National Technical University of Athens
School of Electrical and Computer Engineering
Department of Signals, Control and Robotics

Ανίχνευση Κατάθλιψης από Ψυχοθεραπευτικές Συνεδρίες με Μηχανική Μάθηση και Νευρωνικά Δίκτυα

Diploma Thesis

of

Danai Xezonaki

Supervisor: Alexandros Potamianos
Associate Professor NTUA

Speech and Natural Language Processing Lab (SLP)
Athens, July 2020

National Technical University of Athens
School of Electrical and Computer Engineering
Department of Signals, Control and Robotics
Speech and Natural Language Processing Lab (SLP)

Ανίχνευση Κατάθλιψης από Ψυχοθεραπευτικές Συνεδρίες με Μηχανική Μάθηση και Νευρωνικά Δίκτυα

Diploma Thesis

of

Danai Xezonaki

Supervisor: Alexandros Potamianos
Associate Professor NTUA

Approved by the commitee on 17 July, 2020.

....	....	....
Alexandros Potamianos	Konstantinos Tzafestas	Athanasios Katsamanis
Associate Professor NTUA	Associate Professor NTUA	Principal Researcher ATHENA RC

Athens, July 2020

....

Danai Xezonaki
Electrical and Computer Engineer
National Technical University of Athens

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Technical University of Athens.

Περίληψη

Εισαγωγή

Στην παρούσα διπλωματική εργασία ασχολούμαστε με το πρόβλημα της διάγνωσης κατάθλιψης από ψυχοθεραπευτικές συνεδρίες, με χρήση μηχανικής μάθησης και νευρωνικών δικτύων. Συγκεκριμένα, ερευνούμε τρόπους και μεθόδους για την βελτίωση της απόδοσης των αναδρομικών νευρωνικών δικτύων που χρησιμοποιούνται για το συγκεκριμένο πρόβλημα. Αρχικά, πραγματοποιούμε ανάλυση των δεδομένων μας και παρουσιάζουμε τα αποτελέσματα που φανερώνουν χρήσιμες πληροφορίες για τη γλώσσα που χρησιμοποιούν οι άνθρωποι με κατάθλιψη. Ακόμη, υλοποιούμε ένα ιεραρχικό δίκτυο με μηχανισμό προσοχής το οποίο μαθαίνει να προβλέπει εάν ο ασθενής κάθε συνεδρίας νοσεί από κατάθλιψη ή όχι. Στη συνέχεια, ερευνούμε την ενσωμάτωση υπάρχουσας γλωσσικής πληροφορίας στους μηχανισμούς προσοχής. Επίσης, διευρύνουμε τις αρχιτεκτονικές ενσωματώνοντας την περίληψη κάθε συνεδρίας, εφόσον είναι διαθέσιμη. Πραγματοποιούμε μια σειρά από πειράματα με βάση αυτές τις προσεγγίσεις και εξετάζουμε κατά πόσο βοηθά ολόκληρη η συνεδρία ή η γλώσσα του ασθενή και του θεραπευτή ξεχωριστά, στη βελτίωση της απόδοσης των μοντέλων. Τα αποτελέσματά μας δείχνουν ότι οι προτεινόμενες μέθοδοι βοηθούν την επίδοση των μοντέλων, και ειδικά στην περίπτωση που έχουμε μικρό αριθμό από δεδομένα. Τέλος, εισάγουμε το πρόβλημα της μοντελοποίησης του διαλόγου που πραγματοποιείται σε κάθε συνεδρία και συζητάμε πιθανές αρχιτεκτονικές και μελλοντικές επεκτάσεις. Η παρούσα δουλειά οδήγησε στην υποβολή του επιστημονικού άρθρου [81] στο συνέδριο Interspeech 2020.

Θεωρητικό Υπόβαθρο

Διαταραχές Διάθεσης

Οι διαταραχές διάθεσης είναι ένα σύνολο διαταραχών που επηρεάζουν άμεσα τη διάθεση του ατόμου που νοσεί. Μπορεί να εμφανίσει ιδιαίτερα αυξημένη διάθεση, όπως στην περίπτωση της μανίας, ιδιαίτερα μειωμένη διάθεση, όπως στην περίπτωση της κατάθλιψης, ή ένα συνδυασμό των δύο, όπως στη διπολική διαταραχή. Σύμφωνα με το διαγνωστικό και στατιστικό εγχειρίδιο των ψυχικών διαταραχών της Αμερικανικής Ψυχιατρικής Εταιρίας (DSM-IV) [1], οι διαταραχές διάθεσης διακρίνονται στις εξής κατηγορίες:

- Μείζων καταθλιπτική διαταραχή
- Δυσθυμική διαταραχή
- Άτυπη καταθλιπτική διαταραχή
- Διπολικές διαταραχές

- Άτυπη διπολική διαταραχή
- Κυκλοθυμική διαταραχή
- Διαταραχή διάθεσης λόγω ιατρικής κατάστασης
- Διαταραχή διάθεσης λόγω χρήσης ουσιών

Μείζων καταθλιπτική διαταραχή

Η κατάθλιψη ή μείζων καταθλιπτική διαταραχή είναι ένα αίσθημα έντονης λύπης που μπορεί να οφείλεται σε κάποιο λυπηρό γεγονός αλλά είναι δυσανάλογο του γεγονότος που το προκάλεσε τόσο σε ένταση όσο και σε διάρκεια. Σύμφωνα με τον Παγκκόσμιο Οργανισμό Υγείας [55], εκτιμάται ότι 300 εκατομμύρια άνθρωποι νοσούν από κατάθλιψη, το οποίο αντιστοιχεί στο 4.4% του παγκόσμιου πληθυσμού. Παράλληλα σημειώνονται κάθε χρόνο πάνω από 800.000 θάνατοι που οφείλονται σε κατάθλιψη, ενώ για ανθρώπους ηλικίας $15-29$ ετών είναι η κύρια αιτία θανάτου.

Τα συμπτώματα της κατάθλιψης αναπτύσσονται σταδιακά μέσα σε ένα διάστημα εβδομάδων κατά τις οποίες το άτομο που εκδηλώνει καταθλιπική κρίση εμφανίζεται λυπημένο και απογοητευμένο καθόλη τη διάρκεια της ημέρας, και αδυνατεί να απολαύσει καθημερινές του συνήθειες. Οι διαταραχές ύπνου είναι συχνές ενώ παρουσιάζονται επίσης διατροφικές διαταραχές. Ακόμη, τα άτομα τείνουν να παρουσιάζουν χαμηλή συγκέντρωση, αισθήματα ενοχών, ανεπάρκειας και χαμηλής αυτοεκτίμησης. Σε πιο σοβαρές περιπτώσεις μπορεί να υπάρξουν και αυτοκτονικές σκέψεις.

Η κλινική διάγνωση βασίζεται συνήθως στα συμπτώματα, στη διάρκεια και στη σοβαρότητά τους. Μερικές φορές μπορεί να χρησιμοποιηθούν και τυποποιημένα ερωτηματολόγια ώστε να εκτιμηθεί ο βαθμός της κατάθλιψης. Η θεραπεία της διαταραχής είναι ευτυχώς δυνατή, μέσα από τη χορήγηση φαρμακευτικής αγωγής.

