning and representation learning) κατά την εφαρμογή της στην επεξεργασία φυσικής γλώσσας (natural language processing).

2.2 Διαχείριση κρίσεων και καταστροφών

Η χρήση των κοινωνικών δικτύων ως πηγή ελεύθερων δεδομένων η ανάλυση των οποίων μπορεί να αποκαλύψει τάσεις στη κοινωνία, ή να περιγράψει καταστάσεις που λαμβάνουν χώρα στο φυσικό περιβάλλον ή να προβλέψει γεγονότα που επηρεάζουν το σύνολο των μελών μιας κοινωνίας είναι διαδεδομένη [35, 12]. Απρόβλεπτα γεγονότα μαζικής έκτακτης αναγκής όπως φυσικές καταστροφές, μεγάλα ατυχήματα και τρομοκρατικές επιθέσεις που απειλούν ανθρώπινες ζωές έχει παρατηρηθεί ότι ευαισθητοποιούν ένα μεγάλο μέρος του κοινωνικού συνόλου που χρησιμοποιεί τα μέσα κοινωνικής δικτύωσης ως δίαυλο επικοινωνίας για τη διάδοση ειδήσεων πρωτότυπου περιεχόμενου σχετικών με το συμβάν [8, 25, 7, 33]. Η μελέτη της συμπεριφοράς των χρηστών του μικροϊστολογίου Twitter στη διάρκεια διαφορετικών ανρθωπιστικών κρίσεων μας προσφέρει αξιολογές παρατηρήσεις για το είδος των πληροφορών που διαμοιράζονται καθώς και σε ποιό βαθμό είναι χρήσιμα και σχετικά με το συμβάν τα μηνύματα που δημοσιεύουν οι χρήστες στις κρίσιμες ώρες μιας μεγάλης καταστροφής. [1, 28].

Οι διαμοιραζόμενες πληροφορίες που αφορούν έκτακτες καταστάσεις παρουσιάζουν κοινά χαρακτηριστικά, εμφανίζουν όμως και σημαντικές μεταβολές μεταξύ διαφορετικών ανθρωπιστικών κρίσεων. Στις εργασίες των M. Imran et al [9] και A. Otleanu et al [24] γίνεται μια προσπάθεια συλλογής συνόλων δεδομένων που αφορούν διαφορετικού τύπου καταστροφές με

σκοπό την εκπαίδευση μηχανισμών επιβλεπόμενης μάθησης. Τα μηνύματα έχουν ταξινομηθεί σε κλάσεις χρήσιμων πληροφοριών που μεταδίδονται για την κατάσταση που επικρατεί στη διάρκεια της κρίσης. Χρησιμοποιήσαμε αυτά τα σύνολα δεδομένων συνδυάζοντας και προσαρμόζοντας κατάλληλα τις επισημειωμένες κλάσεις αυτών ώστε να προκύψουν μεγάλα ενιαία σύνολα δεδομένων όπως υπαγορεύουν οι ανάγκες εκπαίδευσης των βαθέων νευρωνικών δικτύων.

2.3 Βαθιά νευρώνικά δίκτυα και μάθηση αναπαράστασης

Στην εργασία των R. Collobert et al. [4] έχουμε τη πρώτη προσπάθεια «μεταφοράς» των αλγοριθμικών ιδέων που χρησιμοποιούσαν, με ομολογουμένως μεγάλη επιτυχία, βαθιά νευρωνικά δίκτυα στην υπολογιστική όραση και συγκεκριμένα στην αναγνώριση αντικειμένων όπως το δίκτυο ImageNet, σε εφαρμογές επεξεργασίας φυσικής γλώσσας όπως η αναγνώριση μέρους του λόγου, λεκτικών μονάδων (chunking), ονομαζόμενων οντοτήτων, σημασιολογικών επισημειώσεων κ.α. Ωστόσο πολλά ζητήματα που είχαν να κάνουν με τον τρόπο αναπαράστασης της εισόδου, το πλήθος των κρυφών στρωμάτων, το τρόπο συνδυασμού χαρακτηριστικών ή τις μεθόδους βελτιστοποίησης παρέμεναν και παραμένουν ανοικτά πεδία έρευνας. Δύο εργασίες που έδωσαν σημαντική ώθηση στην αποτελεσματική εφαρμογή των συνελικτικών δικτύων στη κατηγοριοποίηση μικρο-κειμενικών δεδομένων είναι αυτές των Yoon Kim et al. και Kalchbrener et al. [13][11]. Η πρώτη χρησιμοποίησε πολλαπλά κανάλια αναπαράστασης της εισόδου άλλα και πολλαπλά φίλτρα διαφορετικού εύρους ώστε να εξάγει υψηλότερου επιπέδου συνδυασμούς πολυγραμμικών (ngrams). Η δεύτερη βασιζόταν σε μια ιεραρχική σύνθεση πολλαπλών συνελικτικών στρωμάτων και σε μία μέθοδο δυναμικού, ως προς το μήκος της εισόδου, συνδυασμού χαρακτηριστικών. Και οι δύο αρχιτεκτονικές θεωρούνται state-of-the-art και αυτές υλοποιήσαμε προσαρμοσμένες στις ανάγκες των δεδομένων μας.

Επιπλέον η χρήση κατανεμημένων αναπαραστάσεων της εισόδου μέσω διανυσμάτων ενσωματώσιμων χαρακτηριστικών είναι η πλέον καθιερωμένη σε εφαρμογές μάθησης αναπαράστασης. Ένα βασικό πλεονέκτημα των διανυσμάτων χαρακτηριστικών που προκύπτουν από προεκπαιδευμένα βαθιά νευρωνικά μοντέλα είναι πως εμφανίζουν υψηλά ποσοστά επίτυχίας για έργα ταξινόμησης διαφορετικά από εκείνα για τα οποία αρχικά εκπαιδεύτηκαν [27]. Η χρήση διανυσμάτων ενσωματώσιμων χαρακτηριστικών είναι μια μορφή μη επιβλεπόμενης διαστατικής μείωσης που συμβάλλει στη βελτίωση της γενικευτικής ικανότητας των ταξινομητών επιβλεπόμενης μάθησης. Μοντέλα word2vec [18] και glove [26] παρουσιάζουν αξιόλογα αποτελέσματα σε εφαρμογές ταξινόμησησης κειμένου και επεξεργασίας φυσικής γλώσσας ως καθολικοί περιγραφητές συντακτικών και σημασιολογικών κανονικοτήτων που εντοπίζονται σε μεγάλα σώματα μη επισημειωμένων δεδομένων [36][4].

Η εξειδίκευση αυτών των διανυσμάτων χαρακτηριστικών στο εκάστοτε ειδικό έργο ταξινόμησης μπόρει να γίνει συγχρόνως με την εκπαίδευση του νευρωνικού ταξινομητή κειμένου αν ενσωματωθεί ως στρώμμα κρυφών νευρώνων στο συνολικό δίκτυο. Στη προσπάθεια μας να αναπτύξουμε ειδικούς νευρωνικούς ταξινομητές για τα σύντομα μηνύματα κοινωνικών δικτύων

που σχετίζονται με τις διάφορες ανθρωπιστικές κρίσεις χρησιμοποιήσαμε προεκπαιδευμένα μοντέλα ενσωματώσιμων λεκτικών χαρακτηριστικών γενικού σκοπού καθώς και διανύσματα χαρακτηριστικών που εκπαιδεύτηκαν σε ένα μεγάλο σώμα ανεπισημείωτων δεδομένων κρίσεων από διάφορες παλαιότερες καταστροφές.

Κεφάλαιο 3

Θεωρητικό υπόβαθρο

3.1 Εισαγωγή

Η ταχύτατη και ραγδαία ανάπτυξη του διαδικτύου και η μετάδοση των πληροφοριών όπως σχόλια και δημοσιεύσεις σε κοινωνικά δίκτυα και ιστότοπους οδήγησε στην ανάγκη εύρεσης μεθόδων για την αυτόματη ταξινόμηση κειμένου ώστε να διαχειριστούμε και να οργανώσουμε καλύτερα και αποτελεσματικά όλα αυτά τα δεδομένα. Μέθοδοι ταξινόμησης κειμένου χρησιμοποιούνται για την κατηγοριοποίηση νέων ειδήσεων ή για την αναζήτησή και εύρεση πληροφοριών στον ιστό με τη χρήση μηχανών αναζήτησης. Επειδή η ανάπτυξη ταξινομητών κειμένου απαιτεί υψηλή εξειδίκευση και είναι χρονοβόρα και δύσκολη εν γένει διαδικασία καταφεύγουμε στην χρήση μεθόδων στατιστικής μάθησης μέσω παραδειγμάτων.

Για αρχή θα εξετάσουμε τα πλεονεκτήματα των μηχανών διανυσματικής στήριξης στο έργο της αυτόματης ταξινόμησης κειμένου. Αφού εισαγάγουμε μια συνήθη και απλή διανυσματική αναπαράσταση κειμένου και αναγνωρίσουμε τα ιδιαίτερα χαρακτηριστικά αυτής θα αναλύσουμε λεπτομερώς τον αλγόριθμο SVM και όλα εκείνα τα πλεονεκτήματα που τον καθιστούν κατάλληλο και αξιόπιστο για το έργο της κατηγοριοποίησης κειμένου . Θα δούμε ότι οι επιδόσεις του είναι τόσο κάλες και είναι μια αρκετά ευσταθής μέθοδος ώστε να χρησιμοποιείται ως βάση σύγκρισης για την αποδοτικότητα των νευρωνικών μεθόδων που θα προτείνουμε αργότερα.

Προτού προχωρήσουμε στη ανάλυση του προβλήματος της κατηγοριοποίησης κειμένου θα αναπτύξουμε κάποιους ορισμούς και θα περιγράψουμε τα διάφορα μοντέλα και αλγορίθμους που θα μας είναι απαραίτητα στη συνέχεια. Θα μας απασχολήσει εν πρώτοις το πρόβλημα της διαστατικότητας εγγενές σε όλες τις εφαρμογές ταξινόμησης κειμένου. Στη συνέχεια , εφορμώμενοι από τα πιο πρώιμα γραμμικά μοντέλα όπως οι μηχανές διανυσματικής στήριξης γραμμικού πυρήνα θα προχωρήσουμε στη χρήση μη γραμμικών μοντέλων που αξιοποιούν μεθόδους από το πεδίο των βαθέων νευρωνικών δικτύων. Θα αναζητήσουμε ακόμα αποδοτκά μοντέλα αναπαράστασης κειμένου που έχουν μεγαλύτερα περιθώρια γενίκευσης όπως το συνεχές σακίδιο λέξεων (CBOW) και η μέθοδος Skipgram.

3.2 Term weighting διανυσματική αναπαράσταση κειμένου

3.2.1 Εισαγωγή

Το πρώτο βήμα στην αυτόματη ταξινόμηση κειμένου είναι ο μετασχηματισμός από μία ακολουθία χαρακτήρων σε μία αναπαράσταση κατάλληλη να χρησιμοποιηθεί σαν είσοδος στους διάφορους αλγορίθμους μάθησης . Αυτή η αναπαράσταση έχει συνήθως τη μορφή ενός διανύσματος πραγματικών τιμών που αποδίδουν τη σημασία και το νόημα της συγκεκριμένης λέξης (διάνυσμα χαρακτηριστικών). Μία πολύ απλή και αποδοτική αναπαράσταση προκύπτει από το χώρο της ανάκτησης πληροφορίας(information retrieval) και είναι το μοντέλο στάθμισης όρων (term weighting) . Κάθε λέξη w_I αντιστοιχίζεται σε ένα χαρακτηριστικό όπως λόγου χάρη ο αριθμός των εμφανίσεων της στο εν λόγω κείμενο.

Τα συστήματα στάθμισης όρων χρησιμοποιούνταν κατά κύριο λόγο για την ανάκτηση σχετικών εγγράφων με το αίτημα ή λήμμα αναζήτησης (query) χρήστη στις δίαφορες μηχανές αναζήτησης. Τα αποτελέσματα ανάκτησης αυτών των συστημάτων είναι κατά πολύ ανώτερα άλλων πολυπλοκότερων αναπαραστάσεων κειμένου και γι' αυτό έχουν καθιερωθεί ως χαρακτηριστικά λέξεων στις διάφορες μεθόδους ταξινόμησης κειμένου. Η επιτυχία αυτών των συστημάτων σε μεγάλο βαθμό εξαρτάται από την επιλογή αποδοτικών σχημάτων στάθμισης των όρων περιεχομένου (content terms).

3.2.2 Τυπικός ορισμός

Ένα σύστημα ανάκτησης πληροφορίας στηρίζεται στην ομοιότητα των αναγνωριστικών περιεχόμενου (τιμές του διανύσματος χαρακτηριστικών) μεταξύ των στοιχείων του σώματος αποθηκευμένων εγγράφων και των λημμάτων αναζήτησης χρήστη. Στη πράξη, συγκεκριμένες λέξεις θα εξαχθούν από το σώμα των εγγράφων και των λημμάτων αναζήτησης και θα χρησιμοποιηθούν για αναγνώριση περιεχόμενου (content identification). Εναλλακτικά οι αναπαραστάσεις περιεχομένου θα αποδοθούν με επιβλεπόμενο τρόπο από ένα εξειδικευμένο προσωπικό ανά θέμα και σύμφωνα με το περιεχόμενο των αποθηκευμένων εγγράφων. Στο πλαίσιο της αυτόματης ταξίνόμησης κειμένου μπορούμε να θεωρήσουμε το σύνολο των εγγράφων εκπαίδευσης ως το σώμα των αποθηκευμένων εγγράφων και το σύνολο επαλήθευσης ως τα λήμματα αναζήτησης χρήστη.

Σε κάθε περίπτωση το έγγραφο ή το κείμενο προς ταξινόμηση αναπαρίσταται με ένα διάνυσμα όρων περιεχομένου:

$$D = (t_i, t_j, ..., t_p) \tag{3.1}$$

όπου κάθε t_k έχει αναγνωριστεί ως όρος περιεχομένου στο δείγμα κειμένου D .

Μια πιο τυπική αναπαράσταση του διάνυσματος όρων περιεχομένου περιλαμβάνει σε κάθε δίανυσμα όλους τους πιθανούς όρους περιεχομένου μόνο σταθμισμένους με μία τιμή που μετρά τη σχετική συμμετοχή τους στο έγγραφο. Αύτο το σχήμα αναπαράστασης παρέχει επιπλέον δυνατότητες διάκρισης μεταξύ των λέξεων. Για παράδειγμα, αν ο t_k συμμετέχει στο έγγραφο D

σε βαθμό w_{dk} και T το πλήθος των όρων που προσφέρονται για αναπαράσταση περιεχόμενου, το δίανυσμα όρων του εγγράφου περιγράφεται πλέον ως:

$$D = (t_1, w_1; t_2, w_2, ...; t_T, w_T) \tag{3.2}$$

Μια απλή επιλογή για τα βάρη w_{dk} είναι 1 για όρους που έχουμε αναθέσει στο έγγραφο D και 0 για τους υπόλοιπους. Με βάση τώρα την δοσμένη διανυσματική αναπαράσταση δύο εγγράφων μπορούμε να μετρήσουμε την νοηματική συνάφεια τους χρησιμοποιώντας γεωμετρικές μεθόδους. Ένα καλό μετρό είναι το εσωτερικό γινόμενο των διανυσμάτων αναπαράστασης το οποίο στην περίπτωση που τα βάρη είναι 0 και 1 είναι το πλήθος των όρων που έχουν ανατεθεί από κοινού στα δύο έγγραφα.

$$\begin{aligned} similarity(d, b) &= \mathbf{w_d} \cdot \mathbf{w_b} \\ &= \sum_{k=1}^{T} w_{dk} \cdot w_{bk} \end{aligned} \tag{3.3}$$

Στην πράξη μια άλλη επιλογή βαρών που επιτυγχάνει μεγαλύτερη διάκριση των όρων περιεχόμενου είναι τα βάρη να παίρνουν τιμές στο συνεχές διάστημα $[0, 1]$. Η επιλογή της τιμής αποδίδει την σχετική σημασία που έχει ο ορός για το συγκεκριμένο έγγραφο ώστε όσο μεγαλύτερο είναι το βάρος και πιο κοντά στη μονάδα τόσο πιο σημαντικός θεωρείται ο όρος για το συγκεκριμένο έγγραφο. Αντίθετα όσο μικρότερο είναι το βάρος και πιο κοντά στο μηδέν αφορά όρους που είναι λιγότερο σημαντικοί. Συχνά χρειάζεται να κανονικοποιήσουμε τις τιμές των βαρών προκειμένου να συμπεριλάβουμε την εξάρτηση που υπάρχει σε κάποιο βαθμό από τα βάρη των άλλων όρων στο ίδιο διάνυσμα. Κανονικοπιούμε ως προς το μήκος του διανύσματος αναπαράστασης και έχουμε:

$$\frac{w_{dk}}{\sqrt{\sum_{i \in vector} w_{di}^2}} \tag{3.4}$$

Αν χρησιμοποιήσουμε στάθμιση όρων με συνεχή βάρη και κανονικοποιημένα ως προς το μήκος διανυσματός τότε το μέτρο της νοηματικής συνάφειας δύο εγγράφων προκύπτει ότι είναι το πολύ γνωστό μέτρο ομοιότητας συνημιτόνου:

$$\frac{\sum_{k=1}^{T} w_{dk} \cdot w_{bk}}{\sqrt{\sum_{k=1}^{T} w_{dk}^2 \cdot \sum_{k=1}^{T} w_{bk}^2}} \tag{3.5}$$

3.2.3 Παράμετροι σχεδίασης

Στη σχεδίαση ενός συστήματος στάθμισης όρων περιεχομένου δύο σημαντικές αποφάσεις πρέπει να λάβουμε. Αφενός πρέπει να επιλεγούν οι κατάλληλοι όροι ώστε να διατηρείται

το νόημα του εγγράφου και αφετέρου να αποφασίσουμε πως θα αναθέσουμε βάρη ικανά να διακρίνουν ποιοί όροι είναι πιο σημαντικοί για το έγγραφο.

Όσον αφορά στο ποιούς όρους επιλέγουμε να αναθέσουμε στο έγγραφο για αναγνώριση περιεχομένου υπάρχουν πολλές δυνατότητες. Μια φυσική επιλογή που χρησιμοποιήθηκε κατά κόρον στα πρώτα συστήματά ανάκτησης πληροφορίας είναι τα unigrams δηλαδή ατομικές λέξεις που έχουν εξαχθεί από το σύνολο των αποθηκευμένων εγγράφων. Ωστόσο, η αποκλειστική χρήση ατομικών όρων δεν μπορούσε να επιτύχει ολοκληρωμένη αναγνώριση του νοήματος του κειμένου. Σε πολλές περιπτώσεις πιο πολύπλοκα σχήματα αναπαράστασης όπως διγραμμικά και τριγραμμικά (bigrams and trigrams) προσφέρονταν για βελτίωση των συστημάτων ανάκτησης πληροφορίας. Αλλες επιλογές ήταν να χρησιμοποιήσουμε τα ριζικά (stems) των λέξεων για να αποδόσουμε όρους περιεχομένου ή να μην αποδίδουμε όρους σε λέξεις πολύ μικρής συχνότητας (hapax legomena) είτε λέξεις που εμφανίζονται μεν συχνά αλλά δεν προσφέρουν νοηματικά στη πρόταση (stopwords)· τεχνικές που μείωνουν σημαντικά τη διαστατικότητα της αναπαράστασης. Εν τούτοις τα unigrams παραμένουν μια απλή και αποδοτική επιλογή.

Αναφορικά τώρα με το ποιές τιμές θα επιλέξουμε για να αναθέσουμε στα βάρη των όρων αναγνώρισης περιεχομένου, ο αντικειμενικός σκοπός μας πρέπει να είναι η βελτιστοποίηση της απόδοσης ανάκτησης σχετικών εγγράφων ή στην περίπτωση μας της ταξινόμησης των δειγμάτων του κειμένου επαλήθευσης στη πιο συναφή νοηματικά τάξη.

Πρακτικά δύο είναι τα μέτρα αυτά που αξιολογούν την απόδοση του συστήματος και είναι πολύ γνωστά ως ανάκληση (recall) και ακρίβεια (precision) αντίστοιχα. Ανάκληση είναι το ποσοστό των σχετικών εγγράφων που ανακτήθηκαν ή δείγματα κειμένου που ταξινομήθηκαν σωστά και μετράται ως ο λόγος των σχετικών ανακτηθέντων εγγράφων ή σωστά ταξινομημένων αντίστοιχα προς το σύνολο των σχετικών εγγράφων ή ταξινομημένων στη συγκεκριμένη κατηγορία. Αφετέρου η ακρίβεια είναι το ποσοστό των σχετικών ανακτηθέντων εγγράφων προς το σύνολο των ανακτηθέντων.

Η επιτυχία του συστήματος χρειάζεται εν γένει υψηλή ανάκληση για να ανακτά ότι είναι σχετικό και υψηλή ακρίβεια για να απορρίπτει ότι είναι πολύ άσχετο με την αναζήτηση. Για να έχουμε υψηλή ανάκληση επιλέγουμε να θεωρήσουμε σημαντικούς όλους εκείνους τους όρους που είναι γενικοί και εμφανίζονται συχνά στη συλλογή. Τέτοιοι όροι περιμένουμε να ανακτούν πολλά έγγραφα συμπεριλαμβανομένων και πολλών που είναι σχετικά. Την ακρίβεια απ' την άλλη εξυπηρετεί η επιλογή υψηλά εξειδικευμένων όρων ικανών να διακρίνουν τα λίγα σχετικά έγγραφα απ' την πλειοψηφία των μη σχετικών. Στην πράξη ένας συμβιβασμός των παραπάνω χρησιμοποιείται επιλέγοντας όρους νοηματικά ευρείς τόσο ώστε να έχουμε ένα ικανοποιητικό επίπεδο ανάκλησης και συγχρόνως αρκετά ειδικούς ώστε να μη δημιουργούμε αδικαιολόγητα χαμηλή ακρίβεια.

3.2.4 Το $Tf - Idf$ σύστημα στάθμισης όρων

Οι ανταγωνιστικές απαιτήσεις ακρίβειας και ανάκλησης μπορούν να ικανοποιηθούν από σύνθετα σχήματα στάθμισης όρων. Μια καθιερωμένη μέθοδος για την ανάθεση συντελεστών

	συντελεστής βάρους	περιγραφή
binary	1.0	*δυαδικά βάρη* $w_{dk} = 1$ *για όρους που βρίσκονται στο διάνυσμα του εγγράφου (ανεξαρτήτως συχτότητας όρου)*
raw frequency	tf	*αποκλειστικά συχνότητα εμφάνισης όρου (πόσες φορές εμφανίζεται ο όρος στο έγγραφο)*
augmented normalized	$0.5 + 0.5\frac{tf}{\max tf}$	*κανονικοποιημένη συχνότητα όρου tf παράγοντας κανονικοποιημένος ως προς μέγιστο tf και έπιπλέον κανονικοποιίηση ώστε να κυμένεται η τιμή μεταξύ 0.5 και 1.*

βάρους στους όρους ενός κειμένου είναι η $Tf - Idf$. Τρεις είναι οι βασικοί παράγοντες που λαμβάνονται υπόψη για την σύνθεση των συντελεστών βάρους $Tf - Idf$:

1. Όροι που εμφανίζονται συχνά στα διάφορα έγγραφα τέινουν να συμβάλλουν στην βελτίωση της ακρίβειας του συστήματος. Αυτός ο παράγοντας των συντελεστών βάρους μέτρα τη συχνότητα εμφάνισης ενός όρου στο έγγραφο και καλείται συχνότητα όρου **tf** (*term frequency*).

2. Οι παράγοντες συχνότητας όρου από μόνοι τους δεν μπορούν να εξασφαλίσουν ένα αξιόπιστο και αποδοτικό σύστημα ανακτήσης πληροφορίας. Το πρόβλημα εντοπίζεται όταν λέξεις μεγαλης συχνοτητας δε συγκεντρώνονται σε λίγα και συγκεκριμένα έγγραφα αλλά εμφανίζονται σε όλη τη συλλογή εγγράφων. Η αναζήτησή αυτών των όρων τείνει να ανακτήσει όλα τα έγγραφα κι αυτό έχει επιπτωσεις στην ακρίβεια αναζήτησης. Εισάγουμε λοιπόν ένα παράγοντα που σε επίπεδο συνόλου εγγράφων θεωρεί πιο σημαντικούς τους όρους που εντοπίζονται σε λίγα έγγραφα. Αυτός ο νέος παράγοντας είναι γνωστός ως αντίστροφη συχνότητα εγγράφου **idf** (*inverse document frequency*) .

3. Ο τρίτος παράγοντας εξαρτάται από το μήκος του εγγράφου και χρησιμοποιείται σε περιπτώσεις που η συλλογή εμφανίζει μεγάλες μεταβολές στο μέγεθος των διανυσμάτων αναπαράστασης. Κατά μείζονα λόγο τα πιο σύντομα έγγραφα τείνουν να αναπαρασταθούν από μικρά διανύσματα όρων και αντιστοίχως τα πιο εκτεταμένα έγγραφα από μεγαλύτερα διανύσματα. Τα μεγάλα σύνολα όρων έχουν περισσότερες πιθανότητες ανάκτησης και έτσι το σύστημα μεροληπτεί στην ανάκτηση μεγαλύτερων εγγράφων. Αυτο δεν είναι επιθυμητό γιατί θα θέλαμε μια αναζήτηση να ανακτά αδιακρίτως όλα τα σχετικά έγγραφα. Έτσι έχει προταθεί η χρήση ενός παράγοντα κανονικοποίησης στους συντελεστές βάρους που εξασφαλίζει ισότιμη ανάκτηση για τα έγγραφα ανεξάρτητα του μεγέθους

	συντελεστής βάρους	περιγραφή
neutral	1.0	*καμία μεταβολή στους συντελεστές βάρους χρήση των απλών συχνοτήτων όρου*
inverse collection frequency	$\log \frac{N}{n}$	*πολλαπλασιάζει τις συχνότητες όρου με τις αντίστροφες συχνότητες εγγράφου*
probabilistic inverse collection frequency	$\log \frac{N-n}{n}$	*πολλαπλασιάζει τις συχνότητες όρου με τις τυχαιοκρατικές αντίστροφες συχνότητες εγγράφου*

	συντελεστής βάρους	περιγραφή
neutral	1.0	*καμία μεταβολή στους συντελεστές βάρους χρήση των tf-idf συχνοτήτων χωρίς κανονικοποίηση*
cosine normalisation factor	$\frac{1}{\sqrt{\sum_{i \in vector} w_i^2}}$	*κανονικοποίηση συνημιτόνου : διαιρεί κάθε συντελεστή με το ευκλείδιο μήκος διανύσματος*

τους. Ο τελικός συντελεστής βάρους παίρνει τη μορφή :

$$\frac{w}{\sum_{i \in vector} w_i} \quad \text{ή} \quad \frac{w}{\sqrt{\sum_{i \in vector} w_i^2}} \tag{3.6}$$

όπου $w = tf \times idf$

3.3 Μηχανισμοί διανυσματικής στηριξης

3.3.1 Εισαγωγή

Οι μηχανισμοί διανυσματικής στήριξης βασίζονται στην αρχή της ελαχιστοποίησης του συστημικού ρίσκου όπως ορίζεται στην υπολογιστική θεωρία μάθησης. Είναι καθολικοί μηχανισμοί μάθησης και μπορούν να εκτιμήσουν γραμμικές αλλά και πολυωνυμικές συνάρτησεις χρησιμοποιώντας κατάλληλες συναρτήσεις πυρήνα. Μια εξαιρετική ιδιότητα των μηχανισμών αυτών είναι η ικανότητά τους να μαθαίνουν ανεξάρτητα από την διαστατικότητά του χώρου χαρακτηριστικών. Μετρούν την πολυπλοκότητα του χώρου υποθέσεων με βάση το περιθώριο διαχωρισμού των δεδομένων και όχι τη διαστατικότητα του χώρου χαρακτηριστικών. Με βάση το ίδιο περιθώριο διαχωρισμού μπορούμε να επιλέξουμε ένα βέλτιστο σύνολο παραμέτρων για

τον ταξινομητή χωρίς να χρειάζεται μια υπολογιστικά ακριβή διαδικασία διασταυρωτικής επαλήθευσης (cross-validation).

3.3.2 Ο αλγόριθμος SVM

Οι μηχανές διανυσματικής στήριξης [32] είναι μοντέλα επιβλεπόμενης μάθησης που βασίζονται στην ελαχιστοποίηση του εμπειρικού ρίσκου R_{emp} (3.7). Θεωρώντας δεδομένα, μία γεννήτρια τυχαίων διανυσμάτων χαρακτηριστικών ομοιόμορφα και ανεξαρτητα δειγματοληπτημένων από μια σταθερή άγνωστη συνάρτηση κατανομής πιθανότητας $P(x)$, έναν επιβλέποντα που αναθέτει τα διανύσματα εξόδου y_i για κάθε είσοδο x_i βάσει μιας δεσμευμένης συνάρτησης κατανομής πιθανότητας $P(y|x)$ και μια μηχανή μάθησης που υλοποιεί ένα σύνολο συναρτήσεων υπόθεσης $f(x,a), a \in \Lambda$ κάνουμε την επιλογή της συναρτήσεως υπόθεσης που προβλέπει την απόκριση του επιβλέποντα κατά βέλτιστο τρόπο.

$$\begin{aligned} R(\alpha) &= \int Ł\left(y, f\left(x,a\right)\right) dP(x,y) \\ P(x,y) &= P(x)P(y|x) \\ R_{emp}\left(a\right) &= \frac{1}{l}\sum_{i=1}^{l} Ł\left(y_i, f\left(x_i,a\right)\right) \\ \hat{a_0} &= arg\min_{a\in\Lambda} R_{emp}\left(a\right) \end{aligned} \tag{3.7}$$

Πρώτα μπορεί να δειχθεί πως οι λύσεις στο παραπάνω πρόβλημα ελαχιστοποίησης του ρίσκου συγκλίνουν σε βέλτιστη λύση καθώς επίσης και ότι οι τιμές του εμπειρικού ρίσκου συγκλίνουν κατα πιθανότητα στη ελάχιστη τιμή του πραγματικού ρίσκου (3.8).

$$\begin{aligned} R\left(\alpha_l\right) &\longmapsto^{P}_{l\to\infty} R\left(a_0\right) \\ R_{emp}\left(\alpha_l\right) &\longmapsto^{P}_{l\to\infty} R\left(a_0\right) \end{aligned} \tag{3.8}$$

Για την περίπτωση ενός τυπικού ταξινομητή δύο γραμμικά μη διαχωρίσιμων κλάσεων [30] μας αρκεί να προσδιοριστεί ένα βέλτιστο υπερεπίπεδο διαχωρισμού (3.9). Έστω τα διανύσματα χαρακτηριστικών x_i του συνόλου εκπαίδευσης X καθένα από τα οποία ανήκει σε μία από τις δύο κλάσεις ω_1, ω_2. Στόχος είναι να σχεδιαστεί ένα υπερεπίπεδο που ταξινομεί σωστά όλα τα διανύσματα εκπαίδευσης. Όταν περισσότερα του ενός υπερεπίπεδα διαχωρίζουν τα δεδομένα λογικό είναι να επιλέγει το υπερεπίπεδο που αφήνει το μεγαλύτερο περιθώριο (3.10) εκατέρωθεν, έτσι ώστε δεδομένα και από τις δύο κλάσεις να μπορουν να κινηθουν λίγο πιο ελεύθερα με μικρότερο ρίσκο. Ένα τέτοιο υπερεπίπεδο είναι πιο αξιόπιστο και αποδοτικό με τα άγνωστα δεδομένα (unseen data).

$$g\left(x\right) = w^T x + w_0 \tag{3.9}$$

$$z = \frac{|g(x)|}{\|w\|} \tag{3.10}$$

Κλιμακώνοντας κατάλληλα τις παραμέτρους του υπέρεπιπέδου η τιμή της $g(x)$ στα πλησιέστερα σημεία των κλάσεων είναι είτε 1 ή - 1. Πετυχαίνουμε να έχουμε περιθώριο (3.11) ενώ απαιτούμε (3.12).

$$\frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{3.11}$$

$$\begin{aligned} w^T x + w_0 &\geq 1 \quad \forall x \in \omega_1 \\ w^T x + w_0 &\leq -1 \quad \forall x \in \omega_2 \end{aligned} \tag{3.12}$$

Στόχος είναι να μεγιστοποιηθεί το περιθώριο και να διατηρηθεί το πλήθος των σημείων όσο το δυνατόν πιο μικρό . Το πρόβλημα βελτιστοποίησης τώρα είναι η επιλογή των παραμέτρων του υπερεπιπέδου έτσι ώστε να ελαχιστοποιείται ο τετραγωνικός δείκτης (3.13).

$$\begin{aligned} \Im(w, w_0, \xi) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} I(\xi_i) \\ &\min \quad \Im(w, w_0, \xi) \\ y_i\left[w^T x_i + w_0\right] &\geq 1 - \xi_i \\ \xi_i &\geq 0 \quad , i = 1, 2, \ldots, N \end{aligned} \tag{3.13}$$

Αυτό είναι ένα πρόβλημα κυρτού προγραμματισμού και επειδή οι περιορισμοί είναι γραμμικοί ορίζουν ενα κυρτό σύνολο εφικτών λύσεων. Η Lagrangian του προβλήματος δίνεται στη σχέση (3.14) για γραμμικούς ισοτικούς και ανισοτικούς περιορισμούς.

$$Ł(w, w_0, \xi, \lambda, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i\left[y_i\left(w^T x_i + w_0\right) - 1 + \xi_i\right] \tag{3.14}$$

Οι αναγκαίες συνθήκες για δεσμευμένα ακρότατα Karush-Kuhn-Tucker δίνουν τη διεύθυνση του υπρεπιπέδου σύμφωνα με τη σχέση (3.15) και τη θέση από τις συνθήκες συμπληρωματικής χαλαρότητας (3.16).

$$\frac{\partial Ł}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N}\lambda_i y_i x_i \tag{3.15}$$

$$\frac{\partial Ł}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{N}\lambda_i y_i = 0 \tag{3.16}$$

Αξίζει να σημειωθεί ότι αφού η Hessian της συνάρτησης κόστους είναι θετικά ορισμένη ο βέλτιστος ταξινομητης διανυσματικής στήριξης είναι μοναδικός καθώς η αυστηρή κυρτότητα της συνάρτησης κόστους εγγυάται ότι τα τοπικά ακρότατα είναι ολικά και μοναδικά. Η θετική σταθερά C αντισταθμίζει τους δύο ανταγωνιστικούς όρους της (3.13).

$$\frac{\partial Ł}{\partial \xi_i} = 0 \Rightarrow C - \mu_i - \lambda_i = 0 \qquad i = 1, 2, \ldots, N \tag{3.17}$$

$$\begin{aligned} \mu_i \xi_i &= 0 \qquad i = 1, 2, \ldots, N \\ \mu_i \geq 0, \quad \lambda_i &\geq 0 \qquad i = 1, 2, \ldots, N \end{aligned} \tag{3.18}$$

Υπάρχει η δυνατότητα να εκφραστεί το πρόβλημα δυϊκά ως πρόβλημα μεγιστοποίησης κατα Wolfe (3.19) απλοποιώντας αφενός τους περιορισμούς σε αμιγώς ισοτικούς και αφετέρου την υπολογιστική πολυπλοκότητα του αλγορίθμου.

$$\begin{aligned} \max \quad & Ł(w, w_0, \xi, \lambda, \mu) \\ w &= \sum_{i=1}^{N} \lambda_i y_i x_i \\ C - \mu_i - \lambda_i &= 0 \qquad i = 1, 2, \ldots, N \\ \mu_i \geq 0, \quad \lambda_i &\geq 0 \qquad i = 1, 2, \ldots, N \end{aligned} \tag{3.19}$$

Πιο σημαντικό όμως είναι το συμπέρασμα που συνάγεται από την ισοδύναμη σχέση (3.20). Αυτή η σχέση που έχει εσωτερικά γινόμενα των διανυσμάτων εκπαίδευσης τεκμέραιτε πως η συνάρτηση κόστους **δεν εξαρτάται ρητά από την διάσταση του χώρου εισόδου**. Η πολυπλοκότητα του χώρου υποθέσεων εξαρτάται από το περιθώριο διάκρισης των δεδομένων και όχι από το πλήθος των χαρακτηριστικών. Οι μηχανισμοί διανυσματικής στήριξης διατηρούν την ίδια ικανότητα γενίκευσης ανεξάρτητα από το μέγεθος του διανύσματος χαρακτηριστικών. Η ικανότητα των μηχανισμών διανυσματικής στήριξης να μαθαίνουν ανεξάρτητα από το πλήθος των χαρακτηριστικών των δεδομένων τα καθιστούν ιδανικά για εφαρμογές κατηγοριοποίησης κειμένου.

$$\begin{aligned} \max_{\lambda} & \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j \right) \\ 0 \leq \lambda_i \leq C & \qquad i = 1, 2, \ldots, N \\ & \sum_{i=1}^{N} \lambda_i y_i = 0 \end{aligned} \tag{3.20}$$

3.3.3 Κατηγοριοποίηση κειμένου και μηχανισμοί διανυσματικής στήριξης

Αξίζει να συνοψίσουμε σε αυτό το σημείο τους λόγους που καθιστούν τους μηχανισμούς διανυσματικής στήριξης κατάλληλους για το έργο της κατηγοριοποίησης κειμένου.

1. Η υψηλή διαστατικότητα του χώρου χαρακτηριστικών αποτελεί πρόκληση για κάθε ταξινόμητη κειμένου. Οι διαστατικότητες αναπαραστάσεων όπως αυτών που ορίσαμε στη προηγούμενη ενότητα είναι της τάξεως του λεξιλογίου του σώματος εκπαίδευσης (τυπικά πάνω απο 10000 χαρακτηριστικά). Επειδή οι μηχανισμοί διανυσματικής στήριξης είναι σχεδιασμένοι ώστε να αποτρέπεται η υπερπροσαρμογή (overfiting) στα δεδομένα και η απόδοσή τους δεν εξαρτάται ρητά από τον αριθμό των χαρακτηριστικών έχουν τη δυνατότητα να αντιμετωπίζουν χωρίς προβλήματα αυτούς τους πολυδιάστατους χώρους χαρακτηριστικών.

2. Τα διανύσματα χαρακτηριστικών είναι σε μεγάλο βαθμό αραιά αφού για κάθε έγγραφο το αντίστοιχο διάνυσμα χαρακτηριστικών περιέχει λίγα στοιχεία μη μηδενικά. Στην εργασία των Kivinen et al [14] εξηγείται θεωρητικά και πρακτικά γιατί οι «προσθετικοί» (additive)[1] αλγόριθμοι μάθησης που έχουν παρόμοια επαγωγική μεροληψία (inductive bias) με τους μηχανισμούς διανυσματικής στήριξης είναι κατάλληλοι για προβλήματα με σύνθετα δεδομένα [2](dense concepts) και αραιές αναπαραστάσεις.

3. Τα χαρακτηριστικά που είναι ασυσχέτιστα είναι λίγα. Ένας τρόπος άλλωστε να μειώσουμε την διαστατικότητα του χώρου χαρακτηριστικών είναι να υποθέσουμε ότι τα περισσότερα χαρακτηριστικά είναι ασυσχέτιστα. Χρησιμοποιώντας λοιπόν επιλογή χαρακτηριστικών (feature selection) αποφασίζουμε ποιά είναι τα ασυσχέτιστα χαρακτηριστικά και με αυτά επιλέγουμε να αναπαραστήσουμε τα δεδομένα. Ωστόσο, στην κατηγοριοποίηση κειμένου μόνο λίγα χαρακτηριστικά είναι ασυσχέτιστα. Το να θεωρήσουμε ότι όλα τα υπόλοιπα χαρακτηριστικά, που είναι μια συντριπτική πλειοψηφία, είναι εντελώς περιττά φαίνεται τελείως απίθανο και μας οδηγεί στην υπόθεση ότι ένας καλός ταξινομητής πρέπει να συνδυάζει όλα αυτά τα χαρακτηριστικά (dense concepts) . Αντίθετα μια εκτεταμένη περιστολή χαρακτηριστικών (aggressive feature selection) θα οδηγήσει σε απώλεια πληροφορίας.

4. Τα περισσότερα προβλήματα ταξινόμησης κειμένου είναι γραμμικά διαχωρίσιμα και οι μηχανισμοί διανυσματικής στήριξης είναι στη βασική τους μορφή γραμμικοί ταξινομητές. Και πάλι, εμφανίζουν καλές επιδόσεις για τις περισσότερες επιλογές παραμέτρων και συναρτήσεων πυρήνα και αποφεύγουν την υπερπροσαρμογή ακόμα και σε πολύπλοκους χώρους υποθέσεων (όπως συναρτήσεις πυρήνα πολυωνυμικές πέμπτου βαθμού)

[1] Ένας γραμμικός αλγόριθμος πρόβλεψης λέγεται προσθετικός όταν η ενημέρωση των βαρών είναι προσθετική ως προς το διάνυσμα βαρών δηλαδή κάθε διάνυσμα συντελεστών βάρους προκύπτει ως άθροισμα ενός σταθερού αρχικού διανύσματος βαρών και κάποιου γραμμικού συνδυασμού από στιγμιότυπα που έχουμε ήδη δει $w_t = w_1 + \sum_{j=1}^{t-1} \alpha_{t,j} x_j$

[2] χρειάζονται πολλά χαρακτηριστικά για να αναπαρασταθούν πλήρως

ακόμα και όταν χρησιμοποιούν το πλήρες διάνυσμα χαρακτηριστικών (με τυπική διάσταση 10000).

Το συμπέρασμα από όλα τα παραπάνω είναι ότι οι μηχανισμοί διανυσματικής στήριξης είναι μια φυσική λύση για το πρόβλημα της ταξινόμησης κειμένου. Οι επιδόσεις τους με έλαχιστη ρύθμιση παραμέτρων τα καθιστά ανώτερα από άλλους παραδοσιακούς αλγορίθμους ταξινόμησης. Στη συνέχεια θα εξετάσουμε το πρόβλημα της ταξινόμησης κειμένου υπό το πρίσμα των νέων μεθόθων νευρωνικής μάθησης και οι μηχανισμοί διανυσματικής στήριξης θα αποτελέσουν ένα καλό μέτρο σύγκρισης.

3.4 Νευρωνικά Γλωσσικά Μοντέλα

3.4.1 Εισαγωγή

Μοντέλα προβλέψης συμφραζόμενων ή μοντέλα ενσωματωμένων χαρακτηριστικών ή νευρωνικά γλωσσικά μοντέλα είναι η καθιερωμένη ονοματολογία των όψιμων μεθόδων κατανεμημένης σημασιολογικής μοντελοποίησης των λέξεων ενός κειμένου. Στην υπολογιστική γλωσσολογία κατά μακρά παράδοση γινόταν χρήση της πληροφορίας συμφραζόμενων ως μια καλή προσέγγιση της σημασίας των λέξεων αφού σημασιολογικά παρόμοιες λέξεις τείνουν να έχουν παρόμοιες κατανομές συμφραζόμενων. Αρχικά το πρόβλημα προσεγγιζόταν με τη χρήση διανυσμάτων που διατηρούσαν την πληροφορία συμφραζομένων (π.χ συν-εμφανιζόμενων όρων) για τις λέξεις στόχους και εφαρμόζαν γεωμετρικές μεθόδους για την μέτρηση της σημασιολογικής εγγύτητας των αντίστοιχων λέξεων.

Αυτά τα πρώιμα μοντέλα εμφάνιζαν σημαντικούς περιορισμούς αφού απαιτούσαν την εφαρμογή τεχνικών αναπροσαρμογής βαρών για τις μετρήσεις συμφραζόμενων ή μετασχηματισμών μειώσης διαστατικότητας για να δουλέψουν σωστά. Πρόσφατα μονάχα αναπτύχθηκαν μοντέλα που λύνουν το παραπάνω πρόβλημα διανυσματικής αναπαράστασης βάσει συμφραζομένων χρησιμοποιώντας ένα απλό βήμα μη επιβλεπόμενης μάθησης στη θέση των πολύπλοκων ευριστικών μετασχηματισμών των κλασικών απαριθμητικών μοντέλων.

Τα νέα προβλεπτικά μοντέλα αντί να συλλέγουν πρώτα τα διανύσματα συμφραζόμενων και ύστερα να αναπροσαρμόζουν τα βάρη αυτών θέτουν εξαρχής στα διανύσματα αναπαράστασης τα βάρη εκείνα που κατά βέλτιστο τρόπο προβλέπουν τα συμφραζόμενα των αντίστοιχων λέξεων. Αφού παρόμοιες λέξεις εμφανίζονται σε παρόμοια συμφραζόμενα το σύστημα μαθαίνει να αναθέτει παρόμοια διανύσματα αναπαράστασης σε παρόμοιες λέξεις. Παράλληλα η μάθηση είναι μη επιβλεπόμενη χωρίς το επιπλέον κόστος ανθρωπογενούς επισημειώσης αφού τα παράθυρα συμφραζόμενων που χρησιμοποιούνται για την εκπαίδευση μπορούν αυτόματα να εξαχθούν από ένα ανεπισημείωτο σώμα εγγράφων.

3.4.2 Τυπικός ορισμός

Ένα στατιστικό γλωσσικό μοντέλο μπορεί να αναπαρασταθεί από την κατά συνθήκη πιθανότητα επόμενης λέξης δοθεισών όλων των προηγούμενων λέξεων:

$$\hat{P}\left(w_1^T\right) \quad = \quad \prod_{t=1}^{T} \hat{P}\left(w_t | w_1^{t-1}\right) \tag{3.21}$$

όπου w_t είναι η t-οστή λέξη και γράφοντας w_j^i εννοούμε την υπακολουθία $w_j^i = (w_i, w_{i+1}, \ldots, w_j)$. Τέτοια στατιστικά γλωσσικά μοντέλα έχουν χρησιμοποιηθεί αρκετά σε διάφορες εφαρμογές φυσικής γλώσσας όπως αναγνώριση φωνής, μηχανική μετάφραση και ανάκτηση πληροφορίας.

Όταν κάνεις κατασκευάζει ένα στατιστικό γλωσσικό μοντέλο για να μπορέσει να μειώσει σημαντικά την δυσκολία μοντελοποίησης του προβλήματος εκμεταλλεύεται τη σειρά των λέξεων και το γεγονός ότι λέξεις που είναι γειτονικές στην ακολουθία λέξεων είναι στατιστικά πιο εξαρτημένες. Έτσι τα πολυγραμμικά (n-gram) μοντέλα κατασκευάζουν πίνακες κατά συνθήκη πιθανοτήτων εμφάνισης της επόμενης λέξης για κάθε ένα από ένα μεγάλο πλήθος συμφραζόμενων δηλαδή συνδυασμών των τελευταίων $n-1$ λέξεων:

$$\hat{P}\left(w_t | w_1^{t-1}\right) \quad \approx \quad \hat{P}\left(w_t | w_{t-n+1}^{t-1}\right) \tag{3.22}$$

Λαμβάνουμε υπόψη μας μόνο τους συνδυασμούς διαδοχικών λέξεων που εμφανίζονται στο σώμα εκπαίδευσης ή που τουλάχιστον εμφανίζονται συχνά. Όσον αφορά τώρα συνδυασμούς λέξεων που δεν εμφανίζονται στο σώμα εκπαίδευσης τους αναθέτουμε την πιθανότητα που θα προβλέπαμε χρησιμοποιώντας μικρότερο μέγεθος συμφραζόμενων δηλαδή η νέα ακολουθία είναι σα να παράγεται από την συνένωση των μικρότερων και επικαλυπτόμενων υπακολουθίων μήκους $1, .., n$ τις οποίες έχουμε συχνά συνάντησει στο σώμα εκπαίδευσης.

3.4.3 Συνοπτική περιγραφή της μεθόδου

1. Συσχετισμός κάθε λέξης του λεξιλογίου με ένα κατανεμημένο διάνυσμα χαρακτηριστικών λέξεως (ένα πραγματικότιμο διάνυσμα του $\mathbb{R}^m$).

2. Έκφραση της από κοινού κατανομής πιθανότητας της ακολουθίας λέξεων σε όρους διανυσμάτων χαρακτηριστικών.

3. Ταυτόχρονη μάθηση των διανυσμάτων χαρακτηριστικών λέξεως και των παραμέτρων της παραπάνω κατανομής πιθανότητας.

Το διάνυσμα χαρακτηριστικών λέξεως αναπαριστά την κάθε λέξη σαν σημείο ενός διανυσματικού χώρου. Ο αριθμός των χαρακτηριστικών είναι σημαντικά μικρότερος του μεγέθους του λεξιλογίου ($m = 300$ στα πειράματα). Η κατανομή πιθανότητας εκφράζεται ως γινόμενο των κατά συνθήκη πιθανοτήτων εμφάνισης της επόμενης λέξης δοθεισών των προηγούμενων αυτής . Αυτή η συνάρτηση έχει παραμέτρους που επαναληπτικά προσαρμόζονται ώστε να μεγιστοποιούν την λογαριθμική πιθανοφάνεια των δεδομένων εκπαίδευσης και ενδεχομένως κάποιο κριτήριο κανονικοποίησης.

3.4.4 Το Εμπροσθόδρομο (feedforward) Νευρωνικό Γλωσσικό Μοντέλο (ΕΝΓΜ)

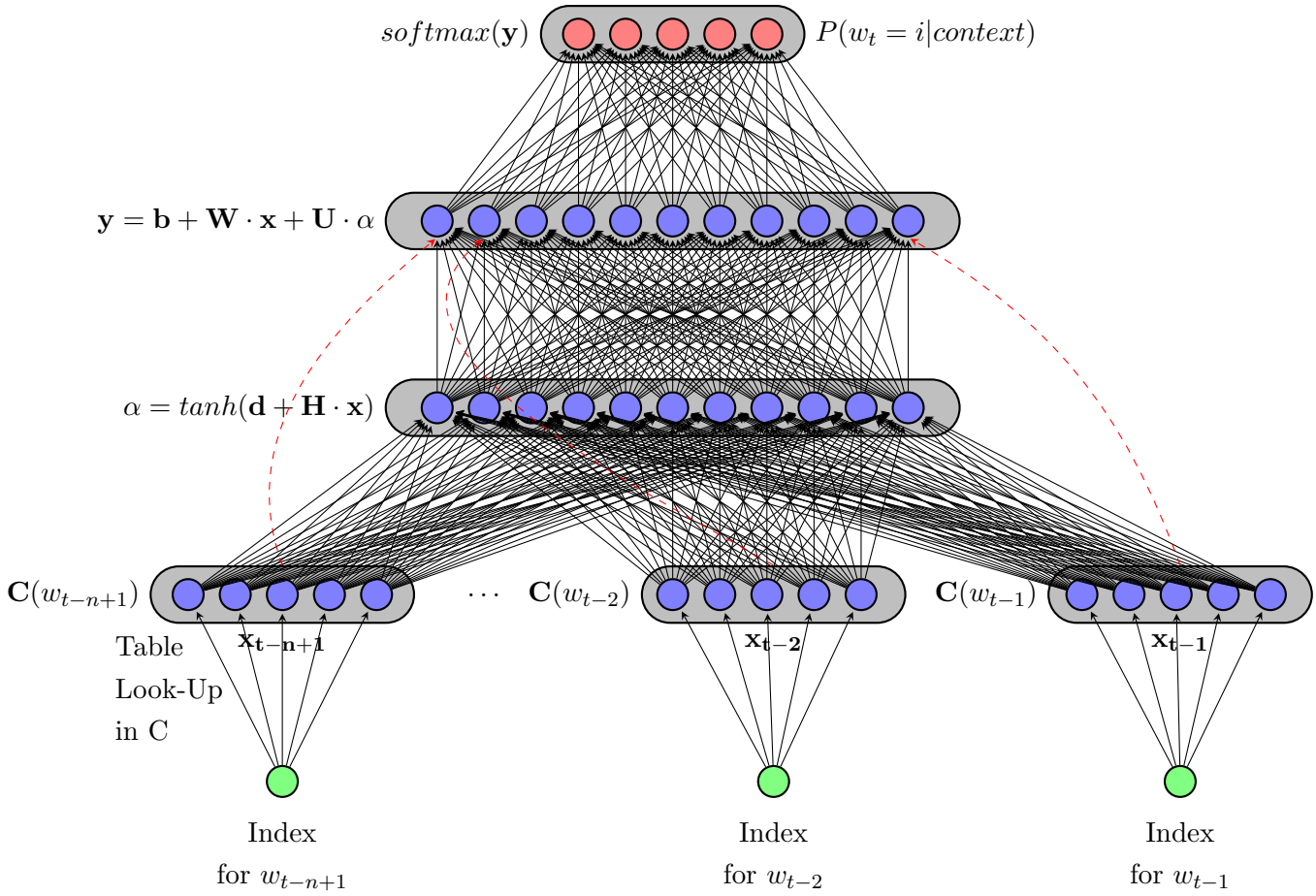
Το 2003 οι Bengio et al. [3] θα προτείνουν μια πυκνή λεξιλογιακή αναπαράσταση που βασίζεται στην εκπαίδευση ενός νευρωνικού πιθανοκρατικού μοντέλου για την εκμάθηση κατανεμημένων λεξιλογιακών χαρακτηριστικών. Ας εξετάσουμε εν γένει ένα νευρωνικό γλωσσικό μοντέλο. Αν θεωρήσουμε την ακολουθία $w_1, ..., w_T$ των λέξεων $w_t \in V$, όπου V το σύνολο λεξιλογίου το οποίο είναι μεγάλο αλλά πεπερασμένο, σκοπός είναι η μάθηση ενός καλού μοντέλου $f(w_t, ..., w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1})$ που θα δίνει υψηλή εκτός δείγματος πιθανοφάνεια. Ο μόνος περιορισμός για το μοντέλο είναι για κάθε επιλογή συμφραζομένων w_1^{t-1}το άθροισμα ως κατανομή πιθανότητας να είναι μονάδα $\sum_{i=1}^{|V|} f(i, w_t, ..., w_{t-n+1}) = 1$. Στη θέση λοιπόν των αραιών πολυδιάστατων αναπαραστάσεων, ο Bengio χρησιμοποιεί ένα χαμηλότερης διάστασης διανυσματικό πεδίο εκπαιδεύοντας ένα νευρώνικό δίκτυο που προβλέπει την επόμενη λέξη w_t για τα συμφραζόμενα w_1^{t-1}. Το δίκτυο παρουσιάζεται στο σχήμα 3.1 και μπορούμε να το διαιρέσουμε δύο τμήματα.

Αποσυνθετουμε την συνάρτηση $f(w_t, ..., w_{t-n+1}) = \hat{P}(w_t|w_1^{t-1})$ σε δύο μέρη:

1. Μια απεικόνιση C από κάθε στοιχείο i του V σε ένα πραγματικότιμο διάνυσμα $C(i) \in \mathbb{R}^m$. Αντιπροσωπεύει ένα κατανεμημένο διάνυσμα χαρακτηριστικών που σχετίζεται με κάθε λέξη στο λεξιλόγιο. Στην πράξη η C αναπαρίσταται σαν $|V| \times m$ πίνακας παραμέτρων προσδιοριστέων.

2. Η κατανομή πιθανότητας της ακολουθίας των λέξεων εκφρασμένη σε όρους του C: μια συνάρτηση g απεικονίζει την ακολουθία εισόδου των διανύσματων χαρακτηριστικών λέξεως για λέξεις εντός συμφραζόμενων, $(C(w_{t-n+1}, ..., C(w_{t-1}))$, σε μια κατά συνθήκη κατανομή πιθανότητας των λέξεων του V για την εκάστοτε επόμενη λέξη w_t. Η έξοδος της g είναι ένα διάνυσμα του οποίου το $i-$στο στοιχείο εκτιμά την πιθανότητα $\hat{P}\left(w_t = i|w_1^{t-1}\right)$ όπως δείχνει και το σχήμα 3.1.

$$f\left(i, w_{t-1}, \cdots, w_{t-n+1}\right) \quad = \quad g\left(i, C\left(w_{t-1}\right), \cdots, C\left(w_{t-n+1}\right)\right) \tag{3.23}$$



Σχήμα 3.1: Αρχιτεκτονική νευρωνικού δικτύου $g(i, C(w_{t-n+1}), \ldots, C(w_{t-2}), C(w_{t-1}))$ όπου g εμπροσθόδρομο νευρωνικό δίκτυο και $C(i)$ το $i-$στο διάνυσμα χαρακτηριστικών.

Η συνάρτηση f προκύπτει τώρα ως σύνθεση των απεικονίσεων (g και C) με το C να διαμοιράζεται μεταξύ όλων των λέξεων στα συμφραζόμενα. Κάθε μέρος έρχεται μαζί με ένα σύνολο παραμέτρων προσδιοριστέων. Οι παράμετροι της απεικονίσεως C είναι τα ίδια τα διανύσματα χαρακτηριστικών που αναπαρίστανται ως $|V| \times m$ πίνακες με $i-$στη γραμμή το διάνυσμα χαρακτηριστικών για την $i-$στη λέξη. Η συνάρτηση g μπορεί να υλοποιηθεί από ένα εμπροσθόδρόμομο νευρωνικό δίκτυο (ή ένα αναδρομικό νευρωνικό δίκτυο) με παραμέτρους ω. Το πλήρες σύνολο παράμετρων είναι το $\theta = (C, \omega)$.

Η εκπαίδευση επιτυγχάνεται αναζητώντας τις παραμέτρους θ που μεγιστοποιούν την λογαριθμική πιθανοφάνεια (3.24) όπου R είναι όρος κανονικοποίησης. Για παράδειγμα το R είναι weight decay ποινή που εφαρμόζεται στα βάρη του νευρωνικού δικτύου και του πίνακα C.

$$L \quad = \quad \frac{1}{T} \sum_t \log f\left(w_t, w_{t-1}, ..., w_{t-n+1}; \theta\right) \quad + \quad R\left(\theta\right) \tag{3.24}$$

Το δίκτυο που εκπαιδεύουμε έχει τρία κρυφά επίπεδα ένα για την κοινή αναπαράσταση των λέξεων σε διανύσματα χαρακτηριστικών, ένα κρυφό επίπεδο υπερβολικής εφαπτομένης για την εξαγωγή μη γραμμικών χαρακτηριστικών και ένα ακόμη κρυφό επίπεδο πλήρως γραμμικό. Τέλος υπολογίζει την συνάρτηση κατανομής με μία $softmax$ (3.25) σε επίπεδο εξόδου που εγγυάται θετικές πιθανότητες που αθροίζονται στη μονάδα.

$$\hat{P}\left(w_t | w_{t-1}, ..., w_{t-n+1}\right) \quad = \quad \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{3.25}$$

Η εκπαίδευση γίνεται με στοχαστική κάθοδο στην διεύθυνση της πιο απότομης βαθμίδας εκτελώντας την εξής επανάληψη ανανέωσης για κάθε νέα λέξη του σώματος εκπαίδευσης και ρυθμό μάθησης ε:

$$\theta \quad \leftarrow \quad \theta \quad + \quad \varepsilon \cdot \frac{\partial \log \hat{P}\left(w_t | w_{t-1}, ..., w_{t-n+1}\right)}{\partial \theta} \tag{3.26}$$

3.4.5 Ανάλυση κόστους για το εμπροσθόδρομο νευρωνικό γλωσσικό μοντέλο

Ένα θέμα που εγείρει η παραπάνω υλοποίηση είναι αυτό της υπολογιστικής πολυπλοκότητας. Εξετάζοντας προσεκτικά το μοντέλο διακρίνουμε τρία βασικά στάδια. Το πρώτο στάδιο είναι η απεικόνιση στο στρώμα εισόδου από τους δείκτες του λεξιλογίου σε μια κοινή κατανεμημένη αναπαράσταση. Για κάθε λέξη εξετάζουμε τις N προηγούμενες και έτσι κάθε στιγμή μόνο N κόμβοι του δικτύου στην είσοδο είναι ενεργοί. Κάθε δείκτης μέσω ενός διαμοιραζόμενου πίνακα απεικονίσης χρησιμοποιείται για τον υπολογισμό των κόμβων του πρώτου κρυφού στρώματος με διάσταση $N \times V$ με το V να κυμαίνεται από 500 ως 2000. Στη συνέχεια κάθε διάνυσμα κατανεμημένων χαρακτηριστικών τροφοδοτείται ως είσοδος στο μη γραμμικό κρυφό επίπεδο που εξάγει καλύτερα χαρακτηριστικά (high level features) με τη βοήθεια V κρυφών μονάδων. Αυτό το στάδιο είναι ιδιαίτερα απαιτητικό αφού μια τυπική τιμή των κρυφών μονάδων

είναι 500 με 1000 και έχει κόστος $N \times V \times H$. Το τελευταίο στάδιο είναι ο υπολογισμός της συνάρτησης κατανομής για κάθε λέξη στο λεξιλόγιο και έχει κόστος $V \times H$. Λαμβάνοντας υπόψη την παραπάνω ανάλυσή το συνολικό κόστος αποτιμάται σε:

$$C_{total} = N \times V + N \times V \times H + H \times V \tag{3.27}$$

Ο όρος που επηρεάζει κατά κύριο λόγο την πολυπλοκότητα είναι ο $V \times H$.Ένας τρόπος να υπολογιστεί πιο αποδοτικά είναι η αποθήκευση του λέξιλογίου σε μορφή δυαδικού δέντρου οπότε ο αριθμός των κρυφών μονάδων που πρέπει να υπολογιστούν γίνεται $N \times \log_2(V)$. Πρακτικά έχουν προταθεί κι άλλες μέθοδοι μείωσης του παραπάνω κόστους όπως η χρήση ιεραρχικού softmax για τον υπολογισμό των μονάδων εξόδου ή η αποφυγή κανονικοποίησης του μοντέλου στο στάδιο της εκπαίδευσης.

Στα πιο αποδοτικά μοντέλα που χρησιμοποιούμε παρακάτω θα χρησιμοποιήσουμε ιεραρχικό softmax και αναπαράσταση του λεξιλογίου με δυαδικά Huffman δέντρα. Έτσι οι πιο συχνοί όροι αναπαρίστανται με ένα σύντομο δυαδικό κωδικό μειώνοντας την πολυπλοκότητα στην έξοδο σε $H \times \log_2(Unigram\_perplexity(V))$. Σε ένα λεξιλόγιο της τάξεως του εκατομμυρίου αυτό προσφέρει έως και διπλάσια επιτάχυνση του χρόνου εκπαίδευσης. Παρότι μια τέτοια μείωση κόστους δεν είναι και τόσο σημαντική συνολικά όσο η υπολογιστική συμφόρηση που προκαλεί ο όρος $N \times V \times H$ θα δούμε ότι με τη χρήση αρχιτεκτονικών χωρίς μη γραμμικά κρυφά επίπεδα η συνολική αποδοτικότητα εξαρτάται σημαντικά από το κόστους αυτού του softmax επιπέδου.

3.5 Γραμμολογαριθμικά (Log-Linear) Νευρωνικά Γλωσσικά Μοντέλα

Δύο νέες αρχιτεκτονικές νευρωνικών που είναι πιο αποδοτικές στην μάθηση διανυσμάτων κατανεμημένων χαρακτηριστικών λέξεων είναι η CBOW και η Skipgram. Με βάση την παρατήρηση από την ανάλυση κόστους για την ανάγκη βελτίωσης του όρου $N \times V \times H$ προτείνονται δύο εναλλακτικές λύσεις. Και οι δύο αφαιρούν το μη γραμμικό κρυφό επίπεδο που είναι ακριβό. Παρόλο που ο υπολογισμός αυτών των μη γραμμικών μετασχηματισμών είναι το βασικό πλεονέκτημα χρήσης των νευρωνικών δικτύων θα δούμε ότι τα απλούστερα αυτά μοντέλα αποδίδουν καλύτερα όταν εκπαιδευτούν πάνω σε μεγαλύτερα σύνολα δεδομένων.

Παρακάτω θα δούμε ότι η εκπαίδευση των διανυσμάτων κατανεμημένων χαρακτηριστικών λέξεων μπορεί να γίνει σε δύο φάσεις. Σε μία πρώτη φάση χρησιμοποιούνται τα απλοποιημένα (γραμμικά) νευρωνικά γλωσσικά μοντέλα για την αρχικοποίηση των διανυσμάτων χαρακτηριστικών λέξεων και σε δεύτερη φάση χρησιμοποιούν τα πολυγραμμικά (n-gram) ΕΝΓΜ , όπως αυτά που περιγράψαμε στη προηγούμενη ενότητα , για την ενημέρωση των προεκπαιδευμένων διανυσμάτων χαρακτηριστικών στα δεδομένα της εκάστοτε εφαρμογής.