Ανίχνευση κατάθλιψης (depression detection)

Η ανίχνευση κατάθλιψης αφορά στη διαδικασία αναγνώρισης καταθληπτικών ενδείξεων σε ασθενείς. Τέτοιες ενδείξεις ανιχνεύονται στα χαρακτηριστικά του προσώπου των ασθενών και στην ομιλία τους, η οποία περιλαμβάνει τόσο τη φωνή όσο και τη χρήση της γλώσσας. Πρόκειται για ένα δύσκολο και πολυπαραγοντικό πρόβλημα, για το οποίο θα πρέπει να ληφθεί υπόψιν η συνολική εικόνα του ασθενή. Έτσι, οι ειδικοί ψυχικής υγείας, προκειμένου να προβούν σε διάγνωση, θα πρέπει να εξετάσουν πόσα συμπτώματα κατάθλιψης έχει ο ασθενής, πόσο καιρό διαρκούν και σε ποιο βαθμό δυσχεραίνουν την καθημερινότητά του.

Τα συχνότερα συμπτώματα κατάθλιψης αποτελούν η καταθληπτική διάθεση καθόλη τη διάρκεια της ημέρας, η μειωμένη όρεξη για οποιαδήποτε δραστηριότητα και απόλαυση (ανηδονία), οι διατροφικές διαταραχές και διαταραχές ύπνου καθώς και αίσθημα κατωτερότητας και ενοχών. Σε σοβαρές περιπτώσεις μπορεί να υπάρχουν ακόμη και αυτοκτονικές τάσεις. Ωστόσο, οι εκφάνσεις της κατάθλιψης είναι πολλές και ως εκ τούτου τα συμπτώματα και η έντασή τους διαφέρουν από άνθρωπο σε άνθρωπο. Για παράδειγμα, κάποιοι άνθρωποι που νοσούν από κατάθλιψη μπορεί να έχουν συμπτώματα άγχους και θυμού, ενώ άλλοι να διακρίνονται από μεγάλη αναποφασιστικότητα. Είναι επομένως αναγκαίο η κάθε περίπτωση να εξετάζεται ξεχωριστά και φυσικά από τους αρμόδιους ειδικούς, ώστε να γίνει σωστή διάγνωση και αντιμετώπιση της νόσου.

Αναγνώριση και Ανάλυση Συναισθημάτων

Ως αναγνώριση συναισθημάτων ορίζουμε τον προσδιορισμό ανθρώπινων συναισθημάτων, όπως

θυμό, φόβο, χαρά, λύπη, έκπληξη και αποστροφή, διαμέσου της έκφρασης. Τα συναισθήματα μπορούν να αναγνωριστούν μέσω των χαρακτηριστικών του προσώπου των ατόμων, στην ομιλία τους ή και στο γραπτό τους λόγο. Παρόλο που είναι ένας σχετικά νεοσύστατος κλάδος, έχει γίνει ιδιαίτερα δημοφιλής λόγω της ικανότητας των υπολογιστικών συστημάτων να ανιχνεύσουν συναισθήματα βάσει αντικειμενικών κανόνων, σε αντίθεση με τους ανθρώπους των οποίων η άποψη διαφέρει πολύ σε τέτοιες προβλέψεις.
Η ανάλυση των συναισθημάτων μπορεί να θεωρηθεί ως η φυσική εξέλιξη της αναγνώρισής τους. Πρόκειται για ένα επίσης πολύ σημαντικό πεδίο του τομέα της επεξεργασίας φυσικής γλώσσας, στο οποίο έχουν βασιστεί χιλιάδες μελέτες. Η αναγνώριση συναισθήματος έχει φανεί εξαιρετικά χρήσιμη σε αντικείμενα που είναι αναγκαία η κατανόηση των συναισθημάτων του δέκτη, όπως στο μάρκετινγκ [6], στα συστήματα συστάσεων [3] και στα συστήματα αυτόματης ερώτησης-απάντησης (quention-answering) [70].

Feature Engineering

Δεδομένης μιας εισόδου κειμένου, χρειάζεται να εξάγουμε χαρακτηριστικά (features) από τα δεδομένα ώστε να τα χρησιμοποιήσουμε για classification. Τέτοιες μέθοδοι παρουσιάζονται στη συνέχεια.

Προεκπαιδευμένα Διανύσματα Λέξεων (Word Embeddings)

Χρησιμοποιούμε τα προεκπαιδευμένα GloVe διανύσματα λέξεων και αρχικοποιούμε το embedding layer του δικτύου. Τα διανύσματα αυτά έχουν προκύψει από την εκπαίδευση ενός δικτύου στο Common Crawl corpus και αποδίδουν μια αναπαράσταση 300 διαστάσεων σε κάθε λέξη. Έχουν το χαρακτηριστικό να αποδίδουν διανύσματα με κοντινές αποστάσεις στον 300-d χώρο σε λέξεις με κοντινό σημασιολογικό περιεχόμενο.

TF-IDF

Η μέθοδος TF-IDF αντανακλά τη σημαντικότητα μιας λέξης σε ένα κείμενο. Το αποτέλεσμα προκύπτει μέσω του πολλαπλασιασμού του αριθμού εμφάνισης της λέξης σε ένα κείμενο με την αντίστροφη συχνότητα της λέξης σε όλα τα κείμενα ενός dataset. Με αυτόν τον τρόπο, λέξεις που είναι συχνές σε όλα τα κείμενα, ακόμη και αν εμφανίζονται πολλές φορές σε κάποιο συγκεκριμένο, δεν θεωρούνται σημαντικές για το νόημα. Από την άλλη μεριά, λέξεις που εμφανίζονται σε λίγα μόνο κείμενα θεωρούμε ότι ενδέχεται να περιέχουν σημαντική πληροφορία.

Η μέθοδος TF-IDF έχει πάρει το όνομά της από τους δύο όρους που την αποτελούν και συμβάλλουν στον υπολογισμό του αποτελέσματος. Ο όρος TF (term frequency) αφορά τη συχνότητα μιας λέξης και υπολογίζεται ως ο αριθμός των εμφανίσεών της. Ο όρος IDF (inverse-document frequency) δείχνει πόσο σπάνια είναι μια λέξη, και υπολογίζεται διαιρώντας το σύνολο των εγγράφων σε ένα dataset με τον αριθμό των εγγράφων που περιέχουν τη λέξη.

Προεπεξεργασία των Δεδομένων

Στα προβλήματα επεξεργασίας φυσικής γλώσσας, η είσοδος σε μορφή κειμένου πρέπει να μετατραπεί σε μορφή που μπορούν να καταλάβουν και να επεξεργαστούν τα υπολογιστικά συστήματα. Για το σκοπό αυτό πρέπει να ακολουθηθούν κάποιες διαδικασίες.

- Tokenization: Δεδομένης μιας εισόδου που αποτελείται από μια ακολουθία χαρακτήρων, ως tokenization ορίζεται η διαδικασία της διάσπασης της ακολουθίας εισόδου σε tokens (υποακολουθίες) που ανήκουν στο διαθέσιμο λεξιλόγιο και απόρριψης συμβολοσειρών που δεν ανήκουν σε αυτό. Σε περίπτωση εισόδου προτάσεων, οι υποακολουθίες αυτές αντιστοιχούν σε λέξεις.

- Αφαίρεση σημείων στίξης: Μόλις η είσοδος προτάσεων έχει διασπαστεί σε λέξεις, ακολούθως πρέπει να αφαιρεθούν τα σημεία στίξης που δεν αποτελούν μέρος κάποιας λέξης, καθώς και ετικέτες του κειμένου (tags) που δεν συμβάλλουν στο νοηματικό περιεχόμενο της πρότασης.

- Αφαίρεση stop words: Ορισμένες λέξεις είναι κοινές και εμφανίζονται πολύ συχνά σε κείμενα, όπως οι λέξεις "εγώ", "και", "είναι", "το". Αφαιρώντας αυτές τις λέξεις που δεν προσδίδουν ιδιαίτερο νόημα στο κείμενο, βοηθούμε τους αλγορίθμους να εστιάσουν σε περισσότερο σημαντικές λέξεις που παραμένουν στο κείμενο. Για την αφαίρεση των συχνών λέξεων μπορούμε είτε να ορίσουμε μια λίστα από όσες επιθυμούμε να αφαιρεθούν, είτε να χρησιμοποιήσουμε τέτοιες έτοιμες διαθέσιμες λίστες.

- Stemming: Με τον όρο stemming αναφερόμαστε στη διαδικασία κατά την οποία κρατάμε μόνο τη ρίζα μιας λέξης πετώντας τους επιπλέον χαρακτήρες της κατάληξης. Έτσι, σχετικές λέξεις μεταξύ τους ελαττώνονται στην ίδια ρίζα και επομένως οι αλγόριθμοι τις θεωρούν συνώνυμες.

- Lemmatization: Ο όρος lemmatization αφορά την μείωση των λέξεων στην βασική τους μορφή (lemma) και την ομαδοποίησή τους βάσει αυτού. Προκειμένου να γνωρίζουν τις βασικές μορφές των λέξεων, οι αλγόριθμοι θα πρέπει να έχουν πρόσβαση σε λεξικά.

Γλωσσικά και Συναισθηματικά Λεξικά (Linguistic and Affective Lexica)

Πρόκειται για γλωσσικές βάσεις δεδομένων καταγεγραμμένες από εμπειρογνώμονες για ένα σύνολο λέξεων. Πιο συγκεκριμένα, σε κάθε λέξη στο λεξικό αποδίδεται ένα σύνολο από τιμές που αφορούν σε ψυχολογικά, συναισθηματικά και γλωσσικά χαρακτηριστικά της λέξης. Η ενσωμάτωση τέτοιας πληροφορίας σε συστήματα επεξεργασίας φυσικής γλώσσας βελτιώνει σημαντικά την επίδοση των αλγορίθμων. Τα λεξικά που χρησιμοποιούνται στην παρούσα εργασία είναι τα ακόλουθα: LIWC [74], Bing Liu Opinion Lexicon [32], AFINN [86], Subjectivity Lexicon (MPQA) [79], SemEval 2015 English Twitter Lexicon (Semeval15) [38] και NRC Emotion Lexicon (Emolex) [52]. Τα AFINN, Semeval15 και Bing Liu παρέχουν ένα $1D$ διάνυσμα που υποδηλώνει θετικό/αρνητικό συναίσθημα για 6,786, 1,515 και 2,477 λέξεις αντίστοιχα. Το MPQA παρέχει ένα $4D$ διάνυσμα με sentiment ratings για 6,886 λέξεις. Το Emolex διαθέτει διάνυσμα $19D$ με emotion ratings για 14,182 λέξεις. Τέλος, το LIWC διαθέτει ένα διάνυσμα $73D$ με ψυχο-γλωσσικές (psycholinguistic) ενδείξεις για 18,504 λέξεις. Ο συνδυασμός των 6 λεξικών καλύπτει λεξιλόγιο μεγέθους 25,534 λέξεων. Τα χαρακτηριστικά που εξάγονται από τα λεξικά για κάθε λέξη συνενώνονται (concatenation) και σχηματίζουν ένα διάνυσμα 99 διαστάσεων (context vector).

Ταξινόμηση Κειμένου

Ταξινόμηση κειμένου (document classification) είναι η διαδικασία της κατηγοριοποίησης ενός κειμένου σε μία ή περισσότερες κατηγορίες, βάσει του περιεχομένου, του είδους ή και του συγγραφέα. Χρησιμοποιείται ευρέως σαν τεχνική με σκοπό την ταξινόμηση και το χειρισμό εγγράφων βασισμένων σε κείμενο, όπως emails, άρθρα ή αποτελέσματα ερευνών. Η διαδικασία μπορεί να γίνει είτε χειροκίνητα, σύμφωνα με τους κανόνες της βιβλιοθηκονομίας, ή αυτόματα με χρήση αλγορίθμων ταξινόμησης. Και οι δύο μέθοδοι έχουν πλεονεκτήματα και μειονεκτήματα. Από τη μία πλευρά, ο ανθρώπινος παράγοντας παίζει σημαντικό ρόλο στην επιλογή των κατηγοριών και στο σωστό έλεγχο της διαδικασίας. Ωστόσο, σε περιπτώσεις που έχουμε μεγάλο αριθμό από έγγραφα, οι αυτόματες διαδικασίες είναι σίγουρα πιο γρήγορες και αποτελεσματικές, και παραμένουν ανεπηρέαστες ως προς τις αποφάσεις.

Υπάρχουν διαφορετικές προσεγγίσεις για ταξινόμηση κειμένου.

- Επιβλεπόμενη μάθηση (supervised learning): παρέχουμε στους αλγορίθμους κείμενα με τις κατηγορίες που τους έχουμε αναθέσει. Τα μοντέλα επομένως μαθαίνουν να συσχετίζουν το περιεχόμενο με τις κατηγορίες και έτσι μπορούν να βγάλουν σωστά αποτελέσματα για κείμενα που δεν έχουν ξανασυναντήσει.

- Μη επιβλεπόμενη μάθηση (unsupervised learning): Σε αυτή την περίπτωση, τα μοντέλα μαθαίνουν τα κατηγοριοποιούν κείμενο χωρίς την ανθρώπινη παρέμβαση. Η ταξινόμηση γίνεται χωρίς αναφορά σε εξωτερική πληροφορία οπότε οι αλγόριθμοι ομαδοποιούν τα κείμενα με παρόμοιες λέξεις ή προτάσεις.

- Rule-based: Η μέθοδος αυτή βασίζεται σε γλωσσικούς, μορφολογικούς, συντατικούς ή σημασιολογικούς κανόνες που ορίζουν την κατηγορία του κάθε κειμένου και δίνουν οδηγίες στους αλγορίθμους. Ακολουθώντας αυτούς τους κανόνες, τα μοντέλα μπορούν να αναθέσουν αυτόματα κατηγορίες.