3.5.1 Συνεχές Σακίδιο Λέξεων (CBOW)

Η πρώτη αρχιτεκτονική καλείται συνεχές σακίδιο λέξεων (Continuous Bag of Words) και χάριν συντομίας (CBOW). Βασίζεται στις εξής δύο απλοποιητικές παρεμβάσεις: την αφαίρεση του μη γραμμικού κρυφού επιπέδου και τον διαμοιρασμό του στρώματος απεικόνισης. Αξίζει να σημειωθεί ότι δεν μιλάμε απλώς για ένα κοινό πίνακα απεικονίσης αλλά για κοινό στρώμα απεικονίσης εισόδου όπου κοινές λέξεις προβάλλονται σε ακριβώς ίδια σημεία του χώρου αναπαράστασης. Καλείται αφενός σακίδιο λέξεων εφόσον τα συμφραζόμενα κάθε λέξης δεν επηρεάζουν την προβολή στο χώρο χαρακτηριστικών και αφετέρου συνεχές αφού χρησιμοποιούν συνεχή κατανεμημένα διανύσματα χαρακτηριστικών. Κάνοντας αυτές τις παραδοχές μπορούμε να κατασκευάσουμε έναν γραμμολογαριθμικό ταξινομητή που με είσοδο τις λέξεις που έπονται και προηγούνται του τρέχοντος όρου προσπαθούν να προβλέψουν την ενδιάμεση λέξη. Η υπολογιστική πολυπλοκότητα εκπαίδευσης αυτού του μοντέλου είναι:

$$C_{total} = N \times V + H \times \log_2(V) \tag{3.28}$$

Σημειώνεται επίσης ότι ο πίνακας απεικονίσης μεταξύ εισόδου και στρώματος προβολής είναι κοίνος για κάθε πιθανό δείκτη λέξεως με τον ίδιο τρόπο όπως και στην περίπτωση του ΕΝΓΜ.

3.5.2 Το Μοντέλο Skip-gram

Η δεύτερη αρχιτεκτονική είναι παρόμοια με το συνεχές σακίδιο λέξεων μόνο που προβλέπει τα συμφραζόμενα γύρω από μία κεντρική λέξη. Για κάθε όρο που παίρνει ως είσοδο ο γραμμολογαριθμικός ταξινομητής προβλέπει τις λέξεις που έπονται και προηγούνται αυτού εντός ενός συγκεκριμένου εύρους. Η υπολογιστική πολυπλοκότητα εκπαίδευσης αυτού του μοντέλου είναι:

$$C_{total} = R \times (D + D \times \log_2(V)) \tag{3.29}$$

όπου R η μέγιστη τιμή του εύρους συμφραζομένων. Είναι σημαντικό να παρατηρήσουμε ότι λέξεις σε μεγαλύτερη απόσταση σχετίζονται λιγότερο με την λέξη εισόδου και γι' αυτό τους αποδίδονται μειωμένα βάρη υποδειγματοληπτώντας αυτούς τους όρους στα παραδείγματα του σύνολου εκπαίδευσης.

Για παράδειγμα, αν επιλέξουμε $R = 5$ τότε για κάθε λέξη μπορούμε να επιλέξουμε αυθαίρετα μια τιμή εύρους $r \in \{1, \ldots, 5\}$ και να χρησιμοποιήσουμε r λέξεις αριστερά και δεξιά του όρου που εξετάζουμε ως σωστές επισημειώσεις. Αυτό απαιτεί τη ταξινόμηση $r \times 2$ λέξεων με τον όρο που εξετάζουμε κάθε φορά ως είσοδο και κάθε μία από τις $r + r$ λέξεις ως έξοδο.

Πιο τυπικά αντικειμενικός σκοπός του μοντέλου Skip-gram είναι η εύρεση των κατάλληλων αναπαραστάσεων για να προβλεφθούν σωστά οι γειτονικές λέξεις (συμφραζόμενα) του όρου που μας ενδιαφέρει. Αν μας δοθεί μια ακολουθία λέξεων εκπαίδευσης $w_1, w_2, \ldots, w_T$ το μοντέλο θα εκπαιδευτεί μεγιστοποιώντας τη μέση λογαριθμική πιθανοφάνεια

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-r\leq j\leq r, j\neq 0} \log \quad p\left(w_{t+j}|w_t\right) \tag{3.30}$$

όπου r πλήθος συμφραζομένων για κάθε λέξη w_t. Η αύξηση του εύρους συμφραζομένων αυξάνει το πλήθος των παραδειγμάτων εκπαίδευσης και συνεπώς βελτιώνει την ακρίβεια του ταξινομητή αλλά και απαιτεί μεγαλύτερο υπολογιστικό κόστος. Η δε κατά συνθήκη πιθανότητα $p\left(w_{t+j}|w_t\right)$ υπολογίζεται από μια συνάρτηση $softmax$.

3.6 Ταξινόμηση κειμένου με συνελικτικά νευρωνικά δίκτυα

3.6.1 Εισαγωγή

Οι Collobert et al. [4] προτείνουν μία ενιαία αρχιτεκτονική νευρωνικων δίκτυων και έναν αλγόριθμο μάθησης που μπορεί να εφαρμοστεί σε διάφορες εφαρμογές επεξεργασίας φυσικής γλώσσας όπως part-of-speech tagging, chunking, named-entity-recognition, semantic role labeling. Πολλές από τις αρχές που θέτει η εργασία τους θα υιοθέτησουμε για το έργο της κατηγοριοποίησης κειμένου (text classification) δίνοντας το βασικό σκελετό για τις προτεινόμενες νευρωνικές αρχιτεκτονικές. Είναι σημαντικό πως στη προσέγγιση αυτή το μοντέλο παραμένει αρκετά ευέλικτο αποφεύγοντας εν γένει μεθόδους εξαγωγής χαρακτηριστικών που είναι εξειδικευμένες στην εκάστοτε εφαρμογή (task-specific) και έτσι δε απαιτεί γνώση του γλωσσικού μοντέλου a priori ούτε κάθε φορά απαιτείται νέα σχεδίαση αντιπροσωπευτικών χαρακτηριστικών. Συνεπώς απαλασσόμαστε από το δύσκολο έργο του να εξάγουμε εμείς τα χαρακτηριστικά με επιβλεπόμενο τρόπο και αφήνουμε το σύστημα να μάθει, βάσει ενός μεγάλου συνόλου ανεπισημείωτων εγγράφων εκπαίδευσης, την αναπαράσταση των χαρακτηριστικών λέξεως (μετέπειτα τα χαρακτηριστικά αύτα προσαρμόζονται εύκολα στο εκάστοτε σώμα δεδομένων εκπαίδευσης). Παρακάτω θα περιγράψουμε σε αρκετά σημεία καινοτομίες που εισάγει η εργασία των Colobert et al. καθώς αποτελεί τον πρώτο νευρωνικό end-to-end ταξινομητή κειμένου με δυνατότητα γενίκευσης σε διαφορετικά έργα.

3.6.2 Το πρόβλημα της επεξεργασίας φυσικής γλώσσας

Σκοπός των μεθόδων επεξεργασίας φυσικής γλώσσας είναι η μετατροπή ενός κειμένου από το φυσικό λόγο σε μια δομή δεδομένων φιλική στον προγραμματιστή που θα περιγράφει το νόημα του κειμένου αυτού. Μέχρι να βρεθεί μια κοινά αποδεκτή λύση για το στοιχειώδες αυτό πρόβλημα τεχνητής νοημοσύνης το αντιμετωπίζουμε μερικώς με την εξαγωγή απλούστερων αναπαραστάσεων που περιγράφουν κάποιες μόνο πτυχές της κειμενικής πληροφορίας.

Τέτοιες απλούστερες περιγραφές χρησιμοποιούνται είτε για εξειδικευμένες εφαρμογές (διάφορες εκδοχές του bag-of-words) όπως εφαρμογές άνλτησης πληροφοριών (information retrieval) είτε για την εξαγωγή γενικών γλωσσολογικών αναπαραστάσεων όπως αυτές που περιγράφουν πληροφορίες σύνταξης ή σημασιολογίας. Συνήθως η πρακτική που ακολουθούσαν στις εφαρμογές επεξεργασίας φυσικής γλώσσας ήταν με βάση σύνολα δεδομένων που

είχαν επισημειωθεί ανθρωπογενώς να συγκρίνουν τις αποδόσεις των διαφόρων αλγορίθμων ταξινόμησης. Τέτοια συστήματα ενεφάνιζαν υψηλές επιδόσεις ωστόσο για τα εξειδικευμένα benchmarks για τα οποία είχαν εκπαιδευτεί.

Συγκεκριμένα η πληθώρα αυτών των μοντέρνων και εξειδικευμένων συστημάτων αντιμετώπιζε ένα ορισμένο έργο ταξινόμησης εφαρμόζοντας γραμμικά στατιστικά μοντέλα σε ad-hoc χαρακτηριστικά. Τα χαρακτηριστικά αυτά προέκυπταν από την έξοδο προϋπαρχόντων συστημάτων και ήταν μία προσέγγιση αποτελεσματική αφού αξιοποιούσε ένα μεγάλο σώμα προϋπάρχουσας γλωσσολογικής γνώσης. Ωστόσο παρείχε μικρές δυνατότητες γενίκευσης αφού βελτιστοποιώντας την απόδοση του συστήματος για ένα εξειδικευμένο έργο πρακτικά συνεισέφερε ελάχιστα στην βελτίωση της απόδοσης του σε άλλα έργα κατηγοριοποίησης και συνεπώς στην επίλυση του γενικού προβλήματος της κατανόησης φυσικής γλώσσας (language understanding).

Κατ΄ αρχήν όλες οι μέθοδοι επεξεργασίας φυσικής γλώσσας μπορούν να αντιμετωπιστούν ως επισημειώσεις λέξεων. Παραδοσιακά το πρόβλημα προσεγγιζόταν ως εξής: εξήγαγαν από τις προτάσεις ένα πλούσιο σύνολο ανθρώπογενως σχεδιασμένων χαρακτηριστικών και αυτά τροφοδοτούνταν σε ένα κλασικό αλγορίθμο ταξινόμησης όπως ο SVM γραμμικού πυρήνα. Η επιλογή των χαρακτηριστικών είναι εντελώς εμπειρική και βασίζεται στην γλωσσολογική διαίσθηση και στη λογική δοκιμής και πλάνης κάνοντας την επιλογή χαρακτηριστικών να συνδέεται απαραίτητα με την υποκείμενη εφαρμογή και αναγκαία την επιπλέον έρευνα για κάθε νέα εφαρμογή γλωσσικής επεξεργασίας. Πολύπλοκα προβλήματα απαιτούσαν πιθανώς ένα μεγάλο πλήθος πολύπλοκων χαρακτηριστικών αυξάνοντας το υπολογιστικό κόστος σε απαγορευτικά επίπεδα στην περίπτωση εφαρμογών μεγάλης κλίμακας.

Στην προσπάθεια λοιπόν βελτιστοποίησης της απόδοσης των αλγορίθμων ταξινόμησης εν γένει και όχι εργο-ειδικά (task-specific) υπήρξε η λογική να αποφευχθεί η σχεδίαση εξειδικευμένων χαρακτηριστικών. Οι Colobert et al. [4] προτείνουν ένα ενιαίο σύστημα μάθησης ικανό να ανκαλύψει την εσωτερική αναπαράσταση αυτών των γλωσσικών χαρακτηριστικών. Χρησιμοποιώντας τα benchmarks σαν έμμεσες μετρήσεις της σχετικότητας των εσωτερικών αναπαραστάσεων που ανακάλυψε το δίκτυο (μια ορισμένη διαδικασία μάθησης) ισχυρίζονται ότι αυτές οι αναπαραστάσεις είναι πιο γενικές από οποιοδήποτε άλλη επιβλεπόμενη αναπαράσταση. Αυτή η απόφαση να αποφευχθεί η σχεδιάση χαρακτηριστικών εξειδικευμένων για την εφαρμογή αποκλείει μεν από ένα μεγάλο σώμα γλωσσολογικής γνώσης που θα είχαν κανονικά διάθεσιμο ανοίγει νέες δυνατότητες δε στην επίλυση του γενικότερου προβλήματος κατανόησης κειμένου.

Συνεπώς προτείνεται μια ριζικά διαφορετική προσέγγιση που περιορίζει την προεπεξεργασία και σχεδιάση χαρακτηριστικών στο ελάχιστο και στη συνέχεια εκπαιδεύεται ένα πολυστρωματικό νευρωνικό δίκτυο για οποιοδήποτε έργο. Μια αρχιτεκτονική που βασίζεται στις παραπάνω αρχές παίρνει ως είσοδο μία πρόταση και αφού εκπαίδευσει διάφορα στρώματα εξαγωγής χαρακτηριστικών, τροφοδοτεί αυτά σε μετέπειτα στρώματα του νευρωνικού δικτύου για το έργο της ταξινόμησης. Προσαρμόζει δε τα υπολογισμένα αυτά χαρακτηριστικά αυτόματα μέσω του αλγορίθμου back-propagation ώστε να είναι όσο το δυνατόν πιο αντιπροσωπευτικά για την εκάστοτε εφαρμογή. Περιγράφουμε παρακάτω αυτή τη γενικού σκοπού αρχιτεκτονική κα-

τάλληλη για κάθε πρόβλημα επεξεργασίας φυσικής γλώσσας και συνεπώς του προβλήματος ενδιαφέροντος μας.

Η προτεινόμενη αρχιτεκτονική συνοψίζεται ως εξής:

- Το πρώτο στρώμα εξάγει χαρακτηριστικά για κάθε ατομική λέξη.
- Το δεύτερο στρώμα εξάγει χαρακτηριστικά από ένα παράθυρο λέξεων η ολόκληρη την πρόταση αντιμετωπίζοντας την σαν ακολουθία με τοπική και ολική δομή (και όχι σαν bag-of-words)
- Ακολουθούν συνήθη στρώματα νευρωνικών δικτύων.

3.6.3 Μετασχηματίζοντας λέξεις σε διανύσματα χαρακτηριστικών

Ένα βασικό στοιχείο για την αρχιτεκτονική είναι η σχεδίαση της ώστε να παραμένει αποδοτική χρησιμοποιώντας σαν είσοδο τις αρχικές λέξεις του κειμένου ελάχιστα προεπεξεργασμένες και γι' αυτό είναι σημαντική η εκμάθηση καλών ενσωματώσιμων χαρακτηριστικών λεξιλογίου (word embeddings). Για λόγους αποδοτικότητας οι λέξεις τροφοδοτούνται στο δίκτυο ως δείκτες ενος πεπερασμένου λεξιλόγιου. Οπωσδήποτε ένας δείκτης δεν παρέχει πολύ χρήσιμη πληροφορία σχετικά με τη λέξη αυτή. Το πρώτο στρώμα του νευρωνικού δικτύου απεικονίζει κάθε δείκτη λέξης σε ένα διάνυσμα χαρακτηριστικών χρησιμοποιώντας έναν πίνακα απεικόνισης. Για κάθε εφαρμογή μια σχετική και αντιπροσωπευτική αναπαράσταση κάθε λέξης δίνεται από τον αντίστοιχο πίνακα διανύσματων χαρακτηριστικών ο όποιος εκπαιδεύεται με έναν αλγόριθμο back-propagation από μια κάποια αρχικοποίηση τιμών (τυχαία ή προεκπαδευμένη).

Για κάθε λέξη μία εσωτερική αναπαράσταση χαρακτηριστικών δίνεται από τον πίνακα απεικονίσης (3.31) (πρώτο στρώμα).

$$LT_W\left[w\right] \quad = \quad \langle W \rangle_w^1 \tag{3.31}$$

Δοθείσης μίας προτάσεως ή μιας ακολουθίας λέξεων από το λεξικό ο πίνακας απεικονίσης εφαρμόζεται ομοίως σε κάθε συνιστώσα της ακολουθίας παράγοντας το ακόλουθο μητρώο (3.32). Ο πίνακας αυτός τροφοδοτείται στα επόμενα επίπεδα του νευρωνικού δικτύου για εξαγωγή ανώτερων χαρακτηριστικών.

$$LT_W\left[[w]_1^T\right] \quad = \quad \left(\langle w \rangle_{[w]_1}^1, \langle w \rangle_{[w]_2}, \cdots, \langle w \rangle_{[w]_T}\right) \tag{3.32}$$

3.6.4 Εξάγοντας υψηλότερου επιπέδου χαρακτηριστικά από τα διανύσματα ενσωματωμένων χαρακτηριστικών

Τα διανύσματα ενσωματωμένων χαρακτηριστικών που παράγονται από τον πίνακα απεικονίσης στη συνέχεια θα πρέπει να συνδυαστούν με μετέπειτα στρώματα του νευρωνικού δικτύου για να παράξουν μία επισημείωση ανά λέξη. Η επισημείωση λέξεων σε ακολουθίες μεταβλητού μήκους είναι βασικό πρόβλημα της μηχανικής μάθησης. Υπάρχουν δύο κοινά αποδεκτές προσεγγίσεις για την λύση του προβλήματος αυτού. Η πρώτη αναφέρεται συχνά ως

τοπική προσέγγιση βάσει ενός παραθύρου γειτονικών λέξεων, όπου κάθε φορά μία νέα λέξη επισημειώνεται, και η δεύτερη είναι μία ολική προσέγγιση ανά πρόταση (συνελικτική).

Εξαγωγή τοπικών χαρακτηριστικών με ένα σταθερό παραθύρο λέξεων

Η εξαγωγή αυτών των χαρακτηριστικών αποδίδει καλά αν γνωρίζουμε ότι η ταξινόμηση κάθε λέξης εξαρτάται αποκλειστικά από γειτονικές λέξεις εντός συγκεκριμένου εύρους. Η διαδικασία ξεκινάει με τη λέξη w_t που επιθυμούμε να ταξινομήσουμε. Επιλέγουμε ένα παράθυρο γειτονικών λέξεων μήκους d_{win} αριστερά και δεξιά της λέξης αυτής . Κάθε λέξη εντός του παραθύρου προβάλλεται από τον πίνακα απεικονίσης χαρακτηριστικών σε ένα διάνυσμα ενσωματώσιμων χαρακτηριστικών λεξιλογίου. Με συνένωση των επιμέρους διανυσμάτων συντελεστών προκύπτει ένα μητρώο χαρακτηριστικών 3.33 που περιγράφει συνολικά το τμήμα αυτο συμφραζομένων.

$$f_1^\theta = \left\langle LT_w\left([w]_1^T\right)\right\rangle_t^{d_{win}} = \begin{pmatrix} \langle W\rangle_{t-d_{win}/2}^1 \\ \vdots \\ \langle W\rangle_t^1 \\ \vdots \\ \langle W\rangle_{t+d_{win}/2}^1 \end{pmatrix} \tag{3.33}$$

Για λόγους ευκολίας συνήθως μετατρέπουμε το παραπάνω μητρώο σε διάνυσμα χαρακτηριστικών με απλή παράθεση των στηλών του μητρώου. Το παραπάνω διάνυσμα έχει σταθερά μήκος $d_{wrd} \cdot d_{win}$ και τροφοδοτείται σε ένα ή περισσότερα στρώματα του νευρωνικού δικτύου. Αρχικά αυτό το σταθερού μήκους διάνυσμα f_1^θ τροφοδοτείται σε ένα γραμμικό στρώμμα όπως το 3.34. Το πλήθος των κρυφών μονάδων n_{hidden}^l που θα χρησιμοποιηθούν επιλέγεται από το σχεδιαστή κατάλληλα.

$$\begin{aligned} f_\theta^l = W^l f_\theta^{l-1} + b^l \quad , \quad W &\in \mathcal{R}^{n_{hidden}^l \times n_{hidden}^{l-1}} \\ , \quad b &\in \mathcal{R}^{n_{hidden}^l} \end{aligned} \tag{3.34}$$

Πολλά απλά γραμμικά στρώματα συνδυάζονται και μετασχηματίζονται μέσω μιας μη γραμμικής συνάρτησης για την εξαγωγή υψηλά μη γραμμικών χαρακτηριστικών. Αξίζει να τονίσουμε ότι αν δεν εφαρμόσουμε αυτό το μη γραμμικό μετασχηματισμό το δίκτυο θα ήταν ένα απλό γραμμικό μοντέλο. Αντίθετα η εξαγωγή αυτών των καλύτερων μη-γραμμικών χαρακτηριστικών (high level features) θα δούμε ότι βελτίωνει την απόδοση γενίκευσης. Ως μη γραμμικό μετασχηματισμό οι Collobert et al. προτείνουν μια αυστηρή εκδοχή της υπερβολικής συνάρτησης εφαπτομένης η οποία είναι υπολογιστικά αποδοτική και έχει περίπου την ίδια απόδοση γενικεύσης συγκριτικά με την γνωστή συνάρτηση υπερβολικής εφαπτομένης.

$$\begin{aligned} \left[f_\theta^l\right]_i &= HardTanh\left(\left[f_\theta^{l-1}\right]_i\right) \\ HardTanh\left(x\right) &= \begin{cases} -1 & if \quad x < -1 \\ x & if \quad -1 \leq x \leq 1 \\ 1 & if \quad x > 1 \end{cases} \end{aligned} \tag{3.35}$$

Τέλος επιλέγεται μια διάσταση έξοδου ίση με το πλήθος των πιθανών κατηγοριών ταξινόμησης για την εφαρμογή ενδιαφέροντος. Κάθε έξοδος μετρά τη πιθανότητα που δίνει ο ταξινομητής να ανήκει το πρότυπο στην αντίστοιχη κατηγορία βάσει μιας καλά επιλεγμένης συνάρτησης κόστους.

Εξαγωγή τοπικών χαρακτηριστικών με κινητό παράθυρο λέξεων (συνελικτική μέθοδός)

Στις περισσότερες περιπτώσεις τα τοπικά χαρακτηριστικά από το σταθερό παράθυρο λέξεων δίνουν πολύ καλά αποτελέσματα. Υπάρχουν όμως και περιπτώσεις όπου η ταξινόμηση μιας λέξης απαιτεί να ληφθεί υπόψη ολόκληρη η πρόταση. Σε αυτή τη περίπτωση μια αρκετά φυσική επιλογή είναι η χρήση των συνελικτικών δικτύων. Ένα κλασικό παράδειγμα είναι το semantic role labeling όπου η επισημείωση μίας λέξης εξαρτάται από το ρήμα ή καλύτερα από το κατηγόρημα το οποίο πρέπει να έχουμε εντοπίσει εκ των προτέρων. Αν το ρήμα βρίσκεται εκτός παράθυρου τότε είναι λογικό η λέξη να μην ταξινομηθεί σωστά. Παρακάτω περιγράφουμε με λεπτομέρεια τη βασική αρχιτεκτονική ενός συνελικτικού δικτύου για την περίπτωση που χρείαζεται να ληφθέι υπόψη ολόκληρη η πρόταση για να ολοκληρωθεί επιτυχώς το έργο της ταξινόμησης.

Το δίκτυο για αρχή παίρνει ολόκληρη την πρόταση και μετασχηματίζει τους επιμέρους όρους της εφαρμόζοντας τον πίνακα απεικόνισης χαρακτηριστικών. Το μητρώο που προκύπτει για λόγους υπολογιστικούς το μετατρέπουμε όπως και στη περίπτωση που είχαμε σταθερό παράθυρο λέξεων σε ένα ολικό διάνυσμα χαρακτηριστικών το οποίο στη συνέχεια τροφοδοτείται σε μετέπειτα στρώματα αφινικών μετασχηματισμών.

Το συνελικτικό στρώμα είναι μία γενίκευση της μεθόδου εξαγωγής τοπικών χαρακτηριστικών με ένα σταθερό παράθυρο λέξεων. Παίρνοντας ως είσοδο μια ακολουθία στηλών του πίνακα f_θ^{l-1} εφαρμόζει τη σχέση (3.36) (δηλαδή ένα γραμμικό συνδυασμό στις τιμές των διανύσματων χαρακτηριστικών) σε κάθε παράθυρο της ακολουθίας των επικαλυπτόμενων παραθύρων .

$$\left\langle f_\theta^l \right\rangle_t^1 = W^l \left\langle f_\theta^{l-1} \right\rangle_t^{d_{win}} + b^l \quad , \forall t \tag{3.36}$$

Τα τοπικά χαρακτηριστικά που εξάγονται από κάθε παράθυρο της ακολουθίας μπορούν να χρησιμοποιηθούν ως είσοδοι για την εξαγωγή ακόμα πιο υψηλού επιπέδου χαρακτηριστικών τροφοδοτώντας τα σε στρώματα αφινικών και μη γραμμικών μετασχηματισμών.

Τέλος ένα στρώμα μέγιστου είναι απαραίτητο διότι οι εξόδοι των συνελικτικών στρωμάτων έχουν διαστάσεις που μεταβάλλονται με το πλήθος των όρων της πρότασης. Τα διανύσματα τοπικών χαρακτηριστικών που εξάγουν τα συνελικτικά στρώματα πρέπει να συνδυαστούν σε ένα ενιαίο διάνυσμα τοπικών χαρακτηριστικών με μέγεθος ανεξάρτητο του μήκους της πρότασης εισόδου για να τα τροφοδοτήσουμε μετά στα στρώματα αφινικών μετασχηματισμών. Έχουν καθιερωθεί δυο τρόποι για να γίνει αυτό. Ο ένας είναι μια σταθμισμένη μεσοτίμηση των διανυσμάτων τοπικών χαρακτηριστικών κι ο άλλος η εύρεση ενός μεγίστου στη διάσταση του «χρόνου». Μιλάμε για χρόνο γιατί παραδοσιακά τέτοιες αρχιτεκτονικές είχαν χρησιμοποιηθεί για αναγνώριση φωνής [34]. Στη περίπτωση μας το παράθυρο χρόνου (frame) είναι παράθυρο λέξεων στην πρόταση το οποίο κινείται ανα λέξη καλύπτωντας όλη τη πρόταση. Η μεσοτίμηση των διανυσμάτων χαρακτηριστικών λεξιλογίου δεν έχει νόημα εν προκειμένω λόγω της διακριτής φύσης του προβλήματος. Πολλές λέξεις δε συνεισφέρουν άλλωστε καθόλου στο έργο της επισημείωσης ενώ άλλες συνεισφέρουν αποφασιστικά. Με τη συνάρτηση μεγίστου αναγκάζουμε το δίκτυο να ενσωματώσει τα πιο χρήσιμα τοπικά χαρακτηριστικά που έχουν εξαγάγει τα συνελικτικά στρώματα. Έτσι, για το μητρώο εξόδου του $l-1$-στου συνελικτού στρώματος f_θ^{l-1} το στρώμα μεγίστου δίνει έξοδο το διάνυσμα f_θ^l:

$$\left[f_\theta^l\right]_i \quad = \quad \max_t \left[f_\theta^{l-1}\right]_{i,t} \quad 1 \leq i \leq n_{hidden}^{l-1} \tag{3.37}$$

Το προκύπτον ολικό διάνυσμα χαρακτηριστικών έχει σταθερό μήκος και μπορεί να χρησιμοποιήθει ακώλυτα από τα μετέπειτα συνήθη στάδια του νευρωνικού δικτύου. Όπως και στη περίπτωση του σταθερού παραθύρου στην έξοδο του νευρωνικόυ παίρνουμε έναν βαθμό βεβαιότητας για κάθε πιθανή κατηγορία ταξινόμησης.

3.6.5 Εκπαίδευση

Όλα τα μοντέλα εκπαιδεύονται με τη μεγιστοποίηση της πιθανοφάνειας των προτύπων εκπαίδευσης χρησιμοποιώντας τη μέθοδο καθόδου στη διεύθυνση της πιο απότομης βαθμίδας. Αν συμβολίσουμε με θ το σύνολο των προσδιοριστέων παραμέτρων του δικτύου οι οποίες εκπαιδεύονται πάνω σε ένα σύνολο παραδειγμάτων $\mathcal{T}$. Η εκπαίδευση θα γίνει με μεγιστοποίηση της πιθανοφάνειας (3.38) ως πρός το σύνολο των παραμέτρων θ:

$$\theta^* \quad = \quad \underset{\theta}{argmax} \left[\sum_{(x,y)\in\mathcal{T}} \log \quad p\left(y|x,\theta\right) \right] \tag{3.38}$$

όπου x πρότυπο εκπαίδευσης είτε σε επίπεδο παραθύρου λέξεων ή πρότασης ολόκληρης. Θεωρώντας την αρνητική πιθανοφάνεια ως συνάρτηση κόστους έχουμε ένα πρόβλημα ελαχιστοποίησης $\mathcal{J}(f_\theta(\cdot))$:

$$\min_\theta \mathcal{J}(f_\theta(\cdot)) = \min_\theta \left[- \sum_{(x,y)\in\mathcal{T}} \log \quad p\left(y|x,\theta\right) \right] \tag{3.39}$$

Οι πιθανότητες $p(\cdot)$ θα δούμε στην επόμενη ενότητα πως μπορούν να υπολογιστούν από τους νευρώνες εξόδου του δικτύου.

3.6.6 Λογαριθμική πιθανοφάνεια σε επίπεδο ατομικών λέξεων

Ένας τρόπος να ερμηνεύσουμε τη συνάρτηση κατανομής πιθανότητας $p\left(i|x,\theta\right)$ είναι να τη θεωρήσουμε ως το βαθμό βεβαίτοτητας με τον οποίο ο ταξινομητής αποφάσιζει ότι το πρότυπο x ανήκει στην κατηγορία i. Δοθέντος ενός προτύπου εισόδου x η έξοδος του νευρωνικού δικτύου $\left[f_{\theta}^{l}\right]_{i}$ μπορεί να ερμηνευτεί ως η κατά συνθήκη πιθανότητα $p\left(i|x,\theta\right)$ επιλέγοντας ως συνάρτηση ενεργοποίησης τη $softmax$ πάνω σε όλες τις δυνατές κατηγορίες:

$$p\left(i|x,\theta\right) \quad = \quad \frac{e^{[f_{\theta}]_{i}}}{\sum_{j} e^{[f_{\theta}]_{j}}} \tag{3.40}$$

Ορίζεται η συνάρτηση $log-add$ σύμφωνα με τη σχέση (3.41)

$$\log \underset{i}{add}\left(z_{i}\right) \quad = \quad \log \sum_{i} e^{z_{i}} \tag{3.41}$$

Η προκύπτουσα συνάρτηση λογαριθμικής πιθανοφάνειας για το πρότυπο είσοδου (x,y) εκφράζεται ως:

$$\log p\left(y|x,\theta\right) \quad = \quad \left[f_{\theta}\right]_{y} \quad - \quad \log \underset{j}{add}[f_{\theta}]_{j} \tag{3.42}$$

Αυτό το κριτήριο καλλείται διεντροπία (cross-entropy) και χρησιμοποιείται συχνά σε προβλήματα ταξινόμησης κειμένου. Υποθέτουμε ότι οι λέξεις ταξινομούνται ανεξάρτητα μεταξύ τους ωστόσο υπάρχουν και περιπτώσεις που πρέπει να λάβουμε υπόψη σε ένα βαθμό και την ταξινόμηση των άλλων λέξεων στη πρόταση προτού αποφασίσουμε για την ταξινόμηση της λέξης που εξετάζουμε. Σε αυτή τη περίπτωση μιλάμε για κριτήρια λογαριθμικής πιθανοφάνειας σε επίπεδο πρότασης [36].

Σε αυτό το σημείο ολοκληρώσαμε την ανάλυση των βασικών αναπαραστάσεων και αλγορίθμων κατηγοριοποίησης κειμένου. Προσπαθήσαμε να εξοικειώσουμε τον αναγνώστη με τις βασικές αρχές και την υποκείμενη θεωρία που στηρίζει την αποτελεσματικότητα των πιο προχωρημένων αλγορίθμων που θα προτείνουμε στη συνέχεια.

Κεφάλαιο 4

Τα κοινωνικά δίκτυα στη διάρκεια έκτακτων καταστάσεων .