Μηχανισμός αυτο-προσοχής (Self-Attention Mechanism)

Ο μηχανισμός αυτο-προσοχής χρησιμοποιείται ώστε να αποδοθεί διαφορετική σημασία σε κάθε λέξη του κειμένου. Συγκεκριμένα, ωθεί το μοντέλο να δώσει μεγαλύτερο βάρος σε λέξεις που είναι καθοριστικές για το νόημα και μικρότερο σε πιο κοινές λέξεις, που δεν συμβάλλουν σημαντικά στο περιεχόμενο της πρότασης. Αναθέτει επομένως μια τιμή a_i σε κάθε αναπαράσταση λέξης που έχουμε πάρει από το αναδρομικό δίκτυο. Έτσι, η συνολική αναπαράσταση της πρότασης προκύπτει ως το σταθμισμένο άθροισμα των αναπαραστάσεων και δίνεται από τον τύπο:

$$\begin{aligned} f(h_i) &= v_a^T tanh(W_a h_i + b_a) \\ a_i &= softmax(f(h_i)) \\ r &= \sum_i a_i h_i \end{aligned} \tag{0.1}$$

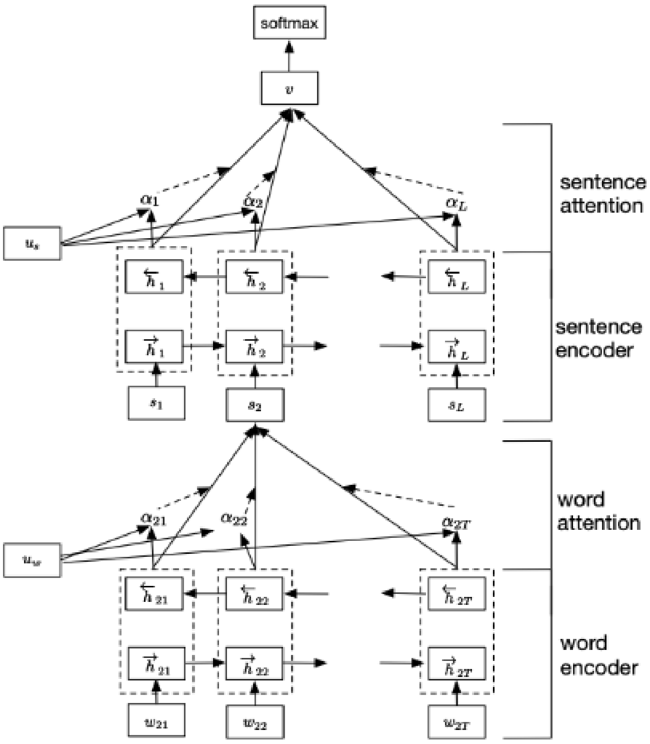
Conditional Self-Attention

Ο μηχανισμός προσοχής επαυξάνεται ώστε να λάβει υπόψιν τις γλωσσικές πληροφορίες που έχουμε πάρει από τα λεξικά. Δίνουμε επομένως ακόμη μεγαλύτερη ώθηση στο μηχανισμό προσοχής ώστε να καθορίσει την κατανομή των σκορ προσοχής σε κάθε πρόταση. Για το σκοπό αυτό, τροφοδοτούμε στο μηχανισμό σαν είσοδο, τόσο την αναπαράσταση κάθε λέξης από το μοντέλο μας όσο και το εξωτερικό διάνυσμα με τις λεξικές αναπαραστάσεις. Τα δύο διανύσματα συνδυάζονται μέσω μιας συνάρτησης συνένωσης (concatenation). Η βασική ιδέα είναι ότι προσθέτοντας επιπλέον διαστάσεις και χαρακτηριστικά σε κάθε λέξη, τόσο πιο διαφοροποιήσιμη την κάνουμε από τις υπόλοιπες. Επομένως, ο αλγόριθμος μπορεί να τις ξεχωρίζει καλύτερα, όπως και το νόημά τους συνολικά στην πρόταση. Η είσοδος που τροφοδοτείται τελικά στο μηχανισμό προσοχής δίνεται από τη σχέση:

$$f(h_i, c(w_i)) = tanh(W[h_i||c(w_i)] + b) \tag{0.2}$$

Ιεραρχικός Μηχανισμός Προσοχής

Πρόκειται για μια παραλαγή του κλασσικού μηχανισμού προσοχής που μπορεί να εφαρμοστεί σε διαφορετικά επίπεδα ενός δικτύου, όπως αυτό του σχήματος 1. Στην περίπτωση του document classification, η είσοδος είναι ένα κείμενο που αποτελείται από προτάσεις, και οι προτάσεις με τη σειρά τους αποτελούνται από λέξεις. Επομένως το δίκτυο θα αποτελείται από δύο στάδια ώστε να μοντελοποιήσει τις ιεραρχίες που υπάρχουν στο κείμενο. Ο μηχανισμός



Σχήμα 1: A hierarchical attention network. Source: Buomsoo Kim.

προσοχής στα δύο επίπεδα αποφασίζει ποιες προτάσεις είναι πιο σημαντικές στο σύνολο του κειμένου και ποιες λέξεις είναι πιο σημαντικές σε κάθε πρόταση.

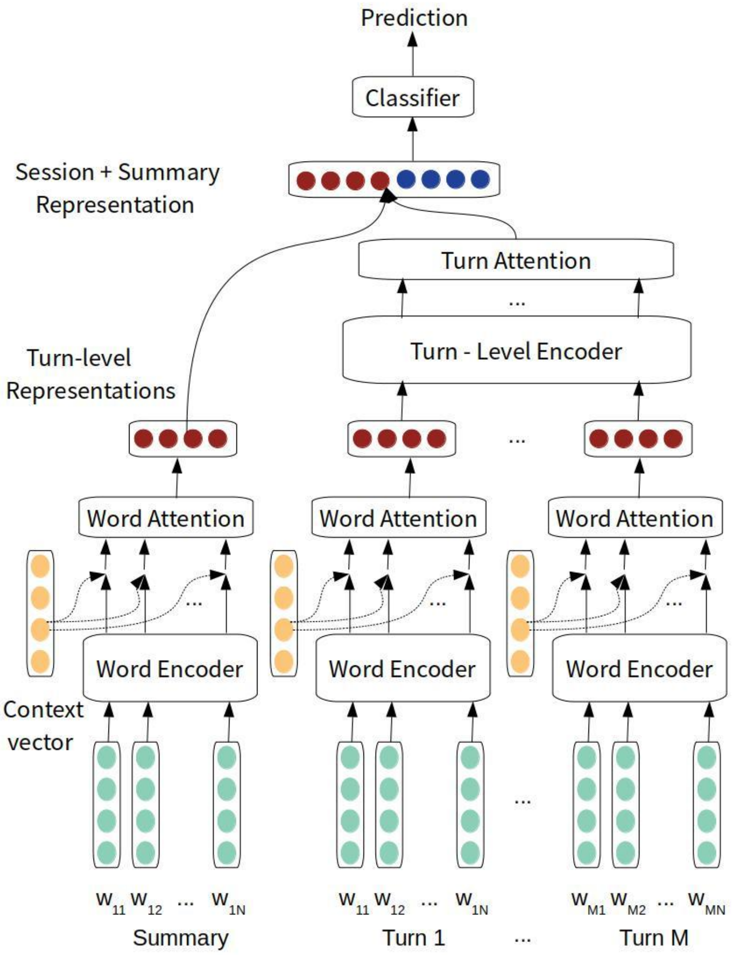
Evaluation Metrics
Κατά την ανάπτυξη και εκπαίδευση μοντέλων, θέλουμε να γνωρίζουμε πόσο καλά έχουν μάθει τα συστήματα να ανταποκρίνονται στο task για τα οποία τα εκπαιδεύουμε. Για να μπορέσουμε να μετρήσουμε πόσο καλά έχουν μάθει το συγκεκριμένο πρόβλημα, χρησιμοποιούμε μετρικές που μετράνε το performance των μοντέλων στο test set και τα βοηθούν να βελτιωθούν μέχρι να φτάσουν ένα επιθυμητό performance. Οι πιο διαδεδομένες μετρικές είναι το Accuracy και το F1-score.
Accuracy
Χρησιμοποιείται κυρίως σε binary classification προβλήματα. Υπολογίζεται διαιρώντας τον αριθμό των σωστών προβλέψεων του μοντέλου με τον συνολικό αριθμό προβλέψεων.
F1-score
Χρησιμοποιείται σε περιπτώσεις multi-class classification ή και σε binary classification προβλήματα που ο αριθμός των παρατηρήσεων δεν είναι ισοκατανεμημένος μεταξύ των κλάσεων. Υπολογίζεται ως ο αρμονικός μέσος όρος των Precision και Recall.

Ανίχνευση Κατάθλιψης με χρήση Αναδρομικών Νευρωνικών Δικτύων

Σε αυτή την εργασία, προσεγγίζουμε ένα δυαδικό πρόβλημα ταξινόμησης. Στόχος είναι να υλοποιήσουμε συστήματα που μπορούν να προβλέψουν εάν ένας ασθενής νοσεί από κατάθλιψη, με βάση το διάλογό του κατά τη διάρκεια μιας συνεδρίας με έναν ειδικό ψυχικής υγείας. Προσεγγίζουμε αυτό το πρόβλημα αναπτύσσοντας ιεραρχικά δίκτυα εξοπλισμένα με μηχανισμό προσοχής. Τα ιεραρχικά δίκτυα, αποτελούνται από στρώματα νευρωνικών δικτύων, που το κάθε στρώμα επικεντρώνεται σε διαφορετικά κομμάτια του κειμένου για να εξάγει πληροφορίες. Στη δεδομένη περίπτωση εκμεταλλευόμαστε την ιεραρχική δομή των κειμένων που περιέχουν το διάλογο της συνεδρίας, τα οποία αποτελούνται από προτάσεις, και η κάθε πρόταση αποτελείται από λέξεις.

Προτεινόμενη Μέθοδος

Αρχικά, υλοποιούμε ένα ιεραρχικό δίκτυο δύο επιπέδων, όπου το πρώτο επίπεδο παίρνει ως είσοδο τις λέξεις μιας πρότασης και βγάζει μια συνολική αναπαράσταση για την πρόταση, ενώ το δεύτερο επίπεδο παίρνει την ακολουθία των προτάσεων και βγάζει μια συνολική αναπαράσταση για όλο το κείμενο. Αυτή η αναπαράσταση κειμένου δίνεται ακολούθως σε έναν ταξινομητή για την τελική πρόβλεψη. Κάθε ένα από τα δύο επίπεδα του δικτύου διαθέτει μηχανισμό αυτο-προσοχής. Στο πρώτο επίπεδο ο μηχανισμός βοηθάει να εστιάσουμε στις σημαντικότερες λέξεις μιας πρότασης, ενώ στο δεύτερο επίπεδο βοηθάει να εστιάσουμε στις σημαντικότερες προτάσεις του κειμένου.Η αρχιτεκτονική του ιεραρχικού μοντέλου φαίνεται στο Σχήμα 2.



Σχήμα 2: Hierarchical Model with Attentional Conditioning

Ενσωμάτωση της γλωσσικής πληροφορίας

Στο κομμάτι αυτό, εισάγουμε τη γλωσσική πληροφορία από τα λεξικά στο μοντέλο. Όπως περιγράφηκε στο θεωρητικό μέρος, εξάγουμε την αναπαράσταση κάθε λέξης w από τα λεξικά σε ένα διάνυσμα 99 διαστάσεων, $c(w_t)$. Ακολούθως συνενώνουμε το διάνυσμα με την αναπαράσταση της λέξης όπως έχει προκύψει από το πρώτο επίπεδο του ιεραρχικού δικτύου, έστω h_t, το οποίο στα πλαίσια της εργασίας είναι ένα διάνυσμα 300 διαστάσεων. Το προκύπτον διάνυσμα επομένως έχει 399 διαστάσεις, και τροφοδοτείται στο μηχανισμό αυτο-προσοχής του πρώτου επιπέδου του δικτύου. Με τον τρόπο αυτό το μοντέλο αποκτά γνώση του συναισθηματικού νοήματος κάθε λέξης και μαθαίνει να δίνει προσοχή στις πιο σημαντικές λέξεις κάθε πρότασης.

Ενσωμάτωση της περίληψης

Για κάθε συνεδρία στο σύνολο των δεδομένων μας, έχει δοθεί η αντίστοιχη περίληψή της. Εισάγουμε την πληροφορία στο σύστημα θεωρώντας ότι η περίληψη παίζει σημαντικό ρόλο αφού συνοψίζει το νόημα του κειμένου. Για το σκοπό αυτό εξάγουμε την αναπαράσταση των προτάσεων της περίληψης από το πρώτο επίπεδο του δικτύου. Παρόλα αυτά, δεν ενώνουμε το αποτέλεσμα με τις υπόλοιπες προτάσεις του κειμένου, γιατί θεωρούμε ότι θα "χάσει" σε προσοχή. Έτσι, συνενώνουμε το διάνυσμα της περίληψης με την τελική αναπαράσταση του κειμένου ακριβώς πριν τον ταξινομητή. Εάν o_t είναι το διάνυσμα της περίληψης, τότε η είσοδος στον ταξινομητή θα είναι το διάνυσμα $(o_t||r)$, όπου r είναι το τελικό διάνυσμα του κειμένου που προκύπτει από το turn-level encoding σε συνδυασμό με το μηχανισμό προσοχής.

Περιγραφή των Δεδομένων

Για τους σκοπούς της εργασίας χρησιμοποιήθηκαν δύο διαφορετικά σύνολα δεδομένων. Πρόκειται για δύο συλλογές από καταγεγραμμένες ψυχοθεραπευτικές συνεδρίες.
Το πρώτο σύνολο είναι το General Psychotherapy Corpus [1], το οποίο περιέχει πάνω από 1300 αρχεία κειμένου, όπου έχει καταγραφεί ο διάλογος των συνεδριών. Οι συνδρίες καλύπτουν ένα μεγάλο φάσμα από θεραπευτικές προσεγγίσεις και συνοδεύονται από ένα σετ πληροφοριών που τους έχουν αποδοθεί από εμπειρογνώμονες. Συγκεκριμένα, για κάθε συνεδρία παρέχονται δημογραφικά χαρακτηριστικά τόσο για τον ασθενή όσο και για τον ψυχολόγο, τα συμπτώματα που αντιμετωπίζει ο ασθενής, καθώς και μια σύντομη περίληψη της συνεδρίας. Καθώς κάποιες από τις συνεδρίες διενεργήθηκαν μεταξύ τριών ή και περισσότερων ατόμων, επιλέγουμε να κρατήσουμε ένα υποσύνολο από 1262 συνεδρίες που αφορούν σε διάλογο μεταξύ ενός ασθενή και ενός ψυχολόγου. Μεταξύ αυτών, 881 έχουν την ένδειξη "not-depressed" για τους ασθενείς, ενώ οι υπόλοιπες 381 έχουν την ένδειξη "depressed".
Το δεύτερο σύνολο δεδομένων είναι το DAIC-WoZ 2017 depression dataset [26]. Περιλαμβάνει κλινικές συνεδρίες που διενεργήθηκαν με σκοπό να βοηθήσουν τη διάγνωση ψυχικών διαταραχών. Η συνέντευξη πραγματοποιήθηκε μεταξύ ενός ασθενή και ενός virtual agent με το ρόλο του θεραπευτή, ονόματι Ellie, το οποίο έλεγχε ένας άνθρωπος σε άλλο δωμάτιο [13]. Τα δεδομένα χωρίζονται σε train, development και test sets και αποτελούνται από 107, 35, 47 δείγματα αντίστοιχα. Ο βαθμός της κατάθλιψης έχει αξιολογηθεί στην κλίμακα PHQ-8.