4.1 Εισαγωγή

Η μελέτη της παρούσας εργασίας επικεντρώνεται στην αυτόματη ταξινόμηση μηνυμάτων που αφορούν εκτακτές καταστάσεις οπώς αυτές που προκύπτουν από φυσικές ή ανθρωπογενείς καταστροφές και στις μοναδικές προκλήσεις που προσφέρουν σε όσους τις μελετάνε δημιουργώντας την ανάγκη για έρευνα συγκεκριμένων μεθόδων διαχείρισης αυτών των δεδομένων. Σε αυτή την εργασία διερευνούμε τις μεθόδους μελέτης αυτών των καταστροφών υπό το πρίσμα της επεξεργασίας και ανάλυσης των χρονικά κρίσιμων πληροφοριών που δημιουργούν οι διάφορες καταστροφές. Συγκεκριμένα μελετάμε τις μεθόδους επεξεργασίας του περιεχομένου μικροκειμενικής πληροφορίας κοινωνικών δικτύων και της ταξινόμησης αυτών σε πληροφοριακές κατηγορίες χρησιμές σε όσους παίρνουν αποφάσεις και αναλαμβάνουν δράση για την αντιμετώπιση των καταστροφών και των συνεπειών τους.

Έκτακτες καταστάσεις εμφανίζονται έπειτα ελάχιστης ή μηδαμινής προειδοποιήσεως και διαμορφώνουν ένα πεδίο όπου κυριαρχεί η αβεβαιότητα και η έλλειψη πληροφοριών ένω είναι απαραίτητη παράλληλα η λήψη κρίσιμων αποφάσεων σε σύντομο χρονικό διάστημα. Η πρόσφατη έρευνα έχει καταδείξει την συνεισφορά των κοινωνικών δικτύων στην άντληση πληροφόρησης και στην επίγνωση της καταστάσεως που προσφέρουν οι χρήστες με τα μηνύματα που ανταλλάσουν και δημοσιεύουν κατα τη διάρκεια ενός κρίσιμου συμβάντος. Ωστόσο η εκτεταμένη χρήση αυτών των δικτύων και η μαζική ροή δεδομένων που παράγονται σε πολύ μικρό χρονικό διάστημα φέρνει τους ειδικούς αντιμέτωπους με ένα συντριπτικό πλήθος δεδομένων που είναι εντελώς απροετοίμαστοι να διαχειριστούν.

Σκοπός της εργασίας αυτής είναι η συνεισφορά στην έρευνα της επιστήμης των υπολογιστών και στο έργο των προγραμματιστών εφαρμογών της μελέτης των υπολογιστικών μεθόδων που μπορούν να καταστούν εργαλεία στα χέρια όλων εκείνων των επίσημων φορέων, των ανθρωπιστικών οργανώσεων και οποιωνδήποτε τελικών χρηστών δραστηριοποιούνται στην αντιμετώπιση των έκτακτων αυτών καταστάσεων. Εν προκειμένω προσφέρει σε αυτούς τα μέσα και τις τεχνικές που είναι απαραίτητα για την επιτυχή αναγνώριση, διαλογή και οργάνωση του

πολύπλοκου αυτού κυκεώνα δεδομένων που παράγονται από τις έκτακτες καταστάσεις όπως αυτές περιγράψαμε σε ανωτέρω χωρίο.

4.2 Διαχείριση ανθρωπιστικών κρίσεων

Οι ανθρώπινες κοινωνίες έρχονται καθημερινά αντιμέτωπες με πολλές και διαφορετικές απειλές που μπορεί να προέρχονται από το φυσικό ή το ανθρωπογενές περιβάλλον. Ο έλεγχος όλων αυτών των απειλών δε είναι πάντοτε εύκολος με αποτέλεσμα κάποιες από αυτές να όδηγησουν στην εκδήλωση κρίσεων, ακραίων και έκρρυθμων καταστάσεων. Κρίσεις συνταράσσουν μια κοινωνία και δοκιμάζουν τις βασικές της υποδομές. Οι ανθρωπιστικές κρίσεις είναι φαινόμενα μικρής συχνότητας που έχουν όμως μεγάλο αντίκτυπο στη κοινωνία και απειλούν τη σταθερότητα και τη βιωσιμότητα του κοινωνικού ιστού. Η διαχείριση αυτών των κρίσεων με την στενή έννοια του όρου συμπεριλαμβανει ένα σύνολο παραγόντων για την καταπολέμηση του ακραίου φαινομένου και την αντιμετώπιση των επιπτώσεων του. Ο όρος αναφέρεται σε κάθε είδους καταστροφή που μπορεί να απειλήσει μια κοινωνία και σε όλους τους δυνατούς τρόπους αντιμετώπισης και διαχείρισης της κρίσης.

Όταν μια κρίση κάνει την εμφάνιση της κινητοποιούνται άμεσα ομάδες και οργανώσεις ιδιωτικού και δημοσίου τομέα για την αντιμετώπιση της. Σε συνεργασία με τις αρχές και τις τοπικές κοινότητες οι ανθρωπιστικές οργανώσεις αναλαμβάνουν να οργανώσουν τη δράση τους για την καταπολέμηση της κρίσης στον άξονα άμβλυνση (mitigation), προετοιμασία(preparedness), αντιμετώπιση(response) και αποκατάσταση(relief) .Παρόλο που οι φάσεις αυτές φαίνεται να είναι σε ακολουθιακή σειρά στην πράξη μπόρει να επικαλύπτονται [31].

1. **Άμβλυνση**: Το στάδιο της άμβλυσνης περιλαμβάνει όλες εκείνες τις πολιτικές και δράσεις που πρέπει να ληφθούν πρίν εκδηλωθεί η κρίση ώστε να ελαχιστοποιηθεί ο βαθμός της καταστροφής ,όταν αυτή συμβεί (π.χ τοποθέτηση ειδικών υπηρεσιών και εγκαταστάσεων, σχεδίαση κατασκευών που είναι ανθεκτικές στις διάφορες καταστροφές, αστικός σχεδιασμός κ.α) .

2. **Προετοιμασία** : Κατά την προετοιμασία έχουμε λήψη προληπτικών μέτρων για αύξηση της ικανότητας αντιμετώπισης (εξέταση και δοκιμή σεναρίων έκτακτης ανάγκης, μελέτη των προβλημάτων που είναι πιθανό να ανακύψουν, εξέταση των τρωτών σημείων και τους τρόπους ελλάτωσης αυτών κ.α).

3. **Αντιμετώπιση**: Μετά την πρώτη αυτή γραμμή άμυνας έχουμε την κυρίως αντιμετώπιση της κρίσης οταν αυτη έχει πια εξελιχθεί και συνίσταται από όλες εκείνες τις ενέργειες που σκοπό έχουν την μείωση των απειλών ανθρώπινης ζωής, άλλων δευτερευόντων απειλών, και ζημιών που σχετίζονται με το συμβάν (όπως προειδοποιήσεις, εκκενώσεις πληθυσμών, προστασία ανθρώπινων ζωών και περιουσίας, αναζητήσεις και διασώσεις, παροχή ιατρικής βοήθειας σε τραυματισμένους, έκτακτα καταφύγια για τους πληγέντες, αφαίρεση κατεδαφισμένων υλικών κ.α).

4. **Αποκατάσταση**: Τέλος έχουμε την αποκατάσταση ,όταν η κρίση έχει παρέλθει, όπου γίνονται διάφορες προσπάθειες μακροπρόθεσμης ανακατασκευής και επανόρθωσης (επισκευές υποδομών και κτισμάτων, προσπάθειες ομαλοποίησης της καθημερινης και επαγγελματικής ζωής των κατοίκων, κάλυψη βασικών αναγκών των πληγέντων όπως στέγη, τροφή, ρουχισμός κ.α.)

Επειδή οι κρίσεις συχνά χαρακτηρίζονται από πολλαπλά αίτια, πολύπλευρες επιπτώσεις και ποικίλα μέσα αντιμετώπισης η διαχείρηση όλων των απραίτητων πληροφοριών είναι ένα δύσκολο έργο. Ένα αποδοτικό και αξιόπιστο σύστημα διαχείρησης πληροφοριών απαιτεί τη συλλογή, ανάλυση και τον διαμοιρασμό των πληροφοριών που σχετίζονται με την επίγνωση της κατάστασης και τη συνεργασία και οργάνωση των δυνάμεων διάσωσης. Οι πληροφορίες που διαχειρίζονται είναι καθορισμένες και τέτοιες ώστε να υποστηρίζουν τις δράσεις σε όλα τα στάδια της κρίσης.

Στη ψηφιακή εποχή οι πολίτες έχουν τη δυνατότητα να επικοινωνουν μεταξύ τους και να πληροφορούν όλους τους άλλους για την κατάσταση που επικρατεί στη διάρκεια μιας κρίσης μέσω κοινωνικών δικτύων. Αυτό και σε συνδυασμό με τα δεδομένα που συλλέγονται και αναλύονται από τις επίσημες οργανώσεις, τους διαχειριστές, τα μέσα ενημέρωσης οδηγεί σε μια έκρηξη του μεγέθους των δεδομένων αυτών θέτοντας νέες πρόκλησεις για τους οργανισμούς που θα ανταποκριθούν στα ακραία αυτά φαινόμενα. Για παράδειγμα, επειδή τα δεδομένα μιας κρίσεως μπορεί σύντομα να γίνουν ξεπερασμένα με τις αλλαγές των συνθηκών της κρίσεως χρειαζόμαστε αποδοτικά συστήματα και δυναμικές ροές πληροφόρησης. Είναι σημαντικό για την κατανόηση όλου αυτού του οικοσύστηματος πληροφοριών να διερευνήσουμε πως οι άνθρωποι χρησιμοποιούν όλες αυτές τις πληροφορίες, πως τις αναζητούν, πως τις μετασχηματίζουν, πως τις μοιράζονται ακομη και για ποιες απ αυτές αδιαφορούν.

Παρά την αυξημένη διαθεσιμότητα των δεδομένων, νέες προκλήσεις έχουν να προστεθούν στο έργο της διαχείρησης ανθρωπιστικών κρίσεων. Υπάρχουν σοβαρές ανησυχίες για την έλλειψη καθορισμένων πληροφορίων καθώς και μηχανισμών για την διασφάλιση της εγκυρότητας των πηγών. Ακόμη προβήματα δημιουργούνται από τον υπερβολικό φόρτο πληροφοριών που πρέπει να επεξεργαστούν και ταυτόχρονα την έλλειψη δεξιοτήτων ανάλυσης μεγάλων δεδομένων από τους διαχειριστές και χρήστες του συστήματος διαχείρισης πληροφοριών . Εν τέλει εγείρονται και θέματα παραβίασης της ιδωτικότητας των δεδομένων αφού κάποιοι με πρόσχημα τη μελέτη των διάφορων κρίσεων,καταστροφών και κοινωνικών ταραχών νομιμοποιούν τη συλλόγη και ανάλυση μεγάλης κλίμακας προσωπικών δεδομένων.

4.3 Τα κοινωνικά δίκτυα στη διάρκεια έκτακτων καταστάσεων

4.3.1 Μια σύντομη ιστορική ανασκόπηση

Η χρήση του διαδικτύου για την συλλογή και διασπορά πληροφοριών που σχετίζονται με συμβάντα καταστροφών και ως δίαυλος επικοινωνίας μεταξύ των ενδιαφερόμενων ανάγεται

ήδη στα τέλη της δεκαετίας του ενενήντα. Ιστορικοί σημειώνουν την δράση ειδησεογραφικών διαδικτυακών ομάδων και ειδικά ενδιαφερόμενων που επικοινωνούν μέσω ηλεκτρονικού ταχυδρομείου για την οργάνωση των διαδηλωτών στην Ινδονησία το 1998 [15]. Επιπλέον, υπήρξαν ιστοσελίδες που δημιουργήθηκαν τύπου wiki όπου χρήστες ανήρτησαν μηνύματα και φωτογραφικό υλικό από καταστροφικά συμβάντα το 2003 [25]. Απ΄ όσο μπορούμε να γνωρίζουμε το 2004 είναι η πρώτη χρονιά όπου χρήστες δημιουργούν ιστότοπο με πρωτότυπο περιεχόμενο το οποίο συντηρούν για χρονικό διάστημα δέκα ημερών μετά το καταστροφικό τσουνάμι στον Ινδικό Ωκεανό στις 26 Σεπτεμβρίου του ιδίου έτους.

Κατά τον απολογισμό του φονικού τυφώνα Κατρίνα το 2005 που έπληξε τη πόλη της νέας Ορλεάνης στις Ηνωμένες Πολιτείες χρησιμοποιήθηκε το κοινωνικό δίκτυο myspace για την πληροφόρηση όσων δραστηριοποιούνταν στην αποκατάσταση των καταστροφών [28]. Από τις πιο πρώιμες περιπτώσεις που γνωρίζουμε τη χρήση υπηρεσίας μικροϊστολογίου Twitter είναι κατά τη διάρκεια των σφοδρών πυρκαϊών στο Σαν Ντιέγο της Καλιφόρνιας Ηνωμένων Πολιτειών εν έτει 2007. Από εκεί και ύστερα το Twitter θα αποτελέσει το κύριο μέσο στο όποιο καταφεύγουν οι πληττόμενοι πληθυσμοί και όσοι ανησυχούν για να επικοινωνήσουν, να θέσουν ερωτήματα, να συλλέξουν και να διασπείρουν πληροφορίες ή να οργανώσουν τις προσπάθειες αποκατάστασης.

4.3.2 Κοινωνικά δίκτυα και νέες προοπτικές στην αντιμετώπιση έκτακτων καταστάσεων

Κατά τη διάρκεια μιας καταστροφής είναι εκτεταμένη η ροή των πληροφοριών και οι επικοινωνίες των πολιτών λόγω της έντονης ανησυχίας για την ασφάλεια των αγαπημένων τους και συγγενικών προσώπων ·αλλά και από γενικότερη αλληλεγγύη προς τους συνανθρώπους τους επιθυμούν να μάθουν πως έχει η κατάσταση και να μένουν ενήμεροι για την εξέλιξη της. Μετά από ένα καταστροφικό συμβάν είναι τυπικό να κατακλύζεται ο τύπος, τα ειδησεογραφικά δελτία από πληροφορίες σχετικά με αυτό, εικόνες, μαρτυρίες , συζητήσεις και για αρκετές μέρες υπάρχει έντονο ενδιαφέρον και πολύ περισσότερο τις κρίσιμες πρώτες ώρες όπου δεν έχουμε ξεκάθαρη εικόνα του τι έχει συμβεί και είναι επιτακτικό να μάθουμε για να πάρουμε αποφάσεις και να αναλάβουμε δράσεις με διακύβευμα ανθρώπινες ζωές ή ανυπολόγιστο κόστος περιβαλλοντικό και οικονομικό.

Λόγω των ακραίων αυτών φαινομένων δημιουργούνται προβλήματα και ασυνέχειες στα δίκτυα ηλεκτρισμού και τηλεπικοινωνιών με αποτέλεσμα οι παραδοσιακοί τρόποι επικοινωνίας να αποτυγχάνουν. Πολίτες που επηρέαζονται από την καταστροφή μπορεί να μην είναι σε θέση να επικοινωνήσουν με τους οικείους τους ή τις διασωστικές αρχές ενώ η ενημέρωση για την εξέλιξη του φαινομένου από τηλεόρασεως ή ραδιόφωνου μπορεί να είναι αδύνατη. Αντίθετα από την επικράτηση των φορητών συσκευών καθως και την καθολική χρήση του διαδικτύου έχουμε να προσβλέπουμε νέες δυνατότητες επικοινωνίας ακόμα και κάτω από τέτοιες αντίξοες συνθήκες.

Η αυξανόμενη αυτή χρήση των κοινωνικών δικτύων στη διάρκεια των καταστροφών προσέθεσε δυνατότητες πληροφόρησης και επικοινωνίας που δεν ήσαν μέχρι και σήμερα. Έχει πια

καταστεί πάγια πολιτική των δυνάμεων αποκατάστασης να δημοσιεύουν πληροφορίες καθώς και προειδοποιητικά και συμβουλές μέσω αυτών των διαύλων. Ωστόσο δεν είναι μονόδρομη η επικοινωνία αυτή καθώς οι ίδιοι οι πολίτες δημοσιεύουν πρωτότυπες πληροφορίες που επηρεάζουν το έργο της αποκατάστασης στη βάση μιας ίσου προς ίσου αλληλεπίδρασης. Οι χρήστες σε πραγματικό χρόνο αναρτούν κρίσιμες πληροφορίες από καταστάσεις που βίωσαν, προσωπικές μαρτυρίες ή αναμεταδίδουν δεδομένα από τρίτους συνεισφέροντας στην καταστασιακή επίγνωση που αναζητούν οι πληττόμενοι αλλά και όσοι βρίσκονται έξω απ την περιοχή της καταστροφής.

Πλατφόρμες κοινωνικών δικτύων όπως το Facebook και το Twitter καλλιεργούν ένα νέο ανοικτό περιβάλλον και διευκολύνουν τους τρόπους που παράγονται, διαδίδονται και καταναλλώνονται οι πληροφορίες. Το περιεχόμενο των πληροφοριών αυτών παράγεται από τους χρήστες του δικτύου και έχει μεγάλη αξία αφού βελτιώνει την επίγνωση καταστάσεως και εξυπηρετεί την καλύτερη αντιμετώπιση της κρίσεως. Σε αντίθεση με τις παραδοσιάκα αυστηρά δομημένες πληροφορίες που μας παρείχαν οργανωμένες άλλα περιορισμένες πηγές (π.χ ειδησεογραφικά πρακτορία, ανθρωπιστικές οργανώσεις, κρατικές αρχές) έχουμε νέες πιο ελεύθερες πηγές πληροφόρησης σε έναν τεχνολογικά επικρατούμενο κόσμο όπου κυριαρχούν η συνεργασία και ο εθελοντισμός. Καλλιεργούν μια τάση για αφθονία πληροφοριών και ένα γνωσιακό πλεόνασμα με τις ταχύτατες επικοινωνιές μεταξύ εκατομμυρίων χρηστών απ όλο το κόσμο.

Πολλοί είναι αυτοί που αναγνωρίζουν τη χρησιμότητα των πληροφοριών που αναρτώνται στα κοινωνικά δίκτυα ως χρονικά κρίσιμες και σημαντικές για την ασφάλεια και την αποκατάσταση των καταστροφών. Πόλλοι είναι αυτοί που παρακολοθούν και ενσωματώνουν πληροφορίες που δημοσιεύονται στα κοινωνικά δίκτυα στο έργο της διάσωσης από τις επίσημες δυνάμεις αποκατάστασης, τις επίσημες ανθρωπιστικές οργανώσεις , τους δημοσιογράφους, τα μέλη του κοινού και άλλους. Υπάρχει λοιπόν η ανάγκη για γρήγορη και εύκολη οργάνωση όλων αυτών των πληροφοριών με σκοπό τον προσπορισμό νέων πρόσθετων πληροφοριών για το έκτακτο συμβάν, την διάγνωση των αναγκών των πληγέντων για την καλύτερη και έγκαιρη παροχή βοήθειας, την αναγώριση τάσεων του κοινού για την καλύτερη πρόβλεψη των αναγκών του άλλα και του καλύτερου σχεδιασμού της όλης προσπάθειας διάσωσης.

Παρόλο που οι επίσημοι φορείς δείχνουν θερμό ενδιαφέρον στην αξιοποίηση των πληροφοριών των κοινωνικών δικτύων δεν λείπουν εμπόδια και περιορισμοί που πρέπει να ληφθούν υπόψη. Εν πρώτοις εγείρονται θέματα ποιότητας των δεδομένων όπως σε τι βαθμό όσα δημοσιεύονται είναι αληθή ή ακριβή ή αξιόπιστα . Απ΄την άλλη, η χρήση εξειδικευμένου προσωπικού για την διαλογή και ανακατεύθυνση στις ενδιαφερόμενες ομάδες των νέων δεδομένων που δημοσιεύονται είναι χρονικά και οικονομικά ασύμφορο. Η χρήση υπολογιστικών μεθόδων μπορεί να υπερπηδήσει τα εμπόδια ελλαττώνοντας το πλήθος των δεδομένων που χρειάζεται να εξεταστεί από τον άνθρωπο. Η αυτοματοποίηση των μεθόδων διαχείρησης και ταξινόμησης των δεδομένων είναι αναγκαία όταν η ανθρώπινη υπολογιστική δύναμη είναι περιορισμένη. Την περιγραγή και ανάλυση αυτών μεθόδων θα εξετάσουμε στην επόμενη ενότητα.

Είναι συνεπώς μια τεχνολογία που στα χέρια των πολιτών ανασχεδιάζει τα όρια μεταξύ επίσημης και ανεπίσημης διασωστικής πολιτικής. Στην πράξη δημιουργεί μια μετάβαση από τα ξεπερασμένα ιεραρχικά γραμμικά μοντέλα πληροφόρησης και δράσης από τις επίσημες δυνάμεις

διάσωσης στους πολίτες σε ένα πολύπλοκο δημοκρατικό μοντέλο από πολίτες προς πολίτες. Η αναγκαιότητα αυτής της μετάβασης είναι πια απαραίτητη στην ολοένα και περισσότερο ψηφιακή εποχή μας.

4.3.3 Χαρακτηριστικά δεδομένων κοινωνικών δικτύων στη διάρκεια μιας καταστροφής

Τυπικές δραστηριότητες που έκαναν δημοφιλείς αυτές τις πλατφόρμες κοινωνικών δικτύων είναι η επαφή με φίλους και συγγενείς καθώς και η δυνατότητα επικοινωνίας με τρίτους. Παρόλο που κάθε κοινωνικό δίκτυο εξυπηρετεί διαφορετικούς σκοπούς,ομοιότητες και κοινά υπάρχουν. Για παράδειγμα, οι τρεις πιο συχνές δραστηριότητες στο Twitter είναι η δημοσίευση που αφορά την καθημερινότητα του χρήστη, το ανέβασμα και ο διαμοιρασμός φωτογραφιών, και ο σχολιασμός στις δημοσιεύσεις των άλλων χρηστών·στο Facebook τώρα το ανέβασμα και ο διαμοιρασμός φωτογραφιών, ανταλλαγή προσωπικών μηνυμάτων με φίλους, και ο σχολιασμός στις δημοσιεύσεις των άλλων.

Ενδιαφέρον παρουσιάζει η μελέτη των εγγενών χαρακτηριστικών που έχουν οι επικοινωνίες των ανθρώπων κατά από συνθήκες γενικής αναταραχής και πως αυτές εκφράζονται στη συμπεριφορά των χρηστών των κοινωνικών δικτύων. Σύμφωνα με τους Crane et al. [6] υπαρχούν πόλλοι λόγοι που πυροδοτούν μια εντονη αύξηση των επικοινωνιών μέσω κοινωνικών δικτύων. Μπορούμε να τους κατηγοριοποιήσουμε σε ενδογενείς και εξωγενείς. Στους ενδογενείς παράγοντες οφείλεται η διάδοση μίας ιδέας , μιας εικόνας, ενός βίντεο που κατακλύζει το δίκτυο καθώς μεταφέρεται από χρήστη σε χρήστη. Στους εξωγενείς παράγοντες υπάγονται τα φαινόμενα μεγάλης κλίμακας που συμβαίνουν στο φυσικό κόσμο και έχουν μεγάλο αντίκτυπο και για τα οποία υπάρχει ένα ευρύτερο ενδιαφέρον. Έχει παρατηρηθεί μια συστηματική αύξηση των επικοινωνιών μέσω κοινωνικών κατά τη διάρκεια μίας καταστροφής καθώς οι πολίτες αναζητούν άμεση και σε βάθος ενημέρωση Fraustino et al. [17].

Παρόλο που τα παραπάνω παραδείγματα μας δίνουν μια αίσθηση του τι είδους πληροφορίες ανταλλάσουν οι χρήστες κατά τη διάρκεια μίας κρίσης σύμφωνα με τον D. Mileti [19] αυτές μπορούν να ενταχθούν σε τρεις βασικές κατηγορίες: ανθρωπογενές περιβάλλον, φυσικό περιβάλλον, κοινωνικό περιβάλλον. Διευκρινίζει ότι το κοινωνικό περιβάλλον αφορά μηνύματα που περιγράφουν την κατάσταση του πληττόμενου πληθυσμού και την αλληλεπίδρασή του με τη κρίση. Το ανθρωπογενές περιβάλλον σχετίζεται με μηνύματα που ενημερώνουν για την κατάσταση των διαφόρων υποδομών, των κοινοφελών υπηρεσιών και των ιδιωτικών περιουσιών. Τέλος, το φυσικό περιβάλλον περιλαμβάνει μηνύματα που αναφέρονται σε απειλές από διάφορα φυσικά φαινόμενα όπως καιρικές συνθήκες και άλλοι περιβαλλοντικοί παράγοντες.

Το να προσδιορίσουμε την κατανομή των διαφόρων τύπων πληροφορίας των μηνυμάτων που ανταλλάσουν μέσω κοινωνικών δικτύων δε είναι εύκολο έργο. Αυτό διότι παρουσιάζουν έντονες μεταβολές όχι μόνο μεταξύ διαφορετικών κρίσεων άλλα και στη διάρκεια των διαφόρων φάσεων της ίδιας κρίσης. Ακόμα μια μεταβλητή που επηρεάζει το είδος της πληροφορίας είναι η προέλευση των δεδομένων αν δηλαδή είναι εντός ή εκτός της ακτίνας επιρροής της κρίσης. Για παράδειγμα , μηνύματα που αφορούν προειδοποιήσεις ή εκκλήσεις βοηθείας προέρχονται

από χρήστες κοντά στο επίκεντρο της καταστροφής αντίθετα από χρήστες απομακρυσμένους που δεν επηρεάζονται άμεσα προέρχονται εν πολλοίς μηνύματα συμπαράστασης και ανησυχίας.

4.4 Μεγάλα δεδομένα ανθρωπιστικών κρίσεων Big Crisis Data

Με τον όρο μεγάλα δεδομένα (big data) αναφερόμαστε σε όλες εκείνες τις μεθόδους και τεχνολογίες συλλογής , επεξεργασίας και ανάλυσης εκτεταμένων συνόλων δεδομένων. Εταιρείες όπως οι Google, Facebook, Twitter, Yahoo κ.α διαχειρίζονται καθημερινά petabytes δεδομένων που παράγουν και καταναλώνουν οι χρήστες τους. Έχει παρατηρηθεί ότι ο όγκος των δεδομένων συνολικά παρουσιάζει μία εκθετική αύξηση εξαιτίας της συνεχούς ψηφιοποίησης της σύγχρονης ζωής, της ευρείας χρήσης των διαδικτυακών υπηρεσιών και της εμπορευματοποίησης της συλλογής δεδομένων. Ωστόσο μόνο ένα μικρό μέρος αυτών είναι ουσιστικά χρήσιμο. Τα δεδομένα αυτά καθαυτά συνήθως είναι σε μεγάλο βαθμό μη αξιοποιήσιμα εξαιτίας της πολυαποσπασματικής φύσης τους· είναι μη δομημένα και προέρχονται από πολλαπλές πηγές. Εμφανίζουν υψηλό πλεονασμό και είναι διαθέσιμα πολυτροπικά· ως ολοκληρωμένο κείμενο , εικόνα , βίντεο ή σχολίο-δημοσίευση σε κάποιο κοινωνικό δίκτυο. Κατά τη διαχείρηση των πληροφοριών ανθρωπιστικών κρίσεων οι ανταποκρινόμενες (κυβερνητικές) οργανώσεις επεξεργάζονται και χρησιμοποιούν μεγάλα δεδομένα. Με τον όρο μεγάλα δεδομένα κρίσεων ανάφερόμαστε σε μια ποσοτική αύξηση του μεγέθους των συνόλων δεδομένων κρίσεων που μπορούν να χρησιμοποιηθούν για αναλυτικούς σκοπούς. Η επεξεργασία και η αποθήκευση αυτών των δεδομένων αποτελεί μια από τις σημαντικότερες προκλήσεις που αντιμετωπίζουν οι δυνάμεις ασφαλείας, η αστυνομία, η πυροσβεστική, τα νοσοκομεία και όσοι άλλοι κυβερνητικοί οργανισμοί καλούνται να διαχειριστούν μία καταστροφή. Είναι ανάγκη για τις ανταποκρινόμενες οργανώσεις να έχουν το συντομότερο δυνατό και σε πραγματικό χρόνο τα δεδομένα κρίσεων έτοιμα για επεξεργασία για να μπορέσουν να δράσουν και να συνεργαστούν αποδοτικά. Άλλωστε η αποδοτική σχεδιάση και διαχείσιση του περιορισμού και αντιμετώπισης μιας κρίσης έξαρτάται σε μεγάλο βαθμό από τη ποιότητα και τη ποσότητα των δεδομένων που είναι διαθέσιμα για επεξεργασία και ανάλυση. Η γρήγορη διαχείριση και αξιοποίηση αυτών των δεδομένων όχι μόνο βελτιώνει τη λήψη αποφάσεων παρέχοντας μία σαφή εποπτεία της καταστροφής άλλα βοηθά και στην ανάληψη των κατάλληλων δράσεων για την αντιμετώπιση και αποκατάσταση της καταστροφής.

Τα μεγάλα δεδομένα κρίσεων εμφανίζουν τις κατεστημμένες διαστάσεις πολυπλοκότητας που χαρακτηρίζουν εν γένει τα μεγάλα δεδομένα. Στη βιβλιογράφία έχει επικρατήσει το σχήμα απομνημόνευσης των τεσσάρων βασικών χαρακτηριστικών των μεγάλων δεδομένων που καθιέρωσε αθελά του ενδεχομένως ο Gartner [7] και έχει γίνει γνωστό ως τα vexing v's (προβληματικές "ποι-ότητες") των δεδομένων. Αυτές είναι η χωριτικ-ότητα (V-olume), η ταχύ-τητα (V-elocity), η ποικιλ-ότητα (V-ariety) και η πιστ-ότητα (V-eracity) των δεδομένων.

1. **Χωρητικότητα**: Μεγάλες κρίσεις μπορεί να προκαλέσουν μία έκρηξη δραστηριότητας των κοινωνικών δικτύων. Το μέγεθος των δεδομένων μπορεί να γίνει ζήτημα άφου

ενσωματώνοντας και συσσωρεύοντας μηνύματα για αρκετές μέρες μπόρει να χρειαστεί να καταγραφούν εκατομμύρια μηνυμάτων. Παρόλο που το μέγεθος ενός τυπικού μηνύματος δε ξεπερνα τα 4 kB λαμβάνοντας υπόψη και τα μεταδεδομένα που ενδεχομένως να έχουν επισυναφθεί, μια συλλογή εκκατομυρίων μηνυμάτων μπορεί να είναι της τάξεως πόλλων εκατοντάδων megabyte ή μερικών gigabyte.

2. **Ταχύτητα**: Οι ταχύτατοι ρυθμοί με τους οποίους παράγονται αυτά τα δεδομένα κατα τη διάρκεια μιας κρίσης είναι μία σημαντική πρόκληση. Ο μεγαλύτερος ρυθμός δεδομένων που έχει καταγραφεί είναι 16000 μηνύματα στη διάρκεια ενός λεπτού κάτα τον φονικό τυφώνα Sandy. Κάτι που είναι ακόμα πιο απαιτητικό είναι ότι ο ρυθμός των δεδομένων αυτών δε είναι σταθερός τουναντίον παρουσιάζει σφοδρότατες μεταβολές.

3. **Ποικιλότητα**: Μηνύματα ετερογενή από πολλαπλές πηγές(π.χ ειδησεογραφικά πρακτορία, λογαριασμοί αυτόπτων μαρτύρων), δομημένα ,ημιδομημένα ή εντελώς αδόμητα με τα περισσότερα μη δομημένα· άρθρα, εικόνες , βίντεο, σχόλια και δημοσιεύσεις σε κοινωνικά δίκτυα με διαφορετικό γλωσσικό και πολιτιστικό υπόβαθρο συνθέτουν ένα πολύπλοκο μωσαικό δεδομένων που δύσκολα διαχειρίζεται.