Αποτελέσματα

Στον πίνακα 1 συγκρίνουμε τα αποτελέσματα των πειραμάτων μας για το General Psychotherapy Corpus, όταν δίνουμε σαν είσοδο στο μοντέλο τα λόγια μόνο του ασθενή (Client), τα λόγια μόνο του θεραπευτή (Therapist) ή ολόκληρη τη συνεδρία (Client+Therapist). Τα αποτελέσματα αναφέρονται στη μετρική F1. Παρατηρούμε ότι η ενσωμάτωση της εξωτερικής γλωσσικής πληροφορίας βελτιώνει την επίδοση των μοντέλων για όλα τα υλοποιημένα συστήματα σε σχέση με τα baselines, SVM και HAN. Ακόμη, παρατηρούμε ότι η ενσωμάτωση της περίληψης επίσης βοηθάει τους αλγορίθμους, και σε κάποιες περιπτώσεις σε μεγαλύτερο

[1] http://alexanderstreet.com

βαθμό από τη γλωσσική πληροφορία. Ο συνδυασμός των δύο τεχνικών επιφέρει σημαντική διαφορά στην περίπτωση του Client. Τέλος, διαπιστώνουμε ότι η γλώσσα του ασθενή είναι πιο σημαντική για την ανίχνευση κατάθλιψης σε σχέση με τη γλώσσα του θεραπευτή. Η καλύτερη επίδοση σημειώνεται για το HAN+S+L μοντέλο, ενώ το HAN+L πηγαίνει καλύτερα εάν δεν έχουμε διαθέσιμη την περίληψη της συνεδρίας.

Πίνακας 1: Results of different architectures on the GPC

Experiment	Client	Therapist	Client+Therapist
SVM	0.478	0.464	0.484
HAN	0.681	0.647	0.695
HAN+S	0.698	0.641	**0.718**
HAN+L	0.693	**0.659**	0.706
HAN+L+S	**0.715**	0.640	**0.716**

Στον πίνακα 2 παρουσιάζουμε τα αποτελέσματα για το DAIC-WoZ dataset. Παρατηρούμε ότι η ενσωμάτωση γλωσσικών πληροφοριών βελτιώνει σημαντικά την επίδοση του μοντέλου σε σχέση με το baseline HAN. Το HAN+L επίσης καταλήγει σε καλύτερα αποτελέσματα για τις μετρικές F1 και UAR σε σχέση με τα μοντέλα που προτείνονται στο /citesame. Συνολικά, διαπιστώνουμε ότι η ενσωμάτωση πληροφορίας (conditioning of external psycholinguistic knowledge) σε αυτό το μικρό dataset (189 δείγματα) βοηθάει πολύ το μοντέλο και τα αποτελέσματα είναι συγκρίσιμα με εκείνα του GPC.

Πίνακας 2: Results of the DAIC-WoZ corpus

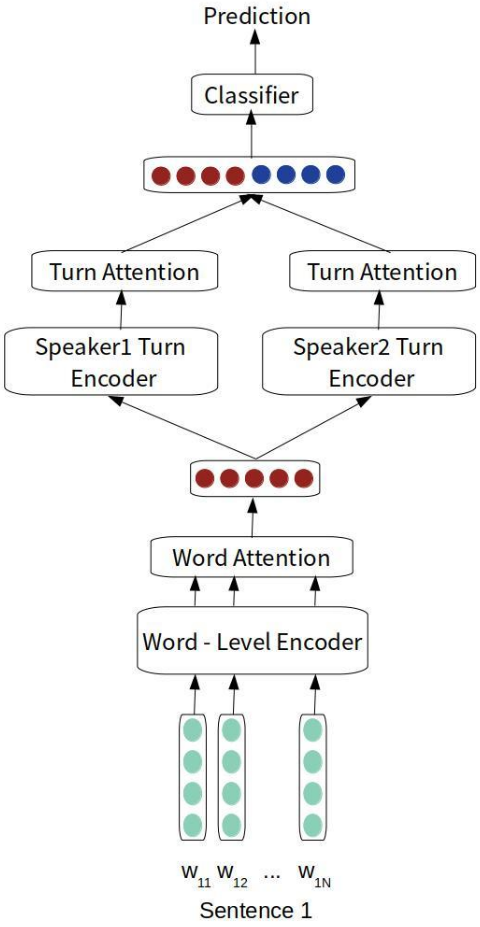
Method	Devel. Set		Test Set	
	F1-macro	UAR	F1-macro	UAR
[48]HCAN	0.51	0.54	0.63	0.66
[48]HLGAN	0.60	0.60	0.35	0.33
HAN	0.46	0.48	0.62	0.63
HAN+L	**0.62**	**0.63**	**0.70**	**0.70**

Μοντελοποίηση Διαλόγου

Τα δεδομένα που χρησιμοποιούνται για την παρούσα εργασία έχουν ληφθεί από ψυχοθεραπευτικές συνεδρίες και αποτελούν τον πλήρη διάλογο που διεξάγεται κατά τη διάρκεια της συνεδρίας μεταξύ θεραπευτή και θεραπευόμενου. Επομένως, υπάρχει η πληροφόρηση του ομιλητή κάθε πρότασης. Στα συστήματα που έχουμε προτείνει μέχρι τώρα θεωρούμε το διάλογο σαν μια διαδοχή προτάσεων χωρίς να δίνουμε την πληροφορία του ομιλητή της κάθε πρότασης. Έτσι, θεωρούμε ότι είναι σημαντικό να αναπτύξουμε συστήματα που θα λαμβάνουν υπόψιν το speaker role κάθε πρότασης ώστε να μπορούν να συλλάβουν το συνολικό context του κειμένου καθώς και την αλληλεπίδραση μεταξύ των ομιλιτών. Συγκεκριμένα, τα συστήματα αυτά θα πρέπει να αξιολογούν όχι μόνο την τελευταία πρόταση που ειπώθηκε κατά τη διάρκεια της συνεδρίας αλλά και την τελευταία πρόταση που ειπώθηκε από τον εκάστοτε ομιλητή.

Υπάρχουν διάφορα συστήματα διαλόγου που έχουν προταθεί στη βιβλιογραφία. Ορισμένα από αυτά βασίζονται σε αναδρομικά νευρωνικά δίκτυα και συγκεκριμένα χρησιμοποιούν διαφορετικούς encoders για κάθε speaker [43, 82, 45]. Στην περίπτωση της αναγνώρισης συναισθήματος από διάλογο, είναι σημαντικό τα συστήματα να μπορούν να αναγνωρίσουν τα υποκείμενα συναισθήματα στη συζήτηση. Τέτοια μοντέλα μπορούν να πραγματοποιούν utterance-level emotion prediction [14], ανίχνευση διαταραχών διάθεσης μέσω της ενσωμάτωσης χαρακτηριστικών από την ομιλία (lexical και prosodic features) [42] καθώς και χαρακτηριστικών του διαλόγου [16, 84].

Στην παρούσα εργασία εξετάζουμε το πρόβλημα της μοντελοποίησης του διαλόγου από θεραπευτικές συνεδρίες και προτείνουμε αρχιτεκτονικές που θεωρούνται ότι θα μπορούσαν να βοηθήσουν σε αυτό το σκοπό. Στο σχήμα 3 παρουσιάζουμε μια τέτοια αρχιτεκτονική. Χρησιμοποιούμε ένα κοινό word-level encoder για το σύνολο των προτάσεων στο κείμενο και ακολούθως υλοποιούμε δύο turn-level encoders, που αντιστοιχούν στα δύο speaker roles (therapist/client). Έτσι, στο turn-level encoding λαμβάνονται υπόψιν κάθε φορά και οι παρελθοντικές προτάσεις που έχει πει ο κάθε speaker. Το συγκεκριμένο μοντέλο εκπαιδεύτηκε με το General Psychotherapy Corpus και έδωσε F1-score 68.9, το οποίο είναι αρκετά κοντινό με τα αποτελέσματα των υπολοίπων απλών ιεραρχικών μοντέλων που υλοποιήσαμε. Το αποτέλεσμα αυτό είναι επομένως αρκετά ενθαρρυντικό και φανερώνει ότι η περαιτέρω μοντελοποίηση του διαλόγου κατά τη διάρκεια της συνεδρίας μπορεί να αποφέρει πολύ καλύτερα αποτελέσματα στην διάγνωση κατάθλιψης.



Σχήμα 3: Hierarchical Model with Attentional Conditioning

Λέξεις Κλειδιά

ανίχνευση κατάθλιψης, θεραπευτικές συνεδρίες, επεξεργασία φυσικής γλώσσας, μηχανική μάθηση, αναδρομικά νευρωνικά δίκτυα, ιεραρχικό δίκτυο, μηχανισμός προσοχής

Abstract

Depression is a common and serious medical illness that affects the way affected people feel, think and act. It can lead to a variety of physical, social and emotional problems and can decrease people's ability to work and function. Fortunately, there is a plethora of available medications able to treat depression. However, it is crucial that clinicians diagnose early the signs of the disorder and prescribe the suitable treatment. The diagnosis procedure is not an easy task. In many cases, the symptoms may not be indicative and thus complicate the process. In this work, we explore the task of Depression Detection from transcribed therapy sessions and propose models and methods that address the task in hand.

Firstly, we consider the transcribed dialogues derived from the sessions between a therapist and a client. In order to leverage the hierarchical structure of documents, we propose a Hierarchical Attention Network and perform document classification. Our task is a binary classification task, so the model decides upon the depression status of clients. However, therapy sessions provide valuable insights of the cognitive and behavioral functioning of clients, which can not be easily captured through processing of the raw document. Indeed, our analysis shows that depressed people use affective language to a greated extent than not-depressed. Therefore, we leverage behavioral and psycholinguistic cues of the client and therapist language to enhance the performance of our models. In particular, we integrate prior word-level psycholinguistic knowledge extracted from affective lexica, into the network architectures. The integration is performed into the self-attention mechanism of the system, which can force higher values for attention weights corresponding to salient affective words. In addition, we also incorporate the summary attributed to each session into the proposed architectures. Our approach improves the performance of the proposed architecture in the General Psychotherapy Corpus and the DAIC-WoZ 2017 depression datasets, achieving state-of-the-art 71.6 and 70.3 using the test set, F1-scores respectively.

Next, we experiment with the problem of dialogue modeling in the context of detecting depression. We present related work for the task of emotion recognition in conversations and propose a model that introduces the speaker-role of utterances into the encoding process. The resulting performance is comparable to the baseline network. Finally, we propose future directions for incorporating the inter-speaker dependencies.

Overall, our work addresses the task of depression detection from therapy sessions and proposes methods that improve the results of our networks, especially in the case we have small amount of data. This fact results in high performing models and improved robustness across two corpora. This work is summarized into the [81] research paper, which is submitted to the Interspeech 2020 conference.

Keywords

depression detection, clinical interviews, natural language processing, machine learning, recurrent neural networks, hierarchical attention networks, affective lexica

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Alexandros Potamianos, for giving me the opportunity to join the NTUA-SLP group and for guiding me throughout the period I conducted my thesis. His support pushed me every time to become better.

Secondly, I would like to thank Giorgos Paraskevopoulos for showing me the way and teaching me everything he knew. His advise and perspectives were precious during all this time. I also thank the rest labmates of the NTUA-SLP group, for the fruitful talks and the memories we shared.

Next, I would like to thank psychologists Anastasios Panopoulos and Evangelia Prassopoulou for their perspectives and psychological insights into Depressive Disorder, which helped me understand deeper the task in hand.

Last but not least, I would like to thank my friends, in and out of NTUA, for the precious moments and their support. Finally, a big thank you to my family for their invaluable support and contributions throughout the years, which helped my achieve my goals.

Danai Xezonaki
July 2020

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Depression is a serious mood disorder that affects the way people think and behave. According to WHO [55], it is estimated that over 300 million people suffer from depression, which corresponds to the 4.4% of the world's population. Indicative symptoms of depression can be the loss of interest in everyday activities, sleeping and eating disorders, feelings of worthlessness, sadness and exhaustion, or even thoughts of suicide [1]. WHO also states that over 800,000 suicide deaths are reported each year due to depression, while for 15-29-year-old people, it is the leading factor of death.

Depression detection is the problem of identifying signs of depression in individuals. These signs might be identified in peoples' speech, facial expressions and in the use of language. However, the process of diagnosing depression is a difficult and complicated task, as depression occurs in a different way across people. The symptoms may vary as well as their severity and duration. For instance, some depressed people might be angry and anxious, while others unable to make decisions. To this end, clinicians have to determine how many symptoms of depression patients have, for how long and how much they interfere with their ability to live life normally. This procedure is thus subjective to some extent and the need of developing intelligent systems in order to help decision making and depression studying is considered imperative.

The growing amount of available online data opens opportunities to perform data driven analyses and develop computational algorithms to assist specialists in the field of psychology, study depression and refine clinical methods and protocols. In addition, through the analysis of therapy sessions, valuable insights can be provided into the cognitive and emotional state of depressed people, and thus help the diagnosing procedure make more reliable predictions. In particular, depression diagnosis is a time-consuming task. Doctors have to simultaneously consider patients' symptoms and medical history, similar disorders and possible treatment. AI systems, however, are capable of processing large amounts of data in a short time. To this end, provided the therapy sessions, systems can give a fast recommendation of the diagnosis, that doctors will then evaluate and use. Apart from providing only the transcribed dialogue, AI systems could also record the whole therapy session and thus have additionaly available the audio and video format of the session. In this case, the diagnosis would be definitely more accurate since models would leverage not only the linguistic information but also the vocal cues and facial expressions, which contribute majorly in the task in hand.

Moreover, depression detection systems can be used effectively for helping patients handle mental health issues. They could be integrated into therapist chatbots, which engage with patients and evaluate the severity of depression. Especially in the case patients are not able to visit a doctor or have to wait a long time for an appointment, these health assistants could propose mental exercises and give self-care advices to patients.

Developing intelligent systems that are able to understand the behavioral functioning of individuals and predict their depression severity status, is a challenging task with a lot to offer to the medical community. However, there exist limitations like the limited number of labeled observations due to medical confidentiality, and the difficulty of models to discriminate cases based on emotional factors. This objective has become an active area of research in the fields of Artificial Intelligence and Natural Language Processing. With the advanced technology used in Machine Learning and Deep Learning, NLP systems are capable of achieving remarkable results in mental health-related tasks whose accuracy can be comparable to an average mental health specialist.

In this work, we aim to research methods and techniques that can help algorithms deal effectively with the task of depression detection. To this end, we present an analysis of therapy sessions and provide the algorithms with psychological and linguistic information so as to discriminate the two classes easier and help models generalize more effectively from unseen data.