4. **Πιστότητα**: Ένα ακόμα ζήτημα είναι η ποιότητα των δεδομένων και ο βαθμός στον οποίο μπορούμε να τα εμπιστευτούμε. Η ποιότητα των δεδομένων περιλαμβάνει μια σειρά από άλλες ιδιότητες όπως η εγκυρότητα (validity) που έχει να κάνει με την αμεροληψία και την αντικειμενικότητα των δεδομένων, η σαφήνεια, η χρονική ετοιμότητα ώστε να διασφαλίζεται το ότι δε διαδίδονται ανυπόστατες φήμες και προπαγανδιστικό υλικό (virality). Εν γένει η αβεβαιότητα για το αληθές και το ακριβές του λόγου που χαρακτηρίζει πληροφορίες από κοινούς χρήστες του δικτύου είναι ένας ακόμα βαθμός ελεθερίας της πολυπλολότητας των δεδομένων που πρέπει να λάβουμε υπόψη.

Πέρα από αυτά υπάρχει μια μεταβλητή που μετρά την χρησιμότητα των δεδομένων και αναφέρεται ως αξία των δεδομένων (data value).Ένα πρότυπο διαχείρισης μεγάλων δεδομένων περιλαμβάνει όλες τις τεχνικές και τα εργαλεία αποθήκευσης, επεξεργασίας και ασφάλειας των δεδομένων. Αντικειμενικός σκοπός όλων των συστημάτων διαχείρισης μεγάλων δεδομένων είναι η ενίσχυση της αξίας και της προσβάσιμότητας των δεδομένων.

4.5 Ταξινόμηση μηνυμάτων κοινωνικών δικτύων

Η ανάλυση περιεχομένου των μηνυμάτων που δημοσιεύονται στα κοινωνικά δίκτυα βρίσκει διάφορες εφαρμογές και μπορεί να χρήσιμοποιηθει για την αναγνώριση πρότυπων συμπεριφοράς των χρηστών ,νέων τάσεων ή την αναζήτησή πρωτότυπων πληροφοριών. Όταν κάνει την εμφάνιση της μια μεγάλη καταστροφή με αντίκτυπο σ όλη τη κοινωνία οι πολίτες μαζικά δημοσιεύουν ενημερώσεις για το πως έχει η κατάσταση και αναζητούν χρήσιμες πληροφορίες για να οργανώσουν τη δράση τους. Η γρήγορη ανάλυση αυτών των μηνυμάτων μπορεί να βοηθήσει τις διάφορες ανθρωπιστικές οργανώσεις όπως τα Ηνωμένα Έθνη ,τους δημοσιογράφους ή τους διάφορους εθελοντές να έχουν μια καλύτερη εικόνα της κατάστασης που επικρατεί

,ποιες επείγουσες ανάγκες των πληττόμενων πρέπει να καλυφθούν ,αν υπάρχουν περιστατικά έκτακτης κατάστασης υγείας κτλ. και να αποφασίσουν κατάλληλα για το ποιες δράσεις πρέπει να ληφθούν.

Χρησιμοποιώντας μεθόδους τεχνητής νοημοσύνης θα προσπαθήσουμε να ανακτήσουμε τις πιο χρήσιμες πληροφορίες για κάθε δράση ενδιαφέροντος. Η αναζήτησή των κατάλληλων πληροφοριών σε πολύ μεγάλα σύνολα δεδομένων δε είναι εύκολη. Συνήθως οι οργανώσεις και τα πρόσωπα προσπαθούν να συλλέξουν τα μηνύματα που τους ενδιαφέρουν βάσει λέξεων κλειδιών. Αυτό μεν μείωνει τον όγκο των δεδομένων προς αναζήτησή αλλά δε επαρκεί για να ανακτήσει τις επιθυμητές πληροφορίες με βάση το περιεχόμενο του μηνύματος. Ένας τρόπος να γίνει αυτό είναι με την εκπαίδευση επιβλεπόμενων ταξινομητών που θα ξεχωρίσουν τα μηνύματα που προσφέρονται για ανάληψη αποφάσεων από τα υπόλοιπα και να τα κατηγοριοποιήσουν κατάλληλα σε τάξεις περιεχομένου .

Η ταξινόμηση αυτών των μικροκειμενικών δεδομένων παρουσιάζει ωστόσο μοναδικές προκλήσεις. Τα περιορισμένης αυτά έκτασης μηνύματα που δημοσιεύονται είναι δύσκολο να τα ταξινομήσουμε νοηματικά βάσει περιεχομένου επειδή έχουμε λίγα συμφραζόμενα. Επιπλέον τα μικροκειμενικά αυτά δεδομένα είναι θορυβώδη και όχι καλά δομημένα. Οι χρήστες χρησιμοποιούν μια κάπως προφορική γλώσσα κι αυτό εγείρει θέματα ποιότητας του λόγου. Χρησιμοποιούνται εν πολλοίς συντμήσεις, ακρώνυμα ή αγοραίες εκφράσεις και εμφανίζουνται γραμματικοί και συντακτικοί τύποί διαφορετικοί από τους αποδεκτούς πράγμα που δημιουργεί αμφισημίες. Ακόμη, τα μηνύματα αυτά πολλές φορές φέρουν τυπογραφικά και ορθογραφικά λάθη ενώ παρατηρείται ότι στη προσπάθεια τους οι χρήστες να μην υπερβούν το όριο χαρακτήρων χρησιμοποιούν λέξεις δίχως κενά είτε ηχομιμητικούς αριθμητικούς χαρακτήρες. Όλες αυτές οι γλωσσικές αλλοιώσεις μπορούν, όπως θα δούμε, να περιοριστούν σε κάποιο βαθμό σε ένα πρώτο στάδιο προεπεξεργασίας και γλωσσικής κανονικοποίησης.

Από την άλλη πλευρά, η κατηγοριοποίηση του περιεχομένου ενός μηνύματος είναι ένα έργο με υψηλό βαθμό υποκειμενικότητας. Οι ειδικοί πολλές φορές δεν μπορούν να συμφωνήσουν για την χρησιμότητα ενός μηνύματος. Ακόμη δυσκολότερο είναι να συμφωνήσουν σε ποιά ακριβώς χρήσιμη κατηγορία πρέπει να εντάξουν ένα μήνυμα ενώ υπάρχουν περιπτώσεις που περρισότερες της μιας κατηγορίας πρέπει να αποδοθούν στο ίδιο μήνυμα. Οι διάφορες δε γλωσσικές αμφισημίες κάνουν δύσκολο να ερμηνευτεί σωστά κάποιο μήνυμα. Όλες αυτές οι εγγενείς δυσκολίες καθιστούν αδύνατο στις μηχανές να συμφωνούν στην νοηματική κατηγοριοποίηση ενός μηνύματος σε βαθμό συγκρίσιμο με εκείνο που συμφωνούν οι ειδικοί επισημειώτες. Αξίζει να αναφερθεί ότι παρά τις εξελίξεις και τις προόδους στην αυτόματη επεξεργασία φυσικής γλώσσας η αυτοματη επεξεργασία σύντομων καθημερινών κειμένων παραμένει ένα δύσκολο πρόβλημα.

Στο έργο της ταξινόμησης μηνυμάτων που αναφέρονται σε μεγαλής κλίμακας καταστροφές μια κλασσική προσέγγιση ήταν η χρήση μεθόδων μάθησης κατά ομάδες (batch learning) και διακριτές διανυσματικές αναπαραστάσεις των όρων περιεχομένου. Αυτή η προσσέγγιση έχει τρείς σημαντικούς περιορισμούς:

1. Κατά την έξαρση του ακραίου φαινομένου δεν υπάρχουν διαθέσιμα επισημειωμένα πρότυπα εκπαίδευσης. Στη συνέχεια τα επισημειωμένα πρότυπα έρχονται σε μικρές ομάδες

ανάλογα με τη διαθεσιμότητα των εθελοντών που είναι κατανεμημένοι γεωγραφικά σε διάφορα μέρη. Η επιτυχία όμως αυτών των μεθόδων εξαρτάται σημαντικά από τη διαθεσιμότητα επισημειωμένων δεδομένων που αφορούν το συμβάν. Επιπλέον, η χρήση διακριτών αναπαραστάσεων των λέξεων σε συνδυασμό με τη ποικιλότητα (variety) των μεγάλων δεδομένων των κρίσεων (big crisis data) από συμβάν σε συμβάν έχει ως αποτέλεσμα ανεπίτρεπτα χαμηλές αποδόσεις των ταξινομητών αυτών αν εκπαιδευθούν μόνο επί ιστορικών δεδομένων άλλων καταστροφών (out-of-event data).

2. Η εκπαίδευση ενός νέου ταξινομητή καθε φορά που μια όμαδα επισημειωμένων προτύπων γίνεται διαθέσιμη είναι ανέφικτη λόγω της ταχύτητας των μεγάλων δεδομένων κρίσεων.

3. Επειδή η κλασσική αυτή προσέγγιση απαιτεί ανθρωπογενή σχεδίαση χαρακτηριστικών όπως η στάθμιση όρων $Tf - Idf$ η προσαρμογή του μοντέλου στις μεταβολές των δεδομένων πρέπει να γίνεται με παρέμβαση του ανθρώπου, πράγμα ανεπιθύμητο και συχνά ανέφικτο.

Αντίθετα τα βαθιά νευρωνικά δίκτυα μπορούν να εκπαιδευτούν πολύ πιο φυσικά και απλά στα μεγάλα δεδομένα κρίσεων. Εκπαιδεύονται συνήθως με μεθόδους συνδεδεμένης μάθησης (online learning) και έχουν την ευελιξία να προσαρμοστούν στα νέα επισημειωμένα πρότυπα χωρίς να χρειάζεται να εκπαιδεύσουμε τον ταξινομητή εξ αρχής. Επιπλέον η χρήση κατανεμημένων αναπαραστέσεων των λέξεων αυξάνει την ικανότητα γενίκευσης τους αξιοποιώντας καλύτερα την εμπειρία από τα ιστορικά δεδομένα άλλων καταστροφών επιταχύνοντας το έργο της ταξινόμησης τις πρώτες ώρες έξαρσης της κρίσεως όπου έχουμε ελάχιστα πρότυπα εκπαίδευσης. Επιπλέον, η αναπαράστση των λέξεων με κατανεμημένα μοντέλα μας απαλάσσει από το δύσκολο έργο της σχεδίασης χαρακτηριστικών. Αυτό έχει μεγαλύτερα περιθώρια γενίκευσης όπως έχει δείξει γενικότερα η χρήση των κατανεμημένων διανυσματικών αναπαραστάσεων σε διάφορες εφαρμογές αυτόματης επεξεργασίας φυσικής γλώσσας.

Κεφάλαιο 5

Ταξινόμηση μηνυμάτων ανθρωπιστικών κρίσεων με συνελικτικά νευρωνικά δίκτυα

5.1 Εισαγωγή

Σ αυτή την ενότητα θα προσπαθήσουμε να αναλύσουμε τα δεδομένα και τους αλγόριθμους που χρησιμοποιήσαμε για την επίλυση του πρόβληματος. Αρχίκα θα εξετάσουμε τον τρόπο με τον οποίο συλλέξαμε τα δεδομένα που είναι απαραίτητα για την εκπαίδευση των μοντέλων που μελετάμε. Αφού ολοκληρωθεί μια διερευνητική ανάλυση των δεδομένων που συνελέγησαν θα αναπτύξουμε τους αλγορίθμους μας συλλογιστικά.

5.2 Το σύνολο δεδομένων DeepCrisis

5.2.1 Προέλευση

Το σύνολο εκπαίδευσης DeepCrisis προέκυψε από αναδρομική δειγματοληψία του 1% των δημόσιων δεδομένων που είναι ελεύθερα διαθέσιμα στο Διαδικτυακό Αρχείο για 37 καταστροφές που έλαβαν χώρα στη διάρκεια των ετών 2012-2017.Πρόκειται για το σύνολο που δημιουργήσαμε για τις ανάγκες της παρούσας εργασίας ενσωματώνοντας δεδομένα από ένα διευρυμμένο σύνολο έτοιμων συλλογών. Το σύνολο αυτό δεδομένων περιλαμβάνει περί τα 50,000 επισημειωμένα μηνύματα μικροιστολογίου Twitter σε 6 κατηγορίες περιεχομένου. Για τη συλλογή των δεδομένων είχε χρησιμοποιηθεί εξειδικευμένο προσωπικό ενώ ένα μέρος είχε ταξινομηθεί από εθελοντές με τη μεθόδο του πληθοπορισμού. [9, 23]

Οι κατηγορίες ταξινόμησης διαφέρουν άναλογα με τις ανάγκες των διάφορων ανθρωπιστικών οργανώσεων και τη φύση της καταστροφής. Το σχήμα ταξινόμησης που χρησιμοποιήθηκε είναι εν γένει ένα υποσύνολο των επισημειώσεων που προτείνει το γραφείο των Ηνωμένων Εθνών για την ρύθμιση ανθρωπιστικών ζητημάτων (United Nations Office for the Coordination of Humanitarian Affairs)(UN OCHA). Περιλαμβάνει έξι κατηγορίες απ τις οποίες δύο

είναι γενικές ώστε να περιλαμβάνουν όσα μηνύματα δε μπορούν να ταξινομηθούν σε καμία απ τις ανωτέρω κατηγορίες.

Η διαδικασία της επισημείωσης περιλαμβάνει τα εξής στάδια: ο ειδικός ή ο εθελοντής που αναλαμβάνει το έργο της ταξινόμησης έχει στη διάθεση του μια λίστα από τις προκαθορισμένες κατηγορίες και το σύνολο των μηνυμάτων προς ταξινόμηση. Αφού διαβάσει το μήνυμα και επισκεφθεί ενδεχομένως διάφορους συνδέσμους αποφασίζει σε ποια κατηγορία θα υπαχθεί το μήνυμα. Μηνύματα που δε ταξινομούνται σε καμία κατηγορία άλλα έχουν σχέση με τη καταστροφή επισημειώνονται στη κλάση «άλλη χρήσιμη πληροφορία» (Other useful information). Η διαδικασία ολοκληρώνεται όταν τρείς τουλάχιστον εθελοντές ή ειδικοί συμφωνήσουν.[9]

Παρακάτω περιγράφουμε το σχήμα ταξινόμησης που χρησιμοποιήσαμε με βάση το περιεχόμενο των μηνυμάτων:

1. Affected Individuals (Πληγέντες): νεκροί· αναφορές σε παγιδευμένους ή αγνοούντες, αναζήτηση αγνοούμενων , αναφορές ανθρώπων που βρέθηκαν ή τους είδαν, τραυματισμένοι ή σε έκτακτη ανάγκη ιατρικής βοήθειας, ενημερώσεις για την κατάσταση των πληγέντων ή προσωπικές ενημερώσεις κτλ.

2. Donations and Volunteering(Εθελοντική προσφορά): προσφορά υλικών αγαθών , τροφίμων , χρημάτων ή υπηρεσιών· δωρεές και εθελοντικές δράσεις για την ανακούφιση του πόνου των πληγέντων· μηνύματα που αιτούνται βοηθείας ή κάνουν προτάσεις και συντονίζουν τις προσπάθειες αποκατάστασης·εκστρατείες συλλογής χρηματικών πόρων·πληροφορίες εθελοντών· προσφορά φιλοξενείας στους πληγέντες και πληροφορίες για διαθεσιμότητα καταφύγιου και τροφής.

3. Infrastructure and Utilities (Υποδομές): υλικές ζημιές και ανθρώπινα θύματα· μηνύματα σχετικά με τις επιπτώσεις της καταστροφής στο φυσικό και ανθρωπογενές περιβάλλον· αναφορές σε διακοπή υπηρεσιών· κατάρρευση κτισμάτων ,προβλήματα και ζημιές στο δίκτυο υδροδότησης ή σε νοσοκομεία και κλινικές · κλείσιμο δρόμων και έκτακτες κυκλοφοριακές ρυθμίσεις.

4. Sympathy and Support (Ηθική Συμπαράσταση): έκφραση ανησυχίας και συμπόνιας στο δράμα των πληγέντων· μηνύματα ευγνωμοσύνης ή συμπαράστασης· ευχολόγια και προσευχές ή συναισθηματικά φορτισμένα μηνύματα στήριξης.

5. Other Useful Information (Άλλες χρησιμες πληροφορίες): χρήσιμα τηλέφωνα ή έκτακτες τηλεφωνικές γραμμές βοήθειας, πληροφορίες σχετικές με το καιρό, την ορατότητα, τους ανέμους·κρίσιμες τοποθεσίες·αναφορές καπνού ή απανθρακωμένων·πρόσθετες πληροφορίες ή μετα-δισσυσσιον,επιβεβαίωση πληροφοριών ή διευκρίνιση ειδικών ζητημάτων.

6. Not related or irrelevant (Μη-σχετικό με τη καταστροφή): εκτός θέματος ή ψευδείς φημολογίες·χιουμοριστικά ή εμπαικτικά σχόλια· άσχετα ή αυτοματοποιημένα ακατανόητα και κακόβουλα μηνύματα.

Πίνακας 5.1: Περιγραφή των τάξεων στα σύνολα δεδομένων. Ο συνολικός αριθμός των επισημειωμένων δεδομένων για κάθε τάξη εμφανίζεται στη στήλη ετικέτες

Κλάση	Ετικέτες	Περιγραφή
Affected Individuals	8179	Αναφορές σε νεκρούς, τραυματίες, εξαφανισθέντες ή μετακινήσεις πληθυσμών
Donations and Volunteering	5168	Μηνύματα που αφορούν προσφορά(φαγητού, στέγης, υπηρεσίων) ή εθελοντικές δράσεις
Infrastructure and Utilities	3932	Αναφορές σε καταστροφές εγκαταστάσεων και υποδομών
Sympathy and Support	7155	Μηνύματα ηθικής και συναισθηματικής υποστήριξης
Other Useful Information	14219	Μηνύματα που περιέχουν χρήσιμες πληροφορίες αλλά δεν εντάσσονται σε καμία απ'τις παραπάνω κατηγορίες
Not related or irrelevant	15302	Μηνύματα μη σχετικά με το συμβάν ή δεν προσφέρουν καμία πληροφόρηση

5.2.2 Στατιστική περιγραφή των δεδομένων

5.3 Προεπεξεργασία Δεδομένων

Οι υπηρεσίες μικροιστολογίου όπως το Twitter αποτελούν την κύρια πηγή ανοιχτών δεδομένων για την διαχείριση ανθρωπιστικών κρίσεων. Έχουν προταθεί διάφορα εργαλεία και τεχνικές για μια πιο προηγμένη και σε βάθος γλωσσολογική ανάλυση με σκοπό τη βελτίωση των επιδόσεων της αυτόματης ταξινόμησης κειμένου. Επειδή όμως αυτά τα εργαλεία συνήθως έχουν εκπαιδευτεί σε μεγάλα καθιερωμένα σώματα δεδομένων που προέρχονται από τον ειδησιογραφικό χώρο όπως το Wall Street Journal corpus παρατηρείται μια υποβάθμιση των επιδόσεων στα μικροκειμενικά δεδομένα κοινωνικών δικτύων εξαιτίας των ιδιαίτερων χαρακτηριστικών του λόγου που χρησιμοποιείται σε αυτά.

Ένα βασικό τμήμα της αρχιτεκτονικής σωλήνωσης για την επεξεργασία φυσικού λόγου είναι η επισημείωση των λεκτικών ως προς το μέρους του λόγου (Part Of Speech tagging) καθώς αποτελεί ένα βασικό εργαλείο συντακτικής ανάλυσης. Ο Brendan O'Connor [22] έχει προτείνει έναν ειδικό επισημειώτη ως προς το μέρος του λόγου για την αγγλική ειδικά σχεδιασμένο για μικροκειμενικά δεδομένα της πλατφόρμας Twitter. Επειδή όμως δε θέλουμε να επωμιστούμε μια εκτεταμμένη σχεδιάση χαρακτηριστικών θα αρκεστούμε σε ένα πρώιμο στάδιο του συστήματος επισημείωσης που προτείνει ο O'Connor. Σ αυτό το πρώτο στάδιο επιτελείται μια αδρομερής επισημείωση ειδικών οντοτήτων (entities) που συστηματικά χρησιμοποιούνται στο Twitter και κάποιες βασικές τεχνικές γλωσσικής κανονικοποίησης (text normalization).

Κατ΄ αρχάς όλα τα λεκτικά κανονικοποιούνται αδιακρίτως σε μικρογράμματη γραφή και περιορίζουμε την επανάληψη χαρακτήρων στο μέγιστο δύο. Στη συνέχεια με τη χρήση κανονικών εκφράσεων αντικαθιστούμε τις αναφορές σε ονόματα χρηστών (at-mentions) με το σταθερό λεκτικό userID, τους συνδέσμους αναζήτησης (urls) αντίστοιχα με το HTTP και όλα τα ψηφία ομοίως με το D. Αφαιρούμε όλα τα σημεία στίξεως εκτός της περιόδου, ημιπεριόδου, ερωτηματικού και θαυμαστικού. Τέλος απομονώνουμε τους λεκτικούς όρους (tokenisation) με το εργαλείο CMU TweetNLP.

Προσπαθήσαμε να προκύψει ένα απλό σύστημα επισημείωσης που προσαρμόζεται στα ιδιαίτερα χαρακτηριστικά των δεδομένων μας ενώ παράλληλα παρέχει ένα ικανοποιητικό επίπεδο γλωσσολογικής εμβάθυνσης. Η επιτυχία αυτής της πρακτικής θα μας επιτρέψει να έχουμε

Πίνακας 5.2: Κατανομή του πληθυσμού των επισημειωμένων μηνυμάτων (tweets) για το σύνολο των καταστροφών που περιλαμβάνονται στο σύνολο DeepCrisis

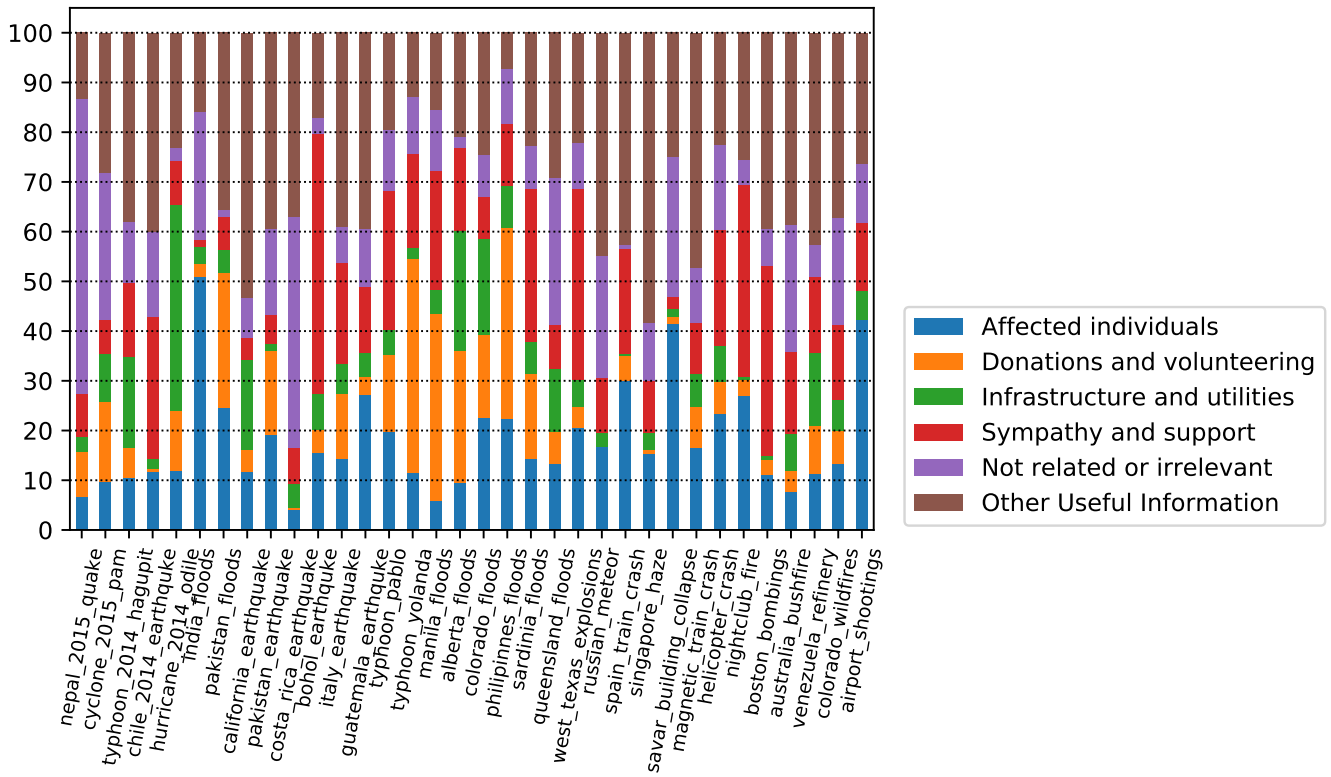
Έτος	Κατηγορία Κρίσεως	Ονομασία	Χώρα	# Tweets
2017	Τυφώνας	Τυφώνας Χάρβεϋ	Τέξας Ηνωμένων Πολιτειών	2970
2017	Τυφώνας	Τυφώνας Ίρμα	Φλώριντα Ηνωμένων Πολιτειών	3754
2017	Τυφώνας	Τυφώνας Μαρία	Δομινίκη	3263
2015	Σεισμός	Σεισμός στο Νεπάλ	Νεπάλ	11314
2015	Τυφώνας	Κυκλώνας Παμ (Pam)	Βανουάτου	2418
2014	Τυφώνας	Τυφώνας Άγουπιτ (Hagupit)	Φιλιππίνες	1924
2014	Σεισμός	Σεισμός στη Χιλή	Χιλή	2123
2014	Τυφώνας	Ανεμοστρόβιλος Οδίλη (Odile)	Μεξικό	2079
2014	Πλημμύρες	Πλημμύρες στην Ινδία	Ινδία	1959
2014	Πλημμύρες	Πλημμύρες στο Πακιστάν	Πακιστάν	1957
2014	Σεισμός	Σεισμός στη Καλιφόρνια	Ηνωμένες Πολιτείες	1929
2013	Σεισμός	Σεισμός στο Πακιστάν	Πακιστάν	1938
2012	Σεισμός	Σεισμός στη Κόστα Ρίκα	Κόστα Ρίκα	1075
2013	Σεισμός	Σεισμός στο Μπόχολ	Φιλιππίνες	939
2012	Σεισμός	Σεισμός στην Ιταλία	Ιταλία	820
2012	Σεισμός	Σεισμός στη Γουατεμάλα	Γουατεμάλα	944
2012	Τυφώνας	Τυφώνας Πάμπλο Pablo	Φιλιππίνες	751
2013	Τυφώνας	Τυφώνας Γιολάντα Yolanda	Φιλιππίνες	951
2012	Πλημμύρες	Πλημμύρες στη Μανίλα	Φιλιππίνες	640
2013	Πλημμύρες	Πλημμύρες στην Αλβέρτα	Καναδάς	826
2013	Πλημμύρες	Πλημμύρες στο Κολοράντο	Ηνωμένες Πολιτείες	903
2012	Πλημμύρες	Πλημμύρες στις Φιλιππίνες	Φιλιππίνες	863
2013	Πλημμύρες	Πλημμύρες στη Σαρδηνία	Ιταλία	861
2013	Πλημμύρες	Πλημμύρες στο Κουίνσλαντ (Queensland)	Αυστραλία	954
2013	Εκρηκτικά	Επίθεση με εκρηκτικά στο Δυτικό Τέξας	Ηνωμένες Πολιτείες	957
2013	Μετεωρίτης	Πτώση Μετεωρίτη στη Ρωσία	Ρωσία	1243
2013	Εκτροχιασμός	Σιδηροδρομικό δυστύχημα στην Ισπανία	Ισπανία	991
2013	Ομίχλη	Ομίχλη στη Σιγκαπούρης	Σιγκαπούρη	582
2013	Κατάρρευση κτιρίου	Κατάρρευση κτιρίου στη Σαβάρ	Μπανγκλαντές	582
2013	Εκτροχιασμός	Lac-Magnetic σιδηροδρομικό δυστύχημα	Καναδάς	975
2013	Πυροβολισμοί	Πυροβολισμοί στο αεροδρόμιο του Λος Άντζελες	Ηνωμένες Πολιτείες	998
2013	Συντριβή	Συντριβή ελικοπτέρου στη Γλασκώυη	Ηνωμένο Βασίλειο	1060
2013	Πυρκαιά	Πυρκαιά σε νυχτερινό κέντρο της Βραζιλίας	Βραζιλία	968
2013	Βομβιστική	Βομβιστική επίθεση στη Βοστώνη	Ηνωμένες Πολιτείες	965
2013	Πυρκαιά	Πυρκαιά στην Αυστραλία	Αυστραλία	978
2012	Εκρηκτικά	Επίθεση με εκρηκτικά στη Βενεζουέλα	Βενεζουέλα	928
2012	Πυρκαιά	Πυρκαιά στο Κολοράντο	Ηνωμένες Πολιτείες	1145

έναν αποδοτικό ταξινομητή επιβλεπόμενης μάθησης που θα προσαρμόζεται ικανοποιητικά σε νέες κρίσεις. Στόχος αυτού του σταδίου είναι η «αποθορυβοποιήση» των δεδομένων από μη κανονικές εκδοχές του λεξιλογίου που υιοθετούν οι χρήστες· φαινόμενο σύνηθες εξαιτίας της φύσης του λόγου που χρησιμοποιείται στο Twitter που είναι σύντομος, άτυπος και μοιάζει πιο πολύ με προφορικό λόγο.

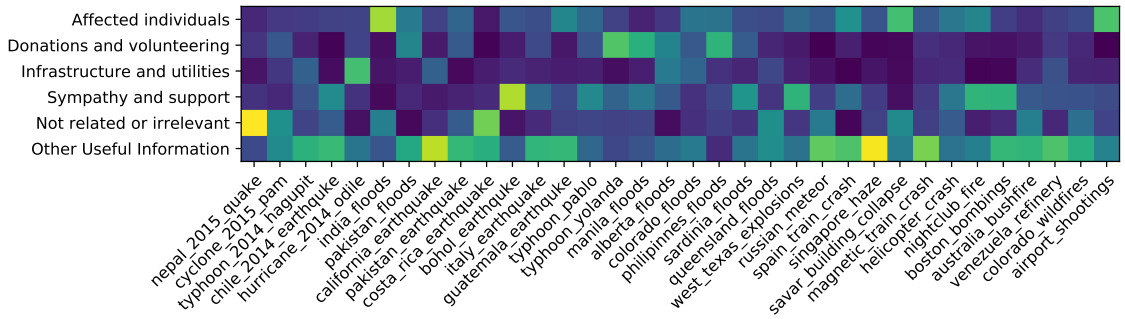
5.4 Συνολική Αρχιτεκτονική των Προτεινόμενων Δικτύων

5.4.1 Εισαγωγή

Με τη χρήση των βαθέων νευρωνικών δικτύων έχει επιτευχθεί σημαντική πρόοδος τόσο σε εφαρμογές υπολογιστικής όρασης όσο και σε εφαρμογές αναγνώρισης ανθρώπινης ομιλίας. Στο πλαίσιο της αυτόματης επεξεργασίας φυσικής γλώσσας εν πολλοίς είχαν χρησιμοποιηθεί για την εκπαίδευση κατανεμημένων αναπαραστάσεων λεξιλογίου (word embeddings) και στην σχεδίαση νευρωνικών γλωσσικών μοντέλων για το έργο της ταξινόμησης κειμένου. Πρόσφατα



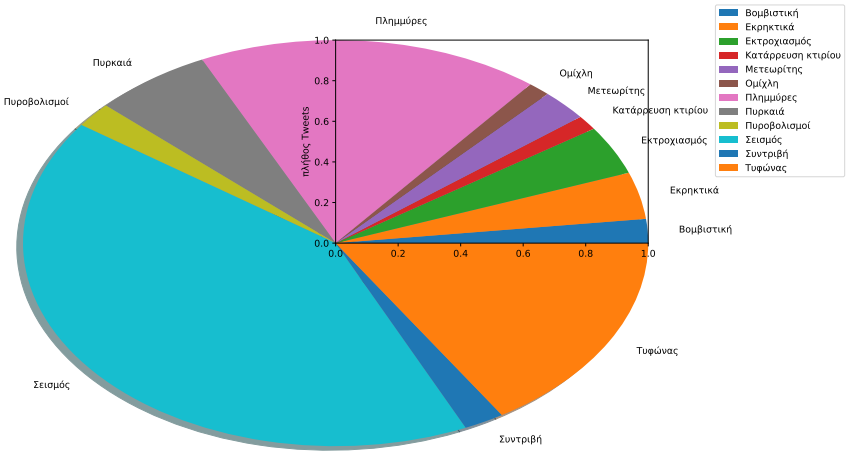
(α′) Ραβδόγραμμα κατανομής κατηγοριών περιεχομένου ανα συμβάν



(β′) HeatMap κατανομής κατηγοριών περιεχομένου ανα συμβάν

Σχήμα 5.1: Κατανομή κατηγοριών πληροφορίας

έγινε διαδεδομένη η χρήση μιας ειδικής μορφής εμπροσθόδρομων νευρωνικών δικτύων για τη μοντελοποίηση της σημασιολογίας σε επίπεδο πρότασης. Πρόκειται για την τάξη των συνελικτικών νευρωνικών δικτύων δηλαδή αρχιτεκτονικών με τουλάχιστον ένα ή περισσότερα στρώματα συνελικτικών φίλτρων για την εξαγωγή τοπικών χαρακτηριστικών.



Σχήμα 5.2: Κατανομή κατηγοριών κρίσεων

Στην παρούσα εργασία εκπαιδεύσαμε ένα απλό και ένα σύνθετο συνελικτικό δίκτυο. Το απλό δίκτυο χρησιμοποιεί ένα στρώμα συνελικτικών φίλτρων και ένα στρώμα τοπικού μεγιστοποιητικού συνδυασμού χαρακτηριστικών (local max pooling). Το σύνθετο δίκτυο απ την άλλη χρησιμοποιεί επιπλέον στρώματα συνελικτικών φίλτρων και στρώματα όλικου μεγιστοποιητικού συνδυασμού χαρακτηριστικών (global max pooling). Και οι δυο αρχιτεκτονικές υποστηρίζουν τις αρχές των (Collobert et al) [4] προσαρμοσμένες στο έργο της ταξινόμησης περιεχομένου σε επίπεδο πρότασης (sentence modeling).

Η εκπαίδευση των δικτύων γίνεται στη βάση διανυσμάτων λεξιλογίου όπως αυτά προκύπτουν από ένα μη επιβλεπόμενο νευρωνικό γλωσσικό μοντέλο. Αυτά τα μοντέλα αναπαράστασης είναι προ εκπαιδευμένα πάνω σε μεγάλα σύνολα δεδομένων γενικού σκοπού. Θα μπορούσαμε με δεδομένη την αναπαράσταση των λέξεων να εκπαιδεύσουμε τις υπόλοιπες παραμέτρους του δικτύου. Πράγματι με μικρές αναπροσαρμογές των υπερπαραμέτρων του δικτύου το απλό αυτό μοντέλο πετυχαίνει εξαιρετικά αποτελέσματα σε πολλαπλά benchmarks και συνεπώς αυτά τα προεκπαιδευμένα διάνυσματα είναι καθολικοί εξαγωγείς χαρακτηριστικών και μπορούν να χρησιμοποιηθούν σε πολλά και διαφορετικά έργα ταξινόμησης κειμένου. Η εκπαίδευση εξειδικευμένων διανυσμάτων λεξιλόγιου με προσαρμογή των συντελεστών της αναπαράστασης στο εκάστοτε έργο ταξινόμησης δίνει ακόμα καλύτερα αποτελέσματα.

5.4.2 Απλό συνελικτικό νευρωνικό δίκτυο (ΣΝΔ)

Η αρχιτεκτονική του δικτύου φαίνεται στο σχήμα 5.3 και είναι βασισμένη στο πολύ γνώστο μεγιστοποιητικό δίκτυο χρονικής καθυστέρησης (Max-TDNN) [34] για την εξαγωγή τοπικών χαρακτηριστικών περιεχομένου και την ταξινόμηση στις προκαθορισμένες κατηγορίες διαχείρισης ανθρωπιστικής κρίσεως.

Έστω $x_i \in \mathbb{R}^d$ το διάνυσμα ενσωματωμένων χαρακτηριστικών που αντιστοιχεί στην $i-$στη λέξη μίας πρότασης προς ταξινόμηση και έστω $\mathbf{X} \in \mathbb{R}^{T \times d}$ η πρόταση αυτή, μήκους T που περιγράφεται από το μητρώο X (5.1). Εν γένει έστω $\mathbf{x}_{i:i+j} = \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+j}$ η παράθεση δηλαδή των j γειτονικών διανυσμάτων χαρακτηριστικών.

$$\mathbf{X} = [\mathbf{x_1}|\dots|\mathbf{x_T}] \tag{5.1}$$

Ο τελέστης της συνέλιξης περιλαμβάνει ένα φίλτρο $h \in \mathbb{R}^{m\times d}$ το οποίο εφαρμόζεται σε ένα παράθυρο λέξεων μήκους m (receptive field) και παράγει ενα νέο χαρακτηριστικό. Εφαρμόζοντάς το σε κάθε δυνατό παράθυρο λέξεων προκύπτει μια διανυσματική απεικόνιση από το χώρο ενσωματωμένων χαρακτηριστικών σε ένα χώρο χαρακτηριστικών υψηλότερου επιπέδου που σχετίζεται με το περιεχόμενο των m-πολυγραμμικών (m-gramms). Για παράδειγμα, το χαρακτηριστικό c_t προκύπτει από ένα παράθυρο $\mathbf{x}_{t:t+m-1}$ όπως στη σχέση 5.2 όπου f μη γραμμική συνάρτηση ενεργοποίησης (π.χRELU, tanh). Εφαρμόζοντας τη σε επικαλυπτόμενα παράθυρα κατά μήκος της πρότασης έχουμε μια απεικόνιση χαρακτηριστικών $\mathbf{c^i} = [c_1, c_2, \dots, c_{T-m+1}]$. Κάθε φίλτρο εκπαιδεύεται ώστε να αναγνωρίζει ένα συγκεκριμένο m-gramm γι αυτό επαναλαμβάνοντας τη παραπάνω διαδικασία για περισσότερα φίλτρα ($i = 1, 2, \dots, N$) ενδεχομένως διαφορετικού εύρους αποκτάμε πολλαπλά χαρακτηριστικά που αναγνωρίζουν ένα σύνολο σημαντικών m-gramms για το έργο της ταξινόμησης.

$$c_t = f(\mathbf{h}.\mathbf{x}_{t:t+m-1} + b_t) \tag{5.2}$$

Στη συνέχεια ακολουθεί ένας τελεστής μεγιστοποιητικού τοπικού συνδυασμού χαρακτηριστικών 5.3 που προσπαθέι να συλλάβει από κάθε απεικόνιση χαρακτηριστικών εκείνο το χαρακτηριστικό με τη μεγαλύτερη ενεργοποίηση δηλάδη αύτο με τη μεγαλύτερη διαχωριστική ικανότητα. Το p είναι ο παράγοντας τοπικού συνδυασμού με την έννοια ότι η απεικόνιση μ_p εξάγει ανά δύο γειτονικές τιμές τις ακολουθίας εισόδου τη μεγαλύτερη . Για $p = 2$ προκύπτει ακολουθία με το ίδιο μήκος με την ακολουθία εισόδου (με κατάλληλη προσθήκη μηδενικών)

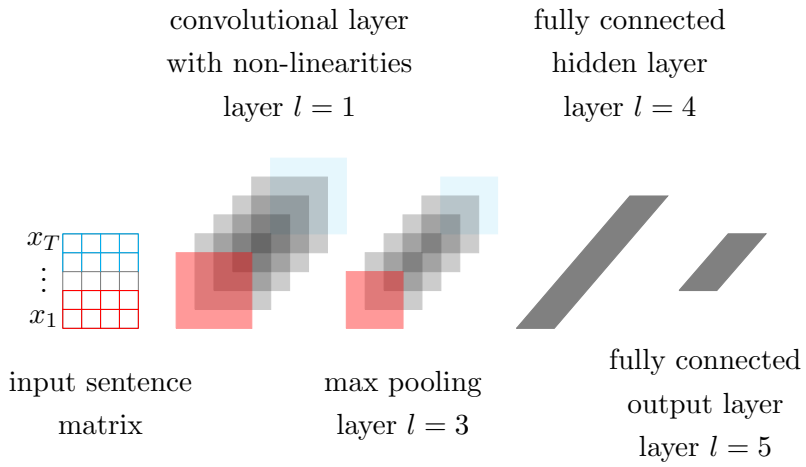
$$\hat{\mathbf{c}} = [\mu_p(\mathbf{c^1}), \mu_p(\mathbf{c^2}), \dots, \mu_p(\mathbf{c^N})] \tag{5.3}$$

Επειδή τα χαρακτηριστικά που εξήχθησαν από τα παραπάνω στρώματα συνελιξης και συνδυασμού εξήχθησαν ανεξάρτητα παρατηρείται ένα αναλλοίωτο στις απόλυτες θέσεις των χαρακτηριστικών και το μοντέλο λειτουργεί σαν σακίδιο πολυγραμμικών (bag-of-ngramms). Αυτό δεν είναι επιθυμητό διότι κατά τη μοντελοποίηση του νοήματος μιας πρότασης οι θέσεις των χαρακτηριστικών παίζουν κάποιο ρόλο και θα προσπαθήσουμε να διατηρήσουμε αυτή την πληροφορία. Χρησιμοποίωντας ένα πλήρως συνδεδεμένο κρυφό επίπεδο νευρώνων εκπαιδεύουμε τον ταξινομητή ώστε να μπορεί να συλλάβει τις αλληλοσχετίσεις μεταξύ των χαρακτηριστικών που επελέγησαν από τα επίπεδα συνέλιξης και συνδυασμού. Το κρύφο αυτό επίπεδο δίνει στην έξοδο του ένα σταθερό αριθμό ενεργοποιήσεων (5.4) που θα τροφοδοτήσουμε στο τελευταίο επίπεδο έξοδου όπου και ολοκληρώνεται το έργο της ταξινόμησης.

$$z = f(\mathbf{Z} \cdot \hat{\mathbf{c}} + \mathbf{b_h}) \tag{5.4}$$

Για την πολυταξική ταξινόμηση το επίπεδο εξόδου υλοποιεί μια συνάρτηση softmax που υπολογίζει την κατανομή πιθανότητας της k-στης επισημείωσης σύμφωνα με τη σχέση 5.5 όπου $\mathbf{w_k}$ είναι τα βάρη από το πλήρως συνδεδεμένο στο επίπεδο εξόδου που σχετίζονται με την k-στη επισημείωση.

$$P(y = k|\mathbf{s}, \theta) = \frac{\exp(\mathbf{w_k} \cdot \mathbf{z} + b_k)}{\sum_{j=1}^{K} \exp(\mathbf{w_j} \cdot \mathbf{z} + b_j)} \tag{5.5}$$



Σχήμα 5.3: Δίκυο CNN

5.4.3 Δυναμικά συνελικτικό νευρωνικό δίκτυο (ΔΣΝΔ)

Το σύνθετο αυτό δίκτυο φαίνεται στο σχήμα 5.4 και χρησιμοποιεί εναλλασσόμενα στρώματα συνέλιξης και δυναμικά $k-$μεγιστοποιητικού ολικού συνδυασμού χαρακτηριστικών (dynamic global k-max pooling). Το πρώτο στρώμα είναι και εδώ μια απεικόνιση των λέξεων της πρότασης εισόδου σε διανύσματα ενσωματωμένων χαρακτηριστικών $x_i \in \mathbb{R}^d$ όπου οι τιμές του x_i είναι παράμετροι προς εκπαίδευση για το δίκτυο. Το κορυφαίο συνελικτικό στρώμα εφαρμόζει ένα φίλτρο $\mathbf{h} \in \mathbb{R}^{m \times d}$ μονοδιάστατης ευρείας συνέλιξης στις συναρτήσεις ενεργοποίησης του προηγούμενου κρυφού επιπέδου ομοίως και το πρώτο συνελικτικό στρώμα απευθείας στο μητρώο ενσωματωμένων χαρακτηριστικών της πρότασης $x \in \mathbb{R}^{T \times d}$.

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_1|\ldots|\mathbf{x}_T], \quad \mathbf{X} \in \mathbb{R}^{d \times T} \\ \mathbf{H} &= [\mathbf{h}_1|\ldots|\mathbf{h}_m], \quad \mathbf{H} \in \mathbb{R}^{d \times m} \end{aligned} \tag{5.6}$$

Η εξαγωγή χαρακτηριστικών ξεκινά εφαρμόζοντας συνέλιξη μεταξύ κάθε γραμμής του πυρήνα του φίλτρου $\mathbf{h}$ και της αντίστοιχης του μητρώου εισόδου $\mathbf{x}$. Το αποτέλεσμα της συνέλιξης εκαφράζεται ως γινόμενο ενός αραιού διαγωνίου πίνακα M με τις τιμές του φίλτρου και κάθε κάθε m-ομάδας τιμών της ακολουθίας εισόδου $\mathbf{x}$. Το αποτέλεσμα της συνέλιξης του

φίλτρου με κάθε παράθυρο λέξεων είναι μια ακολουθία $\mathbf{c}_t$. Ανάλογα με το εύρος τιμών του t η συνέλιξη μπορεί να είναι ευρεία ή στενή. Η στενή συνελίξη απαιτεί το μήκος της πρότασης να είναι μεγαλύτερο του φίλτρου $T \geq m$ και παράγει μια ακολουθία $\mathbf{c} \in \mathbf{R}^{d \times (T-m+1)}$ με το t να κυμαίνεται από m μέχρι T. Αντίθετα η ευρεία συνέλιξη δεν έχει καμία απαίτηση για το μήκος του φίλτρου που θα χρησιμοποιηθεί και παράγει μια ακολουθία $\mathbf{c} \in \mathbb{R}^{d \times (T+m-1)}$ με το t να κυμαίνεται από 1 μέχρι $s+m-1$. Με τη μέθοδο προσθήκης μηδενικών (zero-padding) μπορούμε να ελέγχουμε το είδος της συνέλιξης που θα χρησιμοποιηθεί θεωρώντας πως για τιμές εκτός ορίων της αρχικής ακολουθίας ($t < 1$ ή $t > T$) η ακολουθία εισόδου λαμβάνει μηδενική τιμή.

$$\begin{aligned} \mathbf{c}_t &= \mathbf{M} \cdot \mathbf{x}_{t:t+m-1}, \quad \mathbf{c}_t \in \mathbb{R}^d \\ \mathbf{M} &= [diag(\mathbf{h}_{:,1}) | diag(\mathbf{h}_{:,2}) | \dots | diag(\mathbf{h}_{:,m})] \\ \mathbf{C} &= [\mathbf{c}_1 | \dots | \mathbf{c}_{T+m-1}], \quad \mathbf{c} \in \mathbb{R}^{d \times (T+m-1)} \end{aligned} \tag{5.7}$$

Μετά από κάθε συνελικτικό στρώμα ακολουθεί ένας τελεστής k-μεγιστοποιητικού ολικού συνδυασμού χαρακτηριστικών που αποτελεί γενίκευση του απλού μεγιστοποιητικού ολικού συνδυασμού και διαφορετικός του μεγιστοποιητικού τοπικού συνδυασμού χαρακτηριστικών. Αυτός ο τελεστής υποδειγματοληπτεί τις εξόδους των συνελικτικών στρωμάτων επιλέγοντας εκείνες με τις k πιο υψηλές ενεργοποιήσεις ενώ απορρίπτει όλες τις άλλες. Για παράδειγμα δοθείσης μια ακολουθίας $p \in \mathbb{R}^n$ προκύπτει μετά τη δράση του τελεστή μια υπακολουθία $p_k^{max} \in \mathbb{R}^k$ των k μεγαλύτερων τιμών της p διατηρώντας παράλληλα τη σειρά των τιμών στην αρχική ακολουθία. Η επιλογή του k τώρα μπορεί να αλλάζει συναρτήσει του μήκους της ακολουθίας εισόδου αλλά και του βάθους του δικτύου. Μια απλή επιλογή για το παράγοντα συνδυασμού μεταξύ άλλων συναρτήσεων είναι και η σχέση (5.8) όπου L το συνολικό βάθος του δικτύου και l το συνελικτικό επίπεδο στο οποίο εφαρμόζουμε το τελεστή συνδυασμού χαρακτηριστικών.

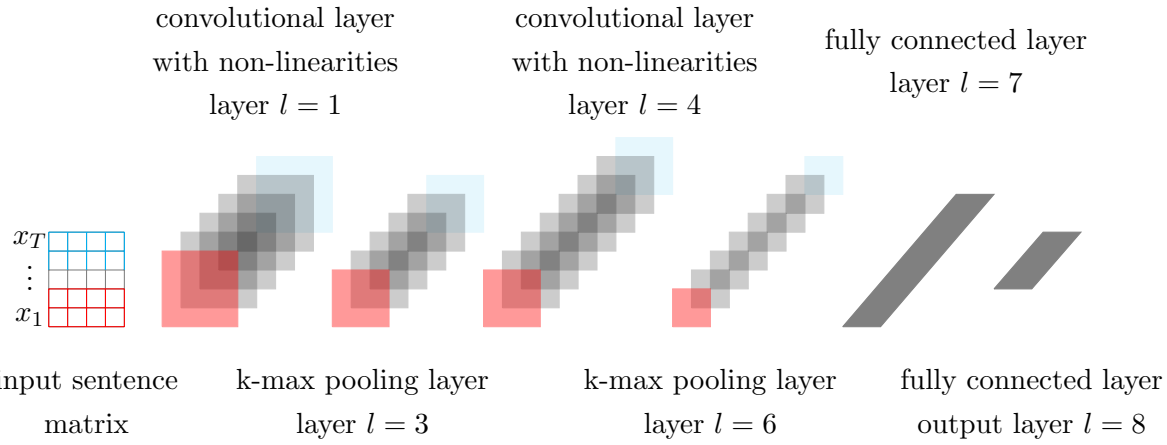
$$\begin{aligned} \mathbf{C}_{max} &= \begin{bmatrix} \mu^{k_l}(\mathbf{c_1}_{,:}) \\ \dots \\ \mu^{k_l}(_{,:}) \end{bmatrix} \quad \mathbf{C}_{max} \in \mathbb{R}^{d \times k_l} \\ k_l &= max(k_{top}, \frac{L-l}{l} \cdot T) \end{aligned} \tag{5.8}$$

Το επίπεδο k-μεγιστοποιητικού συνδυασμού χαρακτηριστικών έχει το νόημα να επιλέξει τα k-καλύτερα χαρακτηριστικά τα οποία όμως μπορεί να απέχουν αρκετά μεταξύ τους στην αρχική ακολουθία $x\cdot$ παράλληλα διατηρείται η σειρά των χαρακτηριστικών δηλαδή οι σχετικές τους θέσεις στην ακολουθία. Έτσι μπορούμε να παρακολουθούμε τις αλληλουχίες υψηλών ενεργοποιήσεων και τις πολλαπλότητες αυτών που μας επιτρέπει μια πιο λεπτομερή και ομαλή ταξινόμηση κειμένου. Στο κορυφαίο συνελικτικό επίπεδο πριν το πλήρες συνδεδεμένο στρώμα ταξινόμησης απαιτείται να έχουμε ένα σταθερό παράγοντα συνδυασμού k_{top} ενώ στα κατώτερα επίπεδα μπορεί να μεταβάλλεται δυναμικά συναρτήσει του μήκους της ακολουθίας εισόδου.

Η απεικόνιση χαρακτηριστικών ολοκληρώνεται εφαρμόζοντας μια μη γραμμική συνάρτηση μετά από κάθε στρώμα δυναμικού συνδυασμού. Προσθέτουμε ένα διάνυσμα πόλωσης $\mathbf{b}_t \in \mathbb{R}^{d \times k_l}$ και εφαρμόζουμε την f σε κάθε στήλη του $\mathbf{C}_{max}$ έτσι προκύπτει κάθε d-διάσταση στήλη του πίνακα χαρακτηριστικών πρώτης τάξης $\hat{\mathbf{C}} \in \mathbb{R}^{d \times k_l}$. Ομοίως για την εξαγωγή χαρακτηριστικών δευτέρας τάξης χρησιμοποιούμε αυτές τις τρεις λειτουργίες, συνελιξης, δυναμικού συνδυασμού, και μη γραμμικής συνάρτησης ενεργοποίησης, σε κάθε παράθυρο $\hat{\mathbf{c}}_{t:t+m-1} \in \mathbb{R}^{d \times m}$ χαρακτηριστικών πρώτης τάξης.

$$\hat{\mathbf{C}} = f(\mathbf{C}_{max} + \mathbf{b}_t), \quad \hat{\mathbf{C}} \in \mathbb{R}^{d \times k_l} \tag{5.9}$$

Μέχρι στιγμής κάθε απεικόνιση χαρακτηριστικών εφαρμόζεται ανεξάρτητα σε κάθε γραμμή του πίνακα της πρότασης. Ένας τρόπος να λάβουμε υπόψη εξαρτήσεις μεταξύ γραμμών είναι να κάνουμε τον πίνακα M πλήρη αντί για αραιά διαγώνιο. Μια απλούστερη μέθοδος για να αποφευχθεί η εισαγωγή επιπλέον παραμέτρων προς εκπαίδευση είναι η αναδίπλωση (folding) του πίνακα χαρακτηριστικών, δηλαδή η αντικατάσταση των γραμμών του με το ανά δύο γραμμών άθροισμα του αρχικού πίνακα, μειώνοντας το μέγεθος της αναπαράστασης κατά $d/2$. Έτσι μετά από κάθε συνελικτικό στρώμα και πριν το στρώμα συνδυασμού εφαρμόζουμε αναδίπλωση στους πίνακες απεικόνισης χαρακτηριστικών. Με αυτο τον τρόπο τα χαρακτηριστικά ανώτερης τάξης εξαρτώνται από τις τιμές των χαρακτηριστικών δύο γραμμών κάθε πίνακα απεικόνισης προηγούμενης τάξης.



Σχήμα 5.4: Δίκτυο DCNN

5.5 Μέθοδοι περιορισμού υπερπροσαρμογής

5.5.1 Εισαγωγή

Οι νευρωνικές αρχιτεκτονικές εξαιτίας της πολυπλοκότητας που τις χαρακτηρίζει έρχονται αναπόφευκτα αντιμέτωπες με το φαινόμενο της υπερπροσαρμογής στα δεδομένα εκπαίδευσης.

Άλλοτε τα δεδομένα μας δεν είναι αρκετά ώστε να εκπαιδεύσουμε την συντριπτική πληθώρα παραμέτρων αυτών των δικτύων. Παρακάτω εξετάζουμε δυο μεθόδους που ενσωματώνονται στην αρχιτεκτονική μας για τη μείωση της υπερπροσαρμογής μέσω της μάθησης πιο αξιόπιστων χαρακτηριστικών.

5.5.2 Χρήση πολυ-κάναλου δικτύου

Η χρήση διανυσμάτων ενσωματωμένων χαρακτηριστικών έχει επιτύχει υψηλά αποτελέσματα σε πολλαπλά benchmarks για αυτό και γενικα θεωρούνται καθολικοί περιγραφιτές χαρακτηριστικών. Τα διανύσματα αυτά έχουν προκύψει από μη επιβλεπόμενα νευρωνικά γλωσσικά μοντέλα όπως το μοντέλο word2vec [18] ή το μοντέλο Glove [26] πάνω σε μεγάλα σύνολα ανεπισημείωτων δεδομένων. Οι Mikolov et al.[18] εκπαίδευσαν διανύσματα ενσωματωμένων χαρακτηριστικών ενός word2vec μοντέλου πάνω σε 100 εκατομμύρια λέξεων του συνόλου Google News. Τα προεκπαιδευμένα αυτά διανύσματα του μοντέλου Google embeddings έγιναν διαθέσιμα σ όλους ομοίως και το μοντέλο Twitter Glove. Επειδή τα δεδομένα από τα ειδησεογραφικά άρθρα και τα μηνύματα του Twitter γενικού σκοπού διαφέρουν σε κάποιο βαθμό από τα δεδομένα κρίσεων οι Imran et al. [9] έχουν εκπαιδεύσει ένα εξειδικευμένο μοντέλο word2vec το Crisis Embeddings πάνω σε 20 εκατομμύρια λέξεων ενός σώματος 57,908 μηνυμάτων Twitter που σχετίζονται με διαφορά καταστροφικά συμβάντα.

Μια καλή αρχικοποιησή της αναπαράστασης της εισόδου συντελεί ώστε το μοντέλο να συγκλίνει πιο γρήγορα αλλά και να αποφύγει τοπικά ελάχιστα της συνάρτησης κόστους. Ωστόσο η χρήση των εξειδικευμένων αυτών διανυσμάτων ενέχει και το κίνδυνό υπερπροσαρμογής στα δεδομένα εκπαίδευσης μετά από λίγες επαναλήψεις. Παράλληλα όμως θα μπορούσαμε να χρησιμοποιήσουμε και στατικά διανύσματα ενσωματωμένων χαρακτηριστικών δηλαδή να παγιώσουμε τις παραμέτρους του νευρώνων απεικόνισης λεξιλογίου και να εκπαιδεύσουμε το υπόλοιπο δίκτυο. Η μέθοδος που χρησιμοποιούμε ουσιαστικά συνδυάζει τις δύο μεθόδους. Μιμούμενοι αρχιτεκτονικές δικτύων που χουν χρησιμοποιηθεί για ταξινόμηση εικόνων όπου η είσοδος διαθέτει πολλά κανάλια (π.χ RGB) έτσι και ο πίνακας πρότασης $\mathbf{X}$ διαθέτει πολλαπλά κανάλια ένα για τα μη στατικά διανύσματα ενσωματωμένων χαρακτηριστικών και ένα με τα στατικά διανύσματα ενσωματωμένων χαρακτηριστικών. Στο δίκτυο που υλοποιήσαμε χρησιμοποιούμε αύτη τη μέθοδο βελτίωσης της γενίκευσης του μοντέλου CNN.

5.5.3 Μέθοδος παράλειψης χαρακτηριστικών

Ο συνδυασμός των προβλέψεων ενός συνόλου διαφορετικών δικτύων μπόρει να περιορίσει σημαντίκα το φαινόμενο της υπερπροσαρμογής. Ωστόσο η εκπαίδευση πολλαπλών δικτύων είναι υπολογιστικά κοστοβόρα σε απαγορευτικά επίπεδα για μεγάλα δίκτυα. Αντ΄ αυτής θα προσπαθήσουμε εκπαιδεύοντας ένα δίκτυο να προσσεγγίσουμε τα αποτελέσματα αυτών των μεθόδων που συνδυάζουν προβλέψεις διαφορετικών δικτύων.

Θα χρησιμοποιήσουμε μία μέθοδο κανονικοποίησης (regularisation) που σε κάθε επανάληψη και για κάθε νέο πρότυπο εισόδου εκτελεί μία τυχαιοκρατική επιλογή νευρώνων που θα «παραλειφθούν» (dropout) [29] υπο την έννοια ότι οι ενεργοποιήσεις αυτών των νευρώνων δε

θα ληφθούν υπόψη ούτε κατά την εμπροσθοδιάδοση ούτε κατά την οπισθοδιάδοση του σφάλματος. Αυτό το πετυχαίνουμε με μία απλή διαδικασία κάλυψης (masking) θέτοντας μηδέν τις ενεργοποιήσεις ενός ποσοστού p (dropout rate) των νευρώνων του στρώματος που υπόκειται σε παράλειψη χαρακτηριστκών (dropout). Για παράδειγμα, αν εφαρμόζαμε παράλειψη χαρακτηριστικών στο προπαραλήγον επίπεδο του απλού συνελικτικού δικτύου θα τροποποιούσαμε την σχέση (5.4) σύμφωνα με την (5.10):

$$z = f(\mathbf{Z} \cdot (\hat{\mathbf{h}} \circ \mathbf{r}) + \mathbf{b_h}) \tag{5.10}$$

όπου $\circ$ σημαίνει στοιχείο προς στοιχείο γινόμενο και όπου r μία τυχαία μεταβλητή bernoulli με πιθανότητα p που είναι ο τανυστής κάλυψης των νευρώνων του Z. Οι βαθμίδες του σφάλματος υπολογίζονται μόνο δια των «ακάλυπτων» μονάδων του δικτύου.

Όταν είναι να χρησιμοποιήσουμε το μοντέλο για να προβλέψουμε την τάξη νέων και άγνωστων προτύπων κλιμακώνουμε τα διανύσματα βαρών εκπαίδευσης κατα p δηλαδή $\hat{w} = p \cdot w$ και με τα νέα αυτά βάρη και χωρίς dropout το δίκτυο υπολογίζει την πιθάνοτητα κάθε τάξης. Αυτό ισοδυναμεί με το να πάρουμε τον γεωμετρικό μέσο των προβλέψεων ενός εκθετικά μεγάλου πλήθους διαφορετικών δικτύων.

Συνεπώς, σε κάθε νέα συστάδα προτύπων εκπαίδευσης προσαρμόζεται σε ένα διαφορετικής αρχιτεκτονικής δίκτυο . Όλα αυτά τα δίκτυα είναι σημαντικό ότι διαμοιράζονται κοίνα βάρη. Αυτό μειώνει την υπερπροσαρμογή εφόσον το δίκτυο αποφεύγει περίπλοκες συπροσαρμογές των νευρώνων και αναγκάζεται να εκπαιδεύσει πιο αξιόπιστα χαρακτηριστικά που δεν εξαρτώνται από την παρουσία άλλων χαρακτηριστικών.

5.6 Προσαρμογή Πεδίου (Domain Adaptation)

5.6.1 Εισαγωγή

Η χρήση μοντέλων με πολλές συνδέσεις και παραμέτρους όπως τα βαθιά νευρωνικά δίκτυα απαιτεί ένα πολυ μεγάλο αριθμό επισημειωμένων προτύπων για την εκπαίδευση τους. Έν τούτοις βασικός περιορισμός για τις μεθόδους επιβλεπομένης μάθησης στην αυτόματη ταξινόμηση κειμένού είναι η έλλειψη επισημειωμένων παραδειγμάτων. Ειδικότερα μία περίπτωση που συχνά αντιμετωπίζουμε είναι να διαθέτουμε άρκετα επισημειωμένα πρότυπα από ένα πεδίο (πηγαίο πεδίο) πχ αρθρογραφία και ελάχιστα έως καθόλου από κάποιο άλλο συγγενικό πεδίο που μας ενδιαφέρει (πεδίο προορισμού) πχ επιστημονική βιβλιογραφία, η οποία περιλαμβάνει σε μεγάλο βαθμό ειδικό λεξιλόγιο. Πρόκειται για το γνωστό πρόβλημα της προσαρμογής καθώς οι κατανομές των δεδομένων μεταξύ των δύο πεδιών παρουσιάζει μία απόκλιση γνωστή ως συμμετάβλητη μετατόπιση (covariate shift).

Στην πράξη εμείς αντιμετωπίζουμε την εξής περίπτωση, έχουμε αρκετά επισημειωμένα δεδομένα από παλαιότερες κρίσεις αλλά ελάχιστα επισημειωμένα πρότυπα τις πρώτες ώρες της εξάρσεως ενός νέου φαινομένου και αυτό γιατί εξαρτώμαστε από το πλήθος και την γεωγραφική κατανομή των ψηφιακών εθελοντών που θα εκτελέσουν το έργο της επισημειώσης. Στην

ενότητα που ακολουθεί θα αναφερθούμε εν γένει στις μεθόδους προσαρμογής και συγκεκριμένα στη μέθοδο που χρησιμοποιήσαμε της επιλογής παραδειγμάτων (instance selection).

5.6.2 Ανάλυση του προβλήματος προσαρμογής

Για αρχή θα αναλύσουμε σε ποιό βαθμό και κατά ποιούς τρόπους επηρεάζει η προσαρμογή πεδίου την αποδοτικότητα των μοντέλων ταξινόμησης κειμένου. Ένα επιλεκτικό μοντέλο (discriminative model) επιτελεί το έργο της ταξινόμησης (πχ. κειμένου) μελετώντας την κατα συνθήκη κατανομή $p(y|x)$, επιλέγοντας από μια παραμετρική οικογένεια κατανομών $p(y|x;\theta)$ τις βέλτιστες παραμέτρους που μεγιστοποιούν την εκτιμώμενη λογαριθμική πιθανοφάνεια (5.11). Πιο τυπικά έστω $f : \mathcal{X} \to \mathcal{Y}$ ένας ταξινομητής δηλαδή μια απεικόνιση από ένα χώρο χαρακτηριστικών σε ένα χώρο επισημειώσεων. Θεωρούμε $[(x_i, y_i)]_{i=1}^{N}$ επισημειώμενα πρότυπα εκπαίδευσης της κατά τα άλλα άγνωστης από κοινού κατανομής πιθανότητας $p(x, y)$.

$$\theta^* = arg\max_{\theta} \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x;\theta) \quad dx \tag{5.11}$$

Μη γνωρίζοντας την κατανομή $p(x, y)$ προσπαθούμε να μεγιστοποιήσουμε την εμπειρική λογαριθμική πιθανοφάνεια (5.12). Αυτό δίνει αξιόπιστα αποτελέσματα υπό την προυπόθεση ότι έχουμε αρκετά επισημειωμένα δεδομένα.

$$\begin{aligned} \theta^* &= arg\max_{\theta} \int_{X} \sum_{y \in Y} \widetilde{p}(x, y) \log p(y|x;\theta) \quad dx \\ &= \frac{1}{N} \sum_{i=1}^{N} \quad \log p(y_i|x_i;\theta) \end{aligned} \tag{5.12}$$

Θεωρητικά το πρόβλημα της προσαρμογής έγγειται στην εκτίμηση της από κοινού κατανομής $p_t(x, y)$ των δειγμάτων του πεδίου προορισμού χρησιμοποιώντας την πληροφορία της γνωστής κατανομής $p_s(x, y)$ αξιοποιώντας το πλήθος των επισημειωμένων προτύπων που διαθέτουμε στο πηγαίο πεδίο. Αν παραγοντοποιήσουμε την κατανομή των δεδομένων εκπαίδευσης σύμφωνα με τη γνωστή σχέση (5.13) συμπεραίνουμε ότι συντρέχουν δύο παράγοντες για την απόκλιση των κατανομών $p_s(x, y)$ και $p_t(x, y)$.

$$p(x, y) = p(y|x) \cdot p(x) \tag{5.13}$$

Γι αυτό και διακρίνουμε εν γένει δύο μεθόδους προσαρμογής πεδίου, την προσαρμογή επισημειώσεων (labeling adaptation) και την προσαρμογή παραδειγμάτων (instance adaptation) [10]. Στην πρώτη περίπτωση έχουμε σημαντική μεταβολή μεταξύ $p_s(y|x)$ και $p_t(y|x)$ το οποίο σημαίνει ότι η χρήση της $p_s(y|x)$ για την ταξινόμηση των παραδειγμάτων του πεδίου προορισμού δεν δίνει ικανοποιητικά αποτελέσματα επειδή αδυνατεί να εκτιμήσει με ακρίβεια την $p_t(y|x)$. Στην άλλη περίπτωση η απόκλιση μετάξυ $p_s(x)$ και $p_t(x)$ ενώ φαινομενικά δεν επηρεάζει τη μεταβολή μεταξύ $p_s(y|x)$ και $p_t(y|x)$, χρήζει πάλι προσαρμογής πεδίου αφού η εκτίμηση

της $p_s(y|x)$ συνδέται με τη εμπειρική κατανομή $\tilde{p}_s(x, y)$ που την επηρεάζει η απόκλιση $p_s(x)$ και $p_t(x)$. Παρακάτω θα αναπτύξουμε μια ειδική περίπτωση της προσαρμογής επισημειώσεων γνωστή ως κλάδεμα ή επιλογή παραδειγμάτων.

5.6.3 Κλάδεμα Παραδειγμάτων (Instance Pruning)

Οι τεχνικές προσαρμογής παραδειγμάτων είναι αναγκαίες όταν τα ανεπισημειώτα πρότυπα που θέλουμε να ταξινομηθούν ανήκουν σε διαφορετική κατανομή από τα επισημειωμένα πρότυπα που διαθέτουμε. Έστω ότι συμβολίζουμε με $p_s(x, y)$ την υποκείμενη κατανομή των επισημειωμένων προτύπων που αφορούν το πηγαίο πεδίο και $p_t(x, y)$ την αντίστοιχη κατανομή των προτύπων του πεδίου προορισμού. Όταν υπάρχει σημαντική απόκλιση μεταξύ $p_s(y|x)$ και $p_t(y|x)$ τότε οι προβλέψεις μας για τα πρότυπα του πεδίου προορισμού δεν είναι καθόλου ικανοποιητικές. Αν γνωρίζουμε τα συγκεκριμένα παραδείγματα για τα οποία οι προβλέψεις βάσει $p_t(y|x)$ διαφέρουν από κείνες βάσει $p_s(y|x)$ μπορούμε να διαγράψουμε από το σύνολο των δεδομένων εκπαίδευσης τα «παραπλανητικά» αυτά παραδείγματα και να εκπαιδεύσουμε στο νέο «προσαρμοσμένο» σύνθετο σύνολο. Αυτή η απλή ιδέα είναι δοκιμασμένη και δίνει καλά αποτελέσματα. Πιο σύνθετα σχήματα «κλαδέματος» που χρησιμοποιούν μη-επιβλεπόμενες μεθόδους μάθησης μπόρουν να δώσουν ακόμα καλύτερες επιδόσεις ωστόσο η μελέτη των μεθόδων αυτών ξεφεύγει των ορίων του παρόντος έργου.

Κεφάλαιο 6

Πειραματική Αξιολόγηση

6.1 Εισαγωγή

Σ αυτή την ενότητα θα παρουσιαστούν και θα σχολιαστούν τα αποτελέσματα της μελέτης. Θα αναφερθούμε σύντομα στο πληθυσμό που μελετήσαμε και συνδυάζοντας τις ερμηνείες των διάφορων μετρικών θα συναγάγουμε γενικότερα συμπεράσματα για την εφαρμοσιμότητα και τα περιθώρια βελτιώσης των προτεινόμενων αλγορίθμων. Προσπαθήσαμε να είμαστε όσο το δυνατόν ποιό αμερόληπτοι χρησιμοποιώντας πάντα ξεχωριστά σύνολα δοκιμής κατα την εκπαίδευση και βελτιστοποίηση των διάφορων παραμέτρων του δικτύου.

6.2 Παράμετροι αξιολόγησης

Η πρώτη παράμετρος που θα εξετάσουμε είναι η ανάκληση ή ευαισθησία και είναι το στατιστικό που μετρά το ποσοστό των αληθώς θετικών περιπτώσεων που ταξινομήθηκαν ορθώς ως θετικές. Εξετάζει δηλαδή κατα πόσο το μοντέλο μπορεί να επιλέξει θετικά δείγματα του πληθυσμού ή σε τι βαθμό ο κάνονας ταξινόμησης «καλύπτει» τα αληθώς θετικά . Ωστόσο δε μπορεί να μας πει τίποτα για τα δείγματα που το μοντέλο θα προβλέψει αρνητικά. Συνηθώς αυτή η μετρική αμελείται ή μεσοτιμείται γιατί δε μας ενδιαφέρει ποιό υποσύνολο θετικών δειγμάτων θα επιλεχθή από το μοντέλο, υποθέτοντας σιώπηρά ότι υπάρχουν πολλά θετικά στο πληθυσμό. Η ανάκληση είναι το ένα βασικό σκέλος των καμπυλών συμβιβασμού (PR) και ROC που θα εξετάσουμε αργότερα. Για να δώσουμε έναν πιο τυπικό ορισμό θεωρούμε, για διάφορες τιμές μιας πιθανότητας διαχωριστικού κατωφλίου T, ότι η τυχαία μεταβλητή X ακολουθεί τη κατανομή f_1 αν $X > T$ οπότε το δείγμα ταξινομείται ως θετικό αλλιώς ακολουθεί την f_0 οπότε το δείγμα ταξινομείται ως αρνητικό.

Ανάκληση ή ευαισθησία τυπικά ορίζεται το ποσοστό των περιπτώσεων που ορθώς προ-

βλέφθηκαν θετικές απ το μοντέλο προς το σύνολο των θετικών παραδειγμάτων .

$$\begin{aligned} recall = sensitivity = tpr &= \frac{TP}{TP+FN} \\ &= \int_T^\infty f_1(x)dx \\ &= P(y=1|\hat{y}=1) \end{aligned} \tag{6.1}$$

Δυϊκά μπορεί να οριστεί το στατιστικό της ακρίβειας ή εμπιστοσύνης ή αληθώς θετικής ορθότητας ως το ποσοστό των περιπτώσεων που προβλέφθηκαν ως θετικές και πράγματι αντιστοιχούν σε θετικές περιπτώσεις:

Ακρίβεια ή εμπιστοσύνη, τυπικά ορίζεται το ποσοστό των περιπτώσεων που ορθώς προβλέφθηκαν θετικές απ το μοντέλο προς το σύνολο των θετικών προβλέψεων.

$$\begin{aligned} precision = confidence = ppv &= \frac{TP}{TP+FP} \\ &= P(\hat{y}=1|y=1) \end{aligned} \tag{6.2}$$

Τόσο η ανάκληση όσο και η ακρίβεια έχουν να κάνουν με τα θετικά πρότυπα και δεν ασχολούνται με τη πληροφορία που μπορεί να μας δώσει η ταξινόμηση των αρνητικών πρότυπων. Ομοίως και οι διάφοροι αριθμητικοί, γεωμέτρικοι και αρμονικοί μέσοι αυτών δε λαμβάνουν υπόψη το ποσοστό των ορθώς αρνήτικών προβλέψεων . Ο πλέον ευρέως χρησιμοποιούμενος μέσος είναι ο αρμονικός $F = \frac{G^2}{A}$ που εκφράζει το ποσοστό ειδικής συμφωνίας ως προς τη θετική κλάση . Ο f_1, εξισορροπημένος αρμονικός μέσος ακρίβειας και ανάκλησης ή δείκτης Sørensen συγκεκριμένα αξιολογεί την αποδοτικότητα του ταξινομητή αποδίδοντας ίση βαρύτητα σε ανάκληση και ακρίβεια:

$$\begin{aligned} f_1, DSC &= \frac{TP}{TP + \frac{FP+FN}{2}} \\ &= 2 * \frac{Precision * Recall}{Precision + Recall} \end{aligned} \tag{6.3}$$

Φυσικά η κλάση των θετικών παραδειγμάτων (αυτών που ανήκουν στην κατηγοριά περιεχομένου που μας ενδιαφέρει) δεν έχει τίποτα το ιδιαίτερο και μπορούμε να ορίσουμε και τις αντίστροφες μετρικές των παραπάνω εναλλάσωντας θετικά με αρνητικά παραδείγματα και ταξινομώντας αντίθετα. Μπορούμε να ορίσουμε την αντίστροφη ανάκληση ή ειδικότητα specificity ως το ποσοστό των αρνητικών περιπτώσεων που ορθώς προβλέφθηκαν αρνητικές και γι αυτό καλλείται ποσοστό αληθώς αρνητικών προτύπων (tnr) . Ακόμη δυϊκά ορίζεται η αντίστροφη ακρίβεια ως το ποσοστό των αρνητικών προβλέψεων που πράγματι αντιστοιχούν σε αρνητικά παραδείγματα του πληθυσμού και καλείται πολλές φορές αληθώς αρνητική ορθότητα.

Αντίθετα με τις παραπάνω μετρικές η ορθότητα rand accuracy λάμβάνει ξεκάθαρα υπόψη της τη ταξινόμηση των αρνητικών παραδειγμάτων. Μπορεί να εκφρασθεί σαν σταθμισμένος μέσος ανάκλησης και αντίστροφης ανάκλησης ή ακρίβειας και αντίστροφης ακρίβειας. Τώρα, ορθότητα accuracy, ορίζεται ως ο λόγος των σωστών προβλέψεων , θετικών και αρνητικών, προς το σύνολο των παραδειγμάτων.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6.4}$$

Οι παραπάνω μετρικές μόλο που είναι καθιερωμένες στα πεδία της μηχανικής μάθησης και της ανάκτησης πληροφοριάς παραμένουν προκατειλημμένες και επηρεάζονται από την σχετική ασσυμετρία των κλάσεων. Κι αυτό γιατί είτε αδυνατούν να διαχειριστούν τη σωστή ταξινόμηση των αρνητικών δειγμάτων είτε την απόδοση των προβλέψεων που στηρίζονται αποκλειστικά στη τύχη. Από την άλλη, η ανάλυση σφάλματος που στηρίζεται στις καμπύλες χαρακτηριστικής λειτουργίας , και χρησιμοποιείται ευρέως στις επιστήμες υγίειας, θεωρείται η πλέον καθιερωμένη μέθοδος επαλήθευσης συγκρίνοντας τον ρυθμό με τον οποίο το μοντέλο αποδέχεται τα αληθώς θετικά με το ρυθμό απόρριψης των αρνητικών δειγμάτων.

Ένας ταξινομητής θεωρείται βέλτιστος στο σημείο (0,1) όπου $tpr = 1$ και $fpr = 0$ και χείριστος στο σημείο (0,0). Όταν $tpr = fpr$ δηλαδή πάνω στη θετική διαγώνιο θεωρείται πως οι προβλέψεις είναι ισοδύναμες με ρίψεις νομίσματος αφού στηρίζονται αποκλειστικά στη τύχη. Πάνω από τη θετική διαγώνιο ο ταξινομητής αρχίζει και προβλέπει πληροφορημένα. Μ αυτό το γεωμετρικό τρόπο μπορούμε να συγκρίνουμε τις επιδόσεις των ταξινομητών μας επειδή είναι όμως δύσκολο να συγκρίνουμε τις διάφορες καμπύλες γεωτρικά , διότι πολύ πληροφορία περιέχουν, τις συγκρίνουμε βάσει του εμβαδού που περικλείουν. Έτσι έχουμε μια απλή μετρική που δεν είναι προκατειλημμένη και εκφράζει σε κάποιο βαθμό την πληροφορισιμότητα των προβλέψεων του ταξινομητή μας. Παρακάτω δίνουμε σύντομα έναν πιο αυστηρό ορισμό του εμβάδου της καμπύλης χαρακτηριστικής λειτουργίας για τις διάφορες τιμές κατωφλίου ταξινόμησης στη περίπτωση της δυαδικής ταξινόμησης που εύκολα μπορούμε να γενικεύσουμε στη περίπτωση της πολυταξικής ταξινόμησης.

Η AUC-ROC(Area Under Curve), είναι το εμβαδόν κάτω απ΄ τη χαρακτηριστική καμπύλη λειτουργίας.

$$\begin{aligned} AUROC &= \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T)f_1(T')f_0(T)dT'dT \end{aligned} \tag{6.5}$$

Με τη καμπύλη χαρακτηριστικής λειτουργίας ROC (Receiver Operating Characteristic Curve μετράμε τη διαγνωστική ικανότητα ενός δυαδικού ταξινομητή καθώς μεταβάλουμε τη τιμή του διαχωριστικού κατώφλιου T.

Μια εναλλακτική επιλογή από την καμπύλη χαρακτηριστικής λειτουργίας είναι η καμπύλη της ακρίβειας συναρτήσει της ευαισθησιάς από το χώρο της ανάκτησης πληροφορίας όπου χρησιμοποιείται κυρίως για την αξιολόγηση συστημάτων ιεραχημένης αναζήτησης (ranked search). Αναφέρεται συχνά ότι μπορεί να αντικαταστήσει την ανάλυση βάσει καμπυλών χαρακτηριστικής λειτουργίας με μεγαλύτερη επιτυχία όταν τα δεδομένα παρουσιάζουν υψηλή ασσυμετρία. Μας επιτρέπει να εντοπίσουμε περιθώρια βελτίωσης που είναι δεν είναι προφανή στο χώρο χαρακτηριστικής λειτουργίας όταν η ασσυμετρία των κλάσεων είναι έντονη. Σ αντίθεση με τις καμπύλες ROC που είναι βέλτιστες στην άνω αριστερή πλευρά του τεταρτημορίου οι καμπύλες PR είναι βέλτιστες στην κάτω δεξιά πλευρά. Και πάλι χάριν απλοποίησης βαθμολογούμε βάσει εμβαδού AU-PR με ορισμό παρόμοιο με αυτό που χρησιμοποιήσαμε ανωτέρω για την AU-ROC.

6.3 Οργάνωση πειραμάτων

Η οργάνωση των πειραμάτων γίνεται σε δύο άξονες. Ο πρώτος είναι δεδομενο-κεντρικός και εξετάζει για τις διάφορες συλλογές τις επιδόσεις των ταξινομητών εν γένει. Για να έχουμε μια εικόνα του δείγματος χρησιμοποιούμε ραβδογράμματα στοίβας όπου βλέπουμε για τα διάφορα σύνολα την εκπροσώπηση των κλάσεων στο πληθυσμό. Απ αυτά τα διαγράμματα αντλούμε πληροφορίες για την επικράτηση ή την ασσυμετρία που εμφανίζουν συγκεκριμένες κλάσεις. Στη συνέχεια για κάθε συμβάν εξετάζουμε την εκπροσώπηση των κλάσεων του πληθυσμού μετά την προσθήκη δεδομένων από παλαιότερα συμβάντα και την επίδραση που έχει τόσο η αύξηση του συνόλου εκπαίδευσης όσο και η εξισσορόπηση των κλάσεων στα συνθετικά αυτά σύνολα. Επειδή οι συλλογές μας αφορούν κυρίως σεισμούς και τυφώνες θα επικεντρωθούμε στις συγκεκριμένες συλλογές και δεν θα συνδυάσουμε συλλογές από άλλα συμβάντα προς περιορισμό της συμμετάβλητης μετατόπισης στις κατανομές των συνθετικών συλλογών.

Ύστερα θα στραφούμε σε μια κατεύθυνση μοντελο-κεντρική. Επειδή το προβλημά μας οφείλει να επιτελέσει το έργο της ταξινόμησης σε κρίσιμα σύντομο χρονικό διάστημα στη διάρκεια που μόλις ξεσπά μία κρίση και στο διάστημα λίγο μετά όταν ο αντίκτυπος της καταστροφής είναι σφοδρός και οι αναφυόμενες ανάγκες επείγουσες θα περιοριστούμε σε αρχιτεκτονικές συνελικτικών δικτύων επειδή είναι οι επιδόσεις τους πολλά υποσχόμενες και ο χρόνος εκπαίδευσης τους μικρός. Απ αυτές δύο βασικές αρχιτεκτονικές θα μας απασχολήσουν αυτές θα υλοποιήσουμε και θα προσαρμόσουμε στις ειδικές ανάγκες των δεδομένων κρίσεων. Η πρώτη που φέρει τον τίτλο CNN και είναι βασισμένη στο υπόδειγμα του Yoon Kim [5] και η άλλη με το τίτλο DCNN και είναι βασισμένη στο υπόδειγμα του Kalchbrener [11]. Τέλος μια μηχανή διανυσματικής στήριξης γραμμικού πυρήνα [32] χρησιμοποιείται ως βασικός μη νευρωνικός ταξινομητής. Θα εξετάσουμε τα πλεονεκτήματα των νευρωνικών μεθόδων ταξινόμησης συγκρίνοντας τις επιδόσεις τους με το βασικό ταξινομητή άλλα και μεταξύ τους για τα διάφορα σύνολα πραγματικά και συνθετικά.

Επειδή τα νευρωνικά δίκτυα δεν είναι καθέ άλλο πάρα έτοιμοι αλγόριθμοι ταξινόμησης χρειάζεται αρκετός πειραματισμός των διάφορων παραμέτρων τους. Αυτό έχει να κάνει τόσο με την ίδια την αρχιτεκτονική τους όπως το πλήθος και το είδος των κρυφών στρωμάτων και μονάδων όσο και με τον χρησιμοποιούμενο αλγόριθμο μάθησης και βελτιστοποίησης. Επιπλέον πρέπει να πειραματιστούμε με τις διάφορες μεθόδους περιορισμού του φαινομένου της υπερπροσαρμογής ώστε να έχουμε ικάνη γενικευσιμότητα στα δεδομένα επαλήθευσης με τα οποία δε θα έχουν εκπαιδευτεί. Παρακάτω δίνουμε πίνακες με τις βελτιοποιήσιμες παραμέτρους όπως ο ρυθμός μάθησης και το μέγεθος μικρο-συστάδας, το εύρος και το πλήθος των συνελικτικών φίλτρων, το πλήθος των κρυφών νευρώνων του πλήρως συνδεδεμένου στρώματος κ.α.

Η ανάγκη να ενσωματώσουμε δεδομένα από άλλα συμβάντα εξετάζει τις δυνατότητες μεταφοράς μάθησης (transfer learning) χρησιμοποιώντας ιστορικά δεδομένα άλλων κρίσεων. Αυτό είναι απαραίτητο γιατί σε πραγματικές συνθήκες έχουμε λίγα επισημειωμένα δεδομένα στην αρχή της κρίσης και είναι και το πιο κρίσιμο διάστημα όπου δεν έχουμε έτοιμες επισημειώσεις

Πίνακας 6.1: Κατανομή κατηγοριών ταξινόμησης για τα συμβάντα προς μελέτη

Τάξη	**Σεισμός Νεπάλ**	**Τυφώνας Χάρβεϊ**	**Σεισμός Καλιφόρνια**	**Τυφώνας Ίρμα**	**All Others**
Affected Individuals	756	106	227	106	4624
Donations and Volunteering	1021	1058	830	389	1752
Infrastructure and Utilities	351	326	351	440	1972
Sympathy and Support	983	60	83	62	4546
Other Useful Information	1505	1280	1028	2126	7709
Not related or irrelevant	6698	140	157	190	418
Grand Total	11314	2970	1929	3754	21021

από τους ψηφιακούς εθελοντές. Αυτό το κενό πληροφορίας που αντιμετωπίζουμε στην αρχή της κρίσης καθιστά δύσκολη την εκπαίδευση των νέων μοντέλων. Τα αποτελέσματα είναι ασθενέστερα αλλά εκτιμούν την ικανότητα γενικευσής των μοντέλων ταξινόμησης σε συνθήκες έλλειψης σχετικών παραδειγμάτων με το συμβάν.

Για να εκπαιδεύσουμε και να αξιολογήσουμε την απόδοση των διάφορων ταξινομητών κειμένου θα διακρίνουμε τέσσερα καταστροφικά φαινόμενα διαμερίζοντας κατάλληλα το σύνολο DeepCrisis· και τα τέσσερα είναι φυσικές καταστροφές μιας και το ενδιαφέρον μας επικεντρώνεται στην αντιμετώπιση κρίσεων μεγάλης κλίμακας που έχουν ηχηρό αντίκτυπο στο κοινωνικό σύνολο. Κάθε σύνολο εμφανίζει μερική ασσυμμετρία κλάσεων, το οποίο γενικά δεν είναι επιθυμητό και οφείλεται σε διάφορους παράγοντες όπως η φύση της απειλής αν δηλαδή είναι γεωλογική (σεισμός), μετεωρολογική (τυφώνας,κυκλώνας), υδρολογική (πλημμύρες, κατολισθήσεις)· άλλοι παράγοντες είναι η χρονική της διάσταση αν δηλαδή είναι στιγμιαία ή εξελισσόμενη ή η γεωγραφική της διασπορά αν είναι εντοπισμένη ή εκτεταμμένη. Ακόμη κάθε σύνολο έχει μια ξεχωριστή εγγενή δυσκολία ταξινόμησης που μετράται με ένα δείκτη διεπισημειωτικής ακρίβειας (inter-annotator accuracy). Το πιο υψηλό IAA έχει το σύνολο California Earthquake με ακρίβεια 0.85 ενώ τη χαμηλότερη 0.70 το Typhoon Hagupit . Τα υπόλοιπα έχουν ακρίβεια περίπου 0.75. Αύτα είναι και τα επίπεδα ακρίβειας που επιθυμούμε να πετύχουμε [21].

Τώρα όσον αφορά την εκπαίδευση διαιρούμε κάθε συλλογή σε υποσυλλογές δεδομένων εκπαίδευσης 70%, δοκιμής 20% και επαλήθευσης 10%. Όταν εκπαιδεύουμε τους ταξινομητές για κάποιο συμβάν τα δεδομένα εκπαίδευσης όλων των υπόλοιπων συλλογών μπορούν να χρησιμοποιηθούν ως πηγαία δεδομένα (source data) που διαμορφώνουν τις out συλλογές. Χρησιμοποιώντας έναν αλγόριθμο διαστρωματικής επαλήθευσης (stratified validation) φροντίζουμε να εξασφαλίζεται μια ισορροπημένη κατανομή όλων των κλάσεων στις υποσυλλογές. Για κάθε σειρά πειραμάτων βελτιστοποιούμε τις υπερ-παραμέτρους των ταξινομητών (hyper-parameters) ώστε να έχουμε βέλτιστη απόδοση για τα διάφορα benchmarks. Η βελτίωση των υπερ-παραμέτρων γίνεται στο development set ώστε να είναι αμερόληπτη. Το σύνολο των βελτιοποιήσιμων παραμέτρων φαίνεται συγκεντρωτικά στους πίνακες πίνακα (6.2) και (6.3). Με έντονη γραμματοσειρά παρουσιάζονται οι τιμές των παραμέτρων που επιτυγχάνουν τη υψηλότερη μέση ορθότητα. Η επιλογή των βέλτιστων υπερ-παραμέτρων γίνεται εν μέρει με εξαντλητική αναζήτηση(grid search) και εν μέρει με ευριστική μέθοδο αναζήτησης (για το ρυθμό μάθησης και το μέγεθος μικροσυστάδας) στα διάφορα benchmarks.

Drop out ratio (ρυθμός παράλειψης)	0.0, 0.2, 0.4, **0.5**
Mini_batch sizes (μέγεθος μικροσυστάδων)	32, 64, **128**
P% (Μέγεθος Λεξιλογίου)	80, 85, **90**
Αριθμός Φίλτρων	100, 125, **150**
Μέγεθος παραθύρου	2, 3, **4**
Pooling Length (εύρος συνδυασμού)	2, 3, **4**
Μονάδες κρυφού στρώματος	100, 128, **150**

Πίνακας 6.2: Παράμετροι πειραμάτων CNN

Drop out ratio (ρυθμός παράλειψης)	0.0, 0.2, **0.5**
Mini_batch sizes (μέγεθος μικροσυστάδων)	32, 64, **128**
P% (Μέγεθος Λεξιλογίου)	80, 85, **90**
Αριθμός Φίλτρων πρώτου συνελικτικού επιπέδου	6, 10, **12**
Αριθμός Φίλτρων δεύτερου συνελικτικού επιπέδου	12, 20, **28**
Μέγεθος παραθύρου	2, 3, **4**
Ρυθμός μάθησης	0.001, 0.02, **0.1**

Πίνακας 6.3: Παράμετροι πειραμάτων DCNN

6.4 Αποτελέσματα της μελέτης

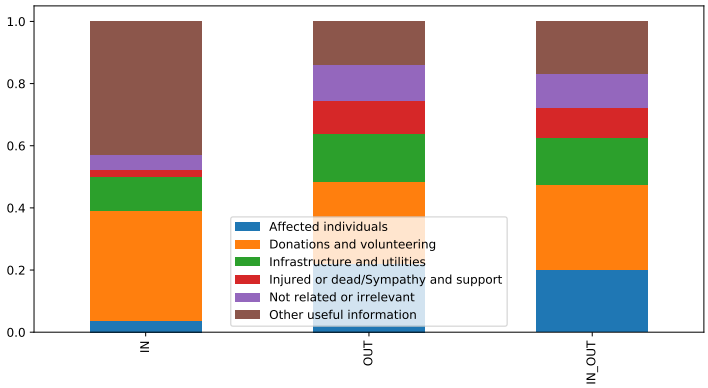
6.4.1 Εισαγωγή

Σ αυτή την ενότητα παρουσιάζουμε τα ποσοτικά και ποιοτικά αποτελέσματα της μελέτης. Θα σχολιάσουμε συνολικά και για κάθε συμβάν ξεχωριστά τις επιδόσεις των διαφόρων αλγορίθμων ποσοτικά βάσει των μετρικών που αναλύσαμε στην παραπάνω υποενότητα. Επιπλέον θα προσπαθήσουμε να αντλήσουμε εποπτικά συμπεράσματα από τις καμπύλες συμβιβασμού και τα ραβδογράμματα κατανομής των κλάσεων.

6.4.2 Παρουσίαση αποτελεσμάτων ανά συμβάν

Στο σχήμα 6.1 παρουσιάζεται η κατανομή κλάσεων για τα δεδομένα που αφορούν τον τυφώνα Χάρβεϊ. Ο τυφώνας Χάρβεϊ σύμφωνα με τη Βικιπαίδεια ήταν τροπικός κυκλώνας κατηγορίας 4 όταν χτύπησε το Τέξας Ηνωμένων Πολιτειών στις 25 Αυγούστου του 2017. Προκάλεσε καταστροφές ύψους 200 δις δολαρίων ιστορικό ρεκόρ για τα δεδομένα των Ηνωμένων Πολιτειών. Μόνο ο τυφώνας Κατρίνα θα μπορούσε να συγκριθεί μαζι του στον ολέθριο αντίκτυπο που είχε. Η συλλογή των δεδομένων ξεκίνησε την 25 Αυγούστου και ολοκληρώθηκε τη 5 Σεπτεμβρίου κατά την οποία συλλέχθηκαν περίπου 7 εκατομμύρια μηνύματα. Απ αύτα επισημειωμένα βρήκαμε περίπου τρεις χιλιάδες. Παρατηρούμε ότι οι κλάσεις που κυριαρχούν είναι μηνύματα που αφορούν προσφορά ανθρωπιστικής βοήθειας και μηνύματα που εντάσσονται στην catch-all κατηγορία μηνυμάτων άλλης χρησιμης πληροφορίας. Κατά δεύτερο λόγο

σημαντική είναι η εκπροσώπηση της κλάσεως που αφορά υλικές καταστροφές σε υποδομές και κτίρια. Μέτα την ενσωμάτωση μηνυμάτων από άλλους τυφώνες παρατηρούμε μείωση της ασσυμετρίας και καλύτερη εκπροσώπηση και των άλλων κλάσεων χρήσιμης πληροφορίας.

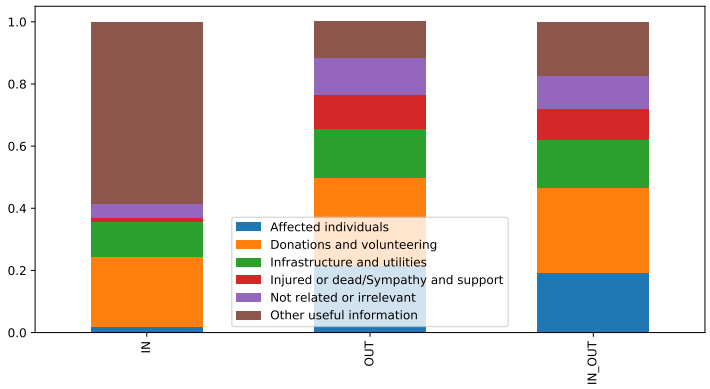


Σχήμα 6.1: Κατανομή κλάσεων για το σύνολο Harvey Hurricane

Την καλύτερη απόδοση έχουν οι νευρωνικοί ταξινομητές. Ο αλγόριθμος CNN με μικρή διαφορά φαίνεται να υπερέχει στο σύνθετο σύνολο ώστοσο στις άλλες δύο διατάξεις δεδομένων ο DCNN έχει το προβάδισμα. Και οι δύο έχουν επιδόσεις κόντα στο βαθμό διεπισημειωτικής ακρίβειας του συνόλου. Παρόλο που η ασσυμετρία των κλάσεων μειώνεται αισθητά στα σύνθετα σύνολα και έχουμε ένα αρκετά μεγαλύτερο πληθυσμό εκπαίδευσης η συμμετάβλητη μετατόπιση λόγω διαφορετικών κατανομων των διαφόρων συλλογών που συνθέτουμε αντισταθμίζουν τα πλεονεκτήματα. Η χαμηλή επίδοση της out συλλογής μας δίνει μια εκτίμηση του επιπέδου απόδοσης του ταξινόμητη στην αρχή του φαινομένου όπου δε διαθέτουμε έτοιμες επισημειώσεις σχετικές με το συμβάν. Από τις καμπύλες συμβιβασμού φαίνεται ότι τη μεγαλύτερη δυσκολία ταξινόμησης παρουσιάζουν η catch-all κλάση των μη σχετικών με το συμβάν μυνημάτων και η κλάση των μηνυμάτων που αφορούν τις επιπτώσεις του τυφώνα στους πληγέντες. Αυτό δικιολογείται από το γεγονός ότι είναι οι δύο κλάσεις που έχουν την χαμηλότερη εκπροσώπηση στο αρχικό σύνολο. Αυτό που φαίνεται και από τους αντίστοιχους πίνακες σύγχυσης είναι ότι οι ταξινομητες είναι προκατειλημμένοι από τις κλάσεις που έχουν μεγαλύτερη επικράτηση (prevalence) στο πληθυσμό.

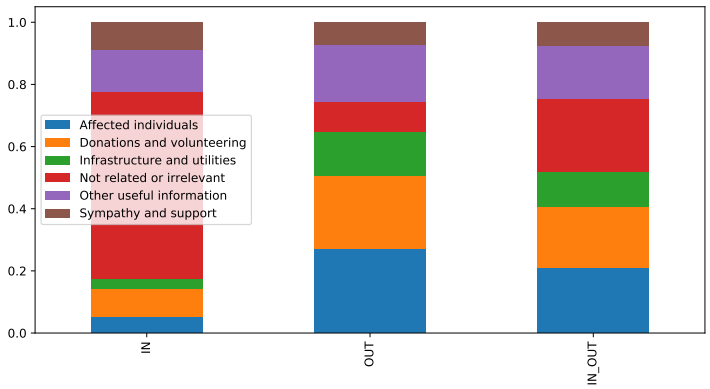
Ακολούθως έχουμε την ανάλυση για τον τυφώνα Ίρμα. Στο σχήμα 6.2 φαίνονται οι διαφορες κατανομές κλάσεων. Ο τυφώνας Ίρμα ήταν τροπικός κυκλώνας κατηγορίας 5 που έπληξε μεταξύ άλλων την Μπαρμπούντα, τον Άγιο Βαρθολομαίο, τον Άγιο Μαρτίνο, την Αγκουίλη και τις Παρθένες Νήσους αφήνοντας στο περασμά του καταστροφές ύψους 66.77 δις δολαρίων. Ο κυβερνήτης της Φλώριντα αναγκάστηκε να προβεί σε μέτρα ασφαλείας όπως κλείσιμο σχολείων και πανεπιστημιακών ιδρυμμάτων για την ασφάλεια των πολιτών. Η συλλογή των δεδομένων ξεκίνησε την 6 Σεπτεμβρίου του 2017 και ολοκληρώθηκε την 19 Σεπτεμβρίου όπου και συλλέχθηκαν περίπου 3.5 εκατομμύρια μηνύματα. Απ αυτά χρησιμοποιήσαμε περίπου 3.4K επισημειωμένα μηνύματα. Την καλύτερη απόδοση έχουν και εδώ οι νευρωνικόι αλγόριθμοι. Το δίκτυο DCNN έχει σαφές προβάδισμα στο αρχικό σύνολο υστέρει όμως λίγο έναντι του

CNN στό σύνθετο. Από τις καμπύλες συμβιβασμού παρατηρούμε ότι οι κατηγορίες των μη σχετικών μηνυμάτων καθώς και των επιπτώσεων στους πληγέντες παρουσιάζουν τη μεγαλύτερη δυσκολία ταξινόμησης. Η λογική και πάλι είναι παρόμοια με όσα παρατηρήσαμε στην περίπτωση του Χάρβεϊ όπου η αρχική ασσυμετρία υποβαθμίζει την απόδοση για τις κλάσεις μειοψηφίας.



Σχήμα 6.2: Κατανομή κλάσεων για το σύνολο Irma Hurricane

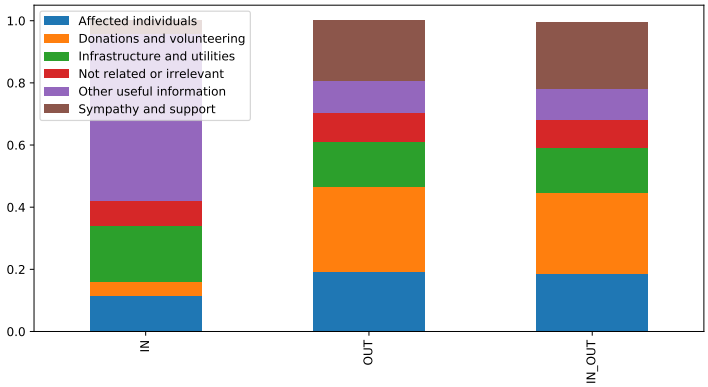
Ο σεισμός στο Νεπάλ ή σεισμός Γκόρκχα συγκλόνισε την υφήλιο το 2015 με περίπου 9K νεκρούς και 22K τραυματισμένους. Με επίκεντρο ανατολικά της Γκόρκχα και ένταση 7.8 M_w προκάλεσε ζημιές 10 δις δολαρίων (περίπου 50% του Νεπαλέζικου ΑΕΠ). Η συλλογή των δεδομένων ξεκίνησε τη 25 Απριλίου και ολοκληρώθηκε τη 19 Μαίου όπου και συνελέχθησαν περίπου 4.2M μηνύματα εκ των οποίων 11K επισημειωμένα. Στο σχήμα 6.3 που ακολουθεί παρατηρούμε την σχετική ασσυμετρία των κλάσεων με αξιοσημείωτη την επικράτηση των μη σχετικών με το συμβάν μηνυμάτων.



Σχήμα 6.3: Κατανομή κλάσεων για το σύνολο Nepal Earthquake

Παρατηρούμε ότι το δίκτυο DCNN επικρατεί σημάντικά των άλλων μεθόδων στο σύνθετο σύνολο ενώ οι επιδόσεις του είναι συγκρίσιμες των άλλων στις άλλες δύο διατάξεις. Από τις καμπύλες συμβιβασμού βλέπουμε ότι η συμπεριφορά του είναι υποδειγματική ως προς όλες

τις κλάσεις. Αυτό πιθανώς οφείλεται στο γεγονός ότι είναι το σύνολο με το μεγαλύτερο πληθυσμό, εως και πέντε φορές συγκρίσει με τα άλλα σύνολα δεδομένων.



Σχήμα 6.4: Κατανομή κλάσεων για το σύνολο California Earthquake

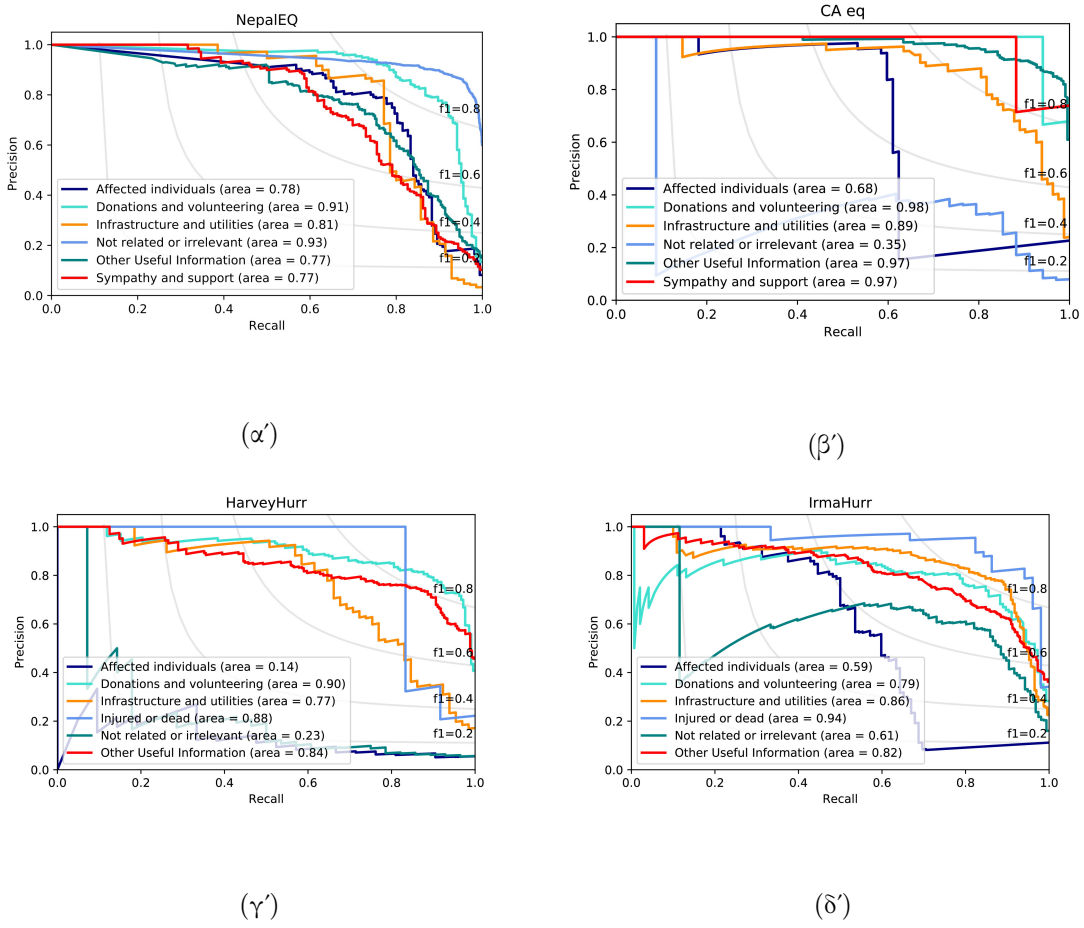
Τέλος ο σεισμός στη νοτία Νάπα έπληξε την περιοχή του Σαν Φραντσίσκο τον Αύγουστο του 2014. Με ένταση περίπου 6 M_w ήταν ο ισχυρότερος στην περιοχή από το 1989. Η συλλογή των δεδομένων ξεκίνησε την 24 Αυγούστου και ολοκληρώθηκε τη 30 του ιδίου μηνός. Η συλλογή απαριθμεί περίπου 250K μηνύματα εκ των οποίων 2K επισημειωμένα. Παρατηρούμε ότι το δίκτυο DCNN εμφανίζει μακράν την καλύτερη επίδοση συγκρίσει των άλλων αρχιτεκτονικών στο απλό σύνολο. Αντίθετα στα σύνθετα σύνολα εμφανίζει συγκρίσιμες επιδόσεις με τον αλγόριθμο CNN. Αυτό οφείλεται κυρίως στο γεγονός ότι τα δεδομένα παρόλο που λίγα στον αριθμό είναι καλύτερης «ποιότητας». Πράγματι το σύνολο παρουσιάζει τον υψηλότερο δείκτη διεπισημειωτικής συμφωνίας όπως αναφέρθηκε και σε παραπάνω ενότητα.

Εν συνόλω παρουσιάζονται στο πίνακα 6.4 οι επιδόσεις μακράς κλίμακας και συγκεκριμένα ανάκληση, ακρίβεια, ορθότητα και f1-score για κάθε αλγόριθμο για το σύνολο των συλλογών που μελετήσαμε. Στο σχήμα (6.5) φαίνονται οι καμπύλες[1] ακρίβειας-ανάκλησης και τα εμβαδά που περικλείουν για κάθε τάξη και για τα τέσσερα υπό μελέτη συμβάντα. Ομοίως στο σχήμα (6.6) παρουσιάζονται οι καμπύλες χαρακτηριστικής λειτουργίας.

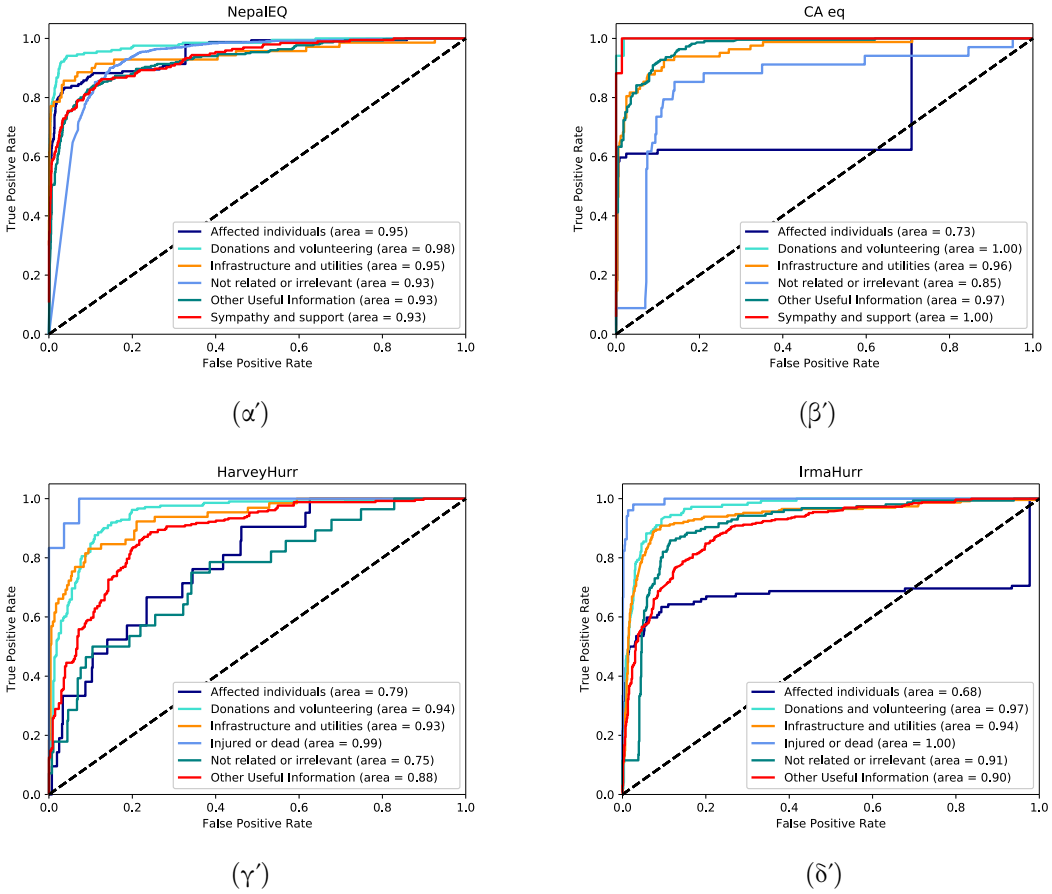
[1]Μελετάμε τις καμπύλες βέλτιστου αλγορίθμου και βέλτιστης διάταξης δεδομένων.

Πίνακας 6.4: Η βαθμολόγηση της αποδοτικότητας των ταξινομήτων βάσει ορθότητας και μακρο-$f1$ για τους αλγορίθμους SVM, CNN και DCNN για τα διάφορα σύνολα εκπαίδευσης

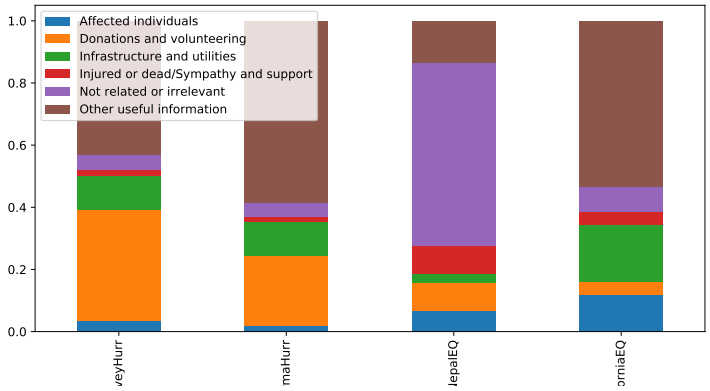
	SVM				CNN				DCNN			
	acc	f1	prec	rec	acc	f1	prec	rec	acc	f1	prec	rec
Τυφώνας Χάρβει												
In	68.847	53.473	66.366	48.931	72.849	39.193	52.541	39.148	74.536	54.678	67.380	52.112
Out	70.496	69.264	73.916	66.062	71.466	68.168	69.332	67.338	72.587	69.990	74.313	67.347
In_Out	68.224	55.584	74.323	51.955	**76.222**	**60.795**	67.287	58.284	75.548	58.503	67.821	54.785
Τυφώνας Ίρμα												
In	71.671	46.969	66.714	41.113	73.866	46.321	80.727	41.363	**76.267**	**53.417**	76.493	48.481
Out	68.611	69.569	72.654	67.266	74.974	76.486	77.920	75.747	76.084	73.714	75.849	74.052
In_Out	71.347	46.387	76.073	40.316	73.466	50.308	60.094	46.550	72.133	47.944	58.894	44.158
Σεισμός Καλιφόρνια												
In	73.118	59.402	60.336	59.138	79.474	74.582	85.647	69.042	**88.066**	**80.089**	83.659	79.408
Out	76.424	70.028	73.070	69.166	77.323	71.077	73.840	56.805	78.372	72.110	74.414	70.611
In_Out	74.151	68.742	71.883	66.743	77.545	68.318	78.094	65.147	77.806	68.825	79.241	64.277
Σεισμός Νεπάλ												
In	68.610	53.637	52.957	55.186	70.632	50.007	60.836	48.999	70.720	50.012	60.400	46.109
Out	77.609	69.216	78.168	65.653	78.915	77.775	78.000	78.286	79.270	74.175	75.009	74.198
In_Out	70.069	54.158	53.183	55.462	71.782	57.554	61.700	55.942	**85.228**	**78.137**	84.602	73.831



Σχήμα 6.5: Καμπύλες συμβιβασμού ακρίβειας-ανάκλησης και AUPR βαθμολογίες για κάθε κλάση



Σχήμα 6.6: Καμπύλες χαρακτηριστικής λειτουργίας και AUROC βαθμολογίες για κάθε κλάση



Σχήμα 6.7: Κατανομή κλάσεων για το σύνολο όλων των κρίσεων

6.5 Σύνοψη συμπερασμάτων αξιολόγησης

Κατά την εκπόνηση της εργασίας διεξήχθη ένα εκτετεταμένο σύνολο πειραμάτων βάσει τόσο κλασικών μεθόδων στατιστικής μάθησης και αναπαραστάσεων διακριτού σακιδίου λέξεων (bag of words) όσο και βαθέων συνελικτικών νευρωνικών μεθόδων που αξιοποιούν κατανεμημένες αναπαραστάσεις ενσωματώσιμων χαρακτηριστικών (word embedding). Η δομή των πειραμάτων λαμβάνει υπόψη διάφορα πραγματικά σενάρια και εφαρμόζεται επί πραγματικών συλλογών δεδομένων για να δείξει ότι η προτεινόμενη προσέγγιση είναι ρεαλιστική και εφικτή. Οι νευρωνικοί ταξινομητές επιδεικνύουν καλύτερες επιδόσεις σ όλα τα πειράματα. Οι διάφορες συλλογές παρουσιάζουν λίγο ή πολύ ασσυμετρία κλάσεων η οποία εξισορροπείται κατά την ενσωμάτωση των ετερογενών συλλόγων. Παρατηρούμε μια μείωση στην μέση ορθότητα αλλά καλύτερη βαθμολογία καμπυλών ακρίβειας -ανάκλησης στα σύνθετα σύνολα. Ο αλγόριθμος DCNN φαίνεται να είναι αποδοτικότερος ωστόσο ο πολύ απλούστερος CNN τον συναγωνίζεται στα περισσότερα πειράματα.

Κεφάλαιο 7

Τεχνικές λεπτομέρειες

7.0.1 Εισαγωγή

Σ αυτή την ενότητα θα μας απασχολήσουν τα τεχνικά ζητήματα του έργου. Αφού αναφερθούμε σε σημεία της υλοποίησης όπου χρειάζονται επεξήγηση θα δούμε συνολικά τις πλατφόρμες και τα προγραμματιστικά εργαλεία όπου χρησιμοποιήθηκαν για την υλοποίηση των αλγορίθμων.

7.1 Λεπτομέρειες υλοποίησης

7.1.1 'Πολλαπλή αναπαράσταση εισόδου'

Στον αλγόριθμο των Yoon Kim et al. [13] προβλέπεται εκτός από το απλό δίκτυο που αναλύσαμε θεωρητικά παραπάνω και η χρήση πολλαπλών «καναλιών», κατά αναλογία με τα δίκτυα υπολογιστκής όρασης, όπου τα «κανάλια» ισοδυναμουν με τους πίνακες παραμέτρων ενσωματώσιμων χαρακτηριστικών. Ο ένας πίνακας έχει σταθεροποιημένες τις παραμέτρους κατα τη διάρκεια της εκπαίδευσης οπότε δεν προσαρμόζεται στα πρότυπα εκπαίδευσης και ο δεύτερος πίνακας είναι προσαρμόσιμος στα πρότυπα. Ουσιαστικά εκπαιδεύουμε τις διπλάσιες παραμέτρους και συνενώνουμε τα διανύσματα χαρακτηριστικών πριν το πλήρως συνδεδεμένο στάδιο ταξινόμησης.

7.1.2 'Δυναμικός συνδυασμός χαρακτηριστικών'

Για τον δυναμικό $k-$μεγιτοποιητικό συνδυασμό χαρακτηριστικών χρησιμοποιήσαμε τη μέθοδο εκπαίδευσης κατά ταξινομημένες συστάδες. Ταξινομούμε τα πρότυπα εκπαιδευσης (ακολουθίες λεκτικών) κατά αύξον μήκος και στη συνέχεια χωρίσαμε σε συστάδες συμπληρώνοντας ενδεχομένως τις ακολουθίες στα όρια με null values ώστε όλα τα παραδείγματα της συστάδας να έχουν ένα και κοινό μήκος. Ειδάλλως θα έπρεπε να εκπαιδεύουμε online ανά παράδειγμα το οποίο δε είναι καθόλου αποδοτικό. Για την εξαγωγή των k-καλύτερων χαρακτηριστικών (k-max-pooling) δεν υπάρχει έτοιμη υλοποίηση επομένως χρησιμοποιούμε τη back-end συνάρτηση *argsort* η οποία όμως δεν υποστηρίζεται από τη βιβλιοθήκη pyCU-

DA και συνεπώς εκτελείται στη CPU . Η υλοποίηση αυτή επιδέχεται βελτιώσεις με τη χρήση υποστηριζόμενων αλγορίθμων ταξινόμησης όπως thrust sort κάτι που σε μελλοντικό χρόνο θα επιμεληθούμε.

7.1.3 Ἀναδίπλωση χαρακτηριστικών᾽

Όπως είδαμε σε παραπάνω ενότητα ο αλγόριθμος των Kalchbrener et al. [11] προβλέπει αναδίπλωση χαρακτηριστικών στα ανώτερα στρώματα δηλαδή άθροισμα γραμμών ανά δύο συνεχόμενων πριν τον μεγιστοποιητικό συνδυασμό και μετά το συνελικτικό στάδιο. Αντί να αθροίζουμε ανά δύο συνεχόμενες σειρές κάνουμε «αναδίπλωση» του πίνακα στη μέση έτσι κάθε γράμμη συνενώνεται με την $n/2$ συμμετρκή της. Στη συνέχεια αθροίζουμε παράλληλα όλες τις γραμμές και «αναδιπλώνουμε πίσω» το πίνακα στην αρχική διάταξη στηλών υποδιπλασιάζοντας έτσι τις γραμμές του.

7.2 Πλατφόρμες και προγραμματιστικά εργαλεία

Όλα τα πειράματα πραγματοποιήθηκαν σε σύγχρονο πολυπύρηνο επεξεργαστικό σύστημα και αξιοποίησαν τις δυνατότητες σύγχρονου εξειδικευμένου υλισμικού. Για την επιτάχυνση της φάσεως εκπαίδευσης χρησιμοποιήθηκε η NVIDIA Tesla K80 με τη βιβλιοθήκη CUDA 8.0 και εξειδικευμένο νευρωνικό optimizer CuDnn 7.0.

Η υλοποίηση των δικτύων CNN έγινε στην πλατφόρμα Tensorflow μέσω της διεπαφής της βιβλιοθήκης keras. Η πλατφόρμα επιτρέπει την παραγωγή αποδοτικού κώδικα και την εκμετάλευση της υψηλής παραλληλοποιησιμότητας των συνελίξεων στη GPU χωρίς περαιτέρω προσαρμογή στο κώδικα.

Για την υλοποίηση του δικτύου DCNN χρησιμοποιήθηκε η πλατφόρμα Theano μέσω της διαπαφής της βιβλιοθήκης lasagne. Η πλατφόρμα προσφέρεται για απλούστερη υλοποίηση στρωμάτων με μέγεθος που άλλαζει δυναμικά . Όποιος C-generated κώδικας υπεισέρχεται κατά την περιγραφή του γράφου του δικτύου που θα εκαπαιδευτεί, χρησιμοποιεί είτε τον GNU GCC compiler είτε όπου απαιτούνται συγκεκριμένες ρυθμίσεις τον LLVM Clang (πχ. αυξηση του βάθους εμφώλευσης nesting bracket depth). Σε μελλοντικό χρόνο θα υλοποιήσουμε και αυτό το δίκτυο στη πλατφόρμα Tensorflow λόγω καλύτερης υποστήριξης.

Για τον αλγόριθμο SVM χρησιμοποιήσαμε την έτοιμη υλοποίηση της βιβλιοθήκης python Libsvm όπως χρησιμοποιείται στην πλατφόρμα sklearn. Ο αλγόριθμος δεν αξιοποιεί τις δυνατότητες της GPU ωστόσο επειδή είναι ευκολότερη η εκπαίδευση του σε σχέση με τα νευρωνικά αυτό δε καθυστέρησε τη φάση της εκπαίδευσης.

Για την τεκμηρίωση των πειραματικών αποτελεσμάτων και την βελτιοποίηση κάποιων υπερπαραμέτρων των νευρωνικών δικτύων χρησιμοποίησαμε την πλατφόρμα Commet-ml [1] . Τα πειράματα εναποτίθενται εκεί και μπορούν να αναπαραχθούν σε μελλοντικό χρόνο.

[1] https://www.comet.ml

Κεφάλαιο 8

Επίλογος

Σ αυτό το σημείο αξίζει να συνοψίσουμε τη συνεισφορά των συνελικτικών νευρωνικών δικτύων στο πρόβλημα της ταχείας ταξινόμησης των σύντομων μηνυμάτων που μοιράζονται οι χρήστες κατα την εμφάνιση μιας εκτεταμένης ανθρωπιστικής κρίσης. Θα επαναλάβουμε τα βασικά συμπεράσματα της συγκριτικής μελέτης ως προς την αποδοτικότητα των διάφορων νευρωνικών αρχιτεκτονικών και τα βασικά τους πλεονεκτήματα έναντι των μη νευρωνικών αλγορίθμων . Τέλος θα δώσουμε μελλοντικές κατευθύνσεις για τη συνέχιση του έργου.

8.1 Σύνοψη και συμπεράσματα

Στα πλαίσια της παρούσας διπλωμάτικής προτείναμε τη χρήση συνελικτικών νευρωνικών δικτύων στο έργο της πραγματικού χρόνου κατηγοριοποίησης μεγάλων δεδομένων που συνδέονται με κάποια ανθρωπιστική κρίση . Η χρήση δυναμικού συνδυασμου χαρακτηριστικών μας επιτρέπει να εξάγουμε μεγαλύτερου εύρους και υψηλότερου επιπέδου δομημένα χαρακτηριστικά , συνάμα όμως η χρήση πολλαπλών φίλτρων και αναπαραστάσεων εισόδου μπορεί να εξάγει εξίσου ανταγωνιστικά χαρακτηριστικά με τους κατάλληλους συνδυασμούς πολυγραμμικών. Πειραματιστήκαμε με τη μείωση της υπερπορσαρμογής όπου και στις δύο περιπτώσεις η μέθοδος παράλειψης παραμέτρων δίνει καλύτερα αποτελέσματα συγκρίσει άλλων μεθόδων (π.χ κανονικοποίηση σε επίπεδο συστάδας). Η χρήση δεδομένων από παλαιότερες καταστροφές και μεθόδων μεταφοράς γνώσης βελτιώνει κάποιες φορές την απόδοση των αλγορίθμων κυρίως όταν διαθέτουμε μικρό σύνολο παραδειγμάτων εκπαίδευσης.

8.2 Μελλοντικές επεκτάσεις

Μελλοντικά θα επιθυμούσαμε να πειραματιστούμε με πιο σύνθετες μεθόδους προσαρμογής των δεδομένων που δεν είναι σχετικά με το συμβάν για παράδειγμα adversarial networks που θα «τιμωρούν» τα «παραπλανητικά» παραδείγματα των εξωγενών σύλλογών. Ακόμη ασχολούμαστε με την αξιοποίηση πολλαπλών αναπαραστάσεων εισόδου από πολλαπλές προεκπαιδευμένες συλλογές διανυσμάτων ενσωματώσιμων χαρακτηριστικών και πιο σύνθετων μεθόδων κανικοποίησης όπως multigroup normalization μεθόδους. Υπάρχουν έτοιμες (off-the-shelf)

συλλογές οι οποίες μοντελοποιούν, εκτός από λεξιλογικά συμφραζόμενα, συντακτικές δομές [16] και ίσως η καλύτερη μοντελοποίηση των δεδομένων βελτιώσει τις επιδόσεις των αλγορίθμων.

Bibliography

Bibliography

[1] A. Acar and Y. Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *IJWBC*, 7(3):392–402, 2011.

[2] F. Alam, F. Ofli, and M. Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.*, pages 465–473. AAAI Press, 2018.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

[5] D. Cosley, A. Forte, L. Ciolfi, and D. McDonald, editors. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*. ACM, 2015.

[6] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[7] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

[8] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4):248–260, 2009.

[9] M. Imran, P. Mitra, and C. Castillo. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* European Language Resources Association (ELRA), 2016.

[10] J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. In J. A. Carroll, A. van den Bosch, and A. Zaenen, editors, *ACL 2007, Proceedings of the*

45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. The Association for Computational Linguistics, 2007.

[11] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 655–665. The Association for Computer Linguistics, 2014.

[12] E. KicKiman and M. Richardson. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 547–556. ACM, 2015.

[13] Y. Kim. Convolutional neural networks for sentence classification. In Moschitti et al. [20], pages 1746–1751.

[14] J. Kivinen and M. K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. In F. T. Leighton and A. Borodin, editors, *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing, 29 May-1 June 1995, Las Vegas, Nevada, USA*, pages 209–218. ACM, 1995.

[15] L. Lambert, C. J. Moschovitis, H. W. Poole, and C. Woodford. *The internet: a historical encyclopedia*, volume 2. ABC-CLIO, 2005.

[16] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308. The Association for Computer Linguistics, 2014.

[17] B. F. Liu, J. D. Fraustino, and Y. Jin. Social media use during disasters. *Communication Research*, 43(5):626–646, 2016.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[19] D. Mileti. *Disasters by design: A reassessment of natural hazards in the United States*. Joseph Henry Press, 1999.

[20] A. Moschitti, B. Pang, and W. Daelemans, editors. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 2014.

[21] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 632–635. AAAI Press, 2017.

[22] B. O'Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. In W. W. Cohen and S. Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press, 2010.

[23] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, editors, *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press, 2014.

[24] A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In Cosley et al. [5], pages 994–1009.

[25] L. Palen and S. B. Liu. Citizen communications in crisis: anticipating a future of ict-supported public participation. In M. B. Rosson and D. J. Gilmore, editors, *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 727–736. ACM, 2007.

[26] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Moschitti et al. [20], pages 1532–1543.

[27] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

[28] I. Shklovski, M. Burke, S. Kiesler, and R. Kraut. Technology adoption and use in the aftermath of hurricane katrina in new orleans. *american Behavioral scientist*, 53(8):1228–1246, 2010.

[29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[30] S. Theodoridis and K. Koutroumbas. *Pattern recognition*. Academic Press, 1999.

[31] K. J. Tierney. Disaster preparedness and response: Research findings and guidance from the social science literature. 1993.

[32] V. Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural Networks*, 10(5):988–999, 1999.

[33] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In

E. D. Mynatt, D. Schoner, G. Fitzpatrick, S. E. Hudson, W. K. Edwards, and T. Rodden, editors, *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, pages 1079–1088. ACM, 2010.

[34] A. H. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.

[35] X. Wang, F. Zhu, J. Jiang, and S. Li. Real time event detection in twitter. In J. Wang, H. Xiong, Y. Ishikawa, J. Xu, and J. Zhou, editors, *Web-Age Information Management - 14th International Conference, WAIM 2013, Beidaihe, China, June 14-16, 2013. Proceedings*, volume 7923 of *Lecture Notes in Computer Science*, pages 502–513. Springer, 2013.

[36] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1168–1175. ACM, 2008.