1.2 Research Contributions

Depression detection is a multifactorial process. In some cases, people may experience the symptoms of depression without suffering from it. On the other hand, patients may experience less symptoms than the expected but medical treatment is considered necessary due to their severity. Therefore, clinicians do not focus only on the symptoms of a client in order to diagnose a disorder but on the whole image. This procedure is difficult to be learned by an artificial intelligence system. Studies have examined depression from audio and video recordings [20, 18, 71, 19] and questionnaire responses [49, 75]. In the case of natural language processing it is crucial that the models learn to discriminate emotionally charged words. However, the available data for diagnosing mental health illnesses are limited, due to confidential reasons. It is thus important to propose methods and systems capable of making right predictions, given a limited number of observations.
In the context of this thesis, we will be looking into methods and techniques for detecting depression from transcribed clinical interviews. In particular, our key contributions are the following:

- We propose a novel model for depression detection by incorporating existing affective information in the proposed models and we show through our results that it improves the performance of the proposed networks, especially in the case we have small amount of data.
- We conduct an analysis in the context of depression on the General Psychotherapy Corpus [1].

[1] http://alexanderstreet.com

- We present state-of-the-art results for the binary depression classification task. Our approach results in high performing models and increased robustness across two corpora.

1.3 Thesis Outline

In chapter 2, we provide the theoretical knowledge required for the subsequent chapters. We introduce the fields of machine learning and natural language processing, along with common methods that are used in this work. Next, we present neural networks and especially the types of models that are used in the thesis. Subsequently, we inform the reader about the task of document classification and some popular approaches. Lastly, we provide insights into mood disorders, the clinical diagnosis procedure and the analysis of therapy sessions.

In chapter 3, we present a detailed analysis of the work *Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews* [81]. We provide a thorough introduction into the problem of depression detection and an analysis of the corpora used. Moreover, we introduce and explain the proposed models and finally present and discuss the obtained results.

In chapter 4, we state the problem of dialogue modeling in the context of depression detection from transcribed clinical interviews. We provide a task-specific literature review and present the results of our baseline model. We also propose alternatives for further improvement.

In chapter 5, we conclude the present work and discuss about future directions.

Chapter 2

Theoretical Background

In this chapter, we present and explain the theoretical knowledge needed in order to understand the methods that were used, to address the task of depression detection. This knowledge concerns the broad sector of Machine Learning and Deep Learning, the required theoretical background of depressive disorder and also techniques used for document classification.

As our work concerns the diagnosis from the transcribed therapy session dialogues, we firstly introduce in 2.1the field of Natural Language Processing and discuss the applications in sentiment and emotion recognition along with the evaluation metrics. Next, we present the sector of Machine Learning, in 2.2 and discuss about learning methods and techniques used to represent text data. In 2.3, we introduce the field of Deep Neural Networks and especially present Recurrent Neural Networks and Attention Mechanisms, in 2.4, that are mostly used in this work. Subsequently, we provide the knowledge of classification models with emphasis on Support Vector Machines and Logistic Regression, and also discuss about loss functions. Due to the document format of session therapies, methods for data representation and classification in the case of documents are described in 2.5. Finally, in 2.6, we provide an analysis of mood disorders and focus on the on a thorough presentation of the depressive disorders.

2.1 Natural Language Processing

2.1.1 Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that concerns the interaction between computers and human language. NLP is the technology used to help computers understand, interpret and even generate natural language data.

NLP is a challenging and quite demanding task. This is due to the nature of human language that makes the task difficult. The rules on which languages are based are not easy to be understood by machines. Some of these rules can be low-leveled, yet some others are more abstract. To this end, in order to understand the language, machines have to know not only the words but also the whole meaning behind them.

Taking into consideration the large corpora that computers can handle, as well as their ability to process information quickly and in an unbiased way, it is obvious that NLP consists a significant method towards extracting valuable hidden information from text bodies. Computers can now communicate with humans in their own language and thus hear speech, understand it, measure sentiment and determine which parts are important. In general terms, the aim of NLP is to split language into shorter units and explore how these pieces work together to create meaning.

Natural Language Processing provides implementations for any task that utilizes text. The most common applications of NLP are:

- **Information Retrieval**: the activity of searching for and obtaining information resources from a database or the Web that are relevant to a given user query.
- **Question Answering**: a field which concerns building systems that automatically answer questions posed by humans in a natural language.
- **Machine Translation**: the process by which a computer software translates text or speech from one human language to another.
- **Document Summarization**: the process of choosing the most important information from a document and produce a summary of it.
- **Natural Language Generation**: the use of AI software to produce written or spoken data based on a text corpus.
- **Information Extraction**: it concerns the recognition and extraction of key-element information from text bodies.

2.1.2 Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify, analyze and extract subjective information (people's opinions, sentiments, emotions and evaluations) towards entities as products, businesses, organizations, topics or other people. It detects polarity within a given text, which means positive, neutral or negative feelings.

As far as businesses and organizations are concerned, sentiment analysis is an essential process since it helps them understand the opinions and emotions of customers and thus understand the social sentiment of their brand, service or product. By automatically analyzing customer feedback, which can be in the form of survey responses or social media conversations, brands are capable of listening to the opinions of their customers and try to further improve their products or services so as to meet the needs of customers. Moreover, with the explosive growth of social media, it is no longer necessary to conduct surveys and opinion polls in order to gather public opinion. The huge amount of online data makes it much easier for organizations to collect all the necessary information. Therefore, automatic sentiment analysis is highly needed. There are different levels on which sentiment analysis can be applied in text bodies.

2.1.3 Emotion Recognition

Emotion Recognition is the process of identifying human emotion through the expression, such as fear, anger, happiness, sadness, surprise and disgust, and is closely related to

Sentiment Analysis. Emotions can be detected in human non-verbal cues, such as facial expressions from video, on spoken expressions from audio and on written expressions from text. Hence, extracting and understanding emotion is of high importance to the interaction between human and machine communication. Although it is a recent field of research, it has gained much popularity due to the ability of machines to make decisions based on objective rules, in contrast to humans who vary widely in their accuracy at recognizing emotions. In the case of conversations, emotion recognition extracts opinions between participants from conversational data which can be found on social media platforms.

2.1.4 Evaluation Metrics

Evaluation metrics are metrics used to measure the quality of the statistical or machine learning model. They provide models with feedback so as to improve until they reach a desirable performance. The performance is usually evaluated on the test set. The most common metrics used in classification tasks are accuracy and F1-score.

Accuracy. It is mostly used in binary classification tasks. Accuracy divides the number of true predictions by the total number of observations and outputs a percentage score. Its mathematical formula is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

where TP, TN, FP, FN are the number of true positive, true negative, false positive and false negative predictions.

F1-score. This metric is used for multi-class classification or for binary classification where the number of observations is not balanced across the two classes. F1-score is the harmonic mean of Precision and Recall, which are illustrated in Figure 2.1, and are defined as follows:

$$\begin{gathered} Precision = \frac{TP}{TP + FP} \\ Recall = \frac{TP}{TP + FN} \\ F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \end{gathered} \tag{2.2}$$

In the case we are dealing with a multi-class classification task, we need to combine the F1-score of each class into a single F1-score. This is done using the **macro-F1** metric, which is computed as the arithmetic mean of the N classes' F1-scores. Let $F1_i$ be the F1 score of the i-th class computed using Equation 2.2. Then, the F1-macro score is calculated as follows:

$$F1macro = \frac{\sum_{i=1}^{N} F1_i}{N}$$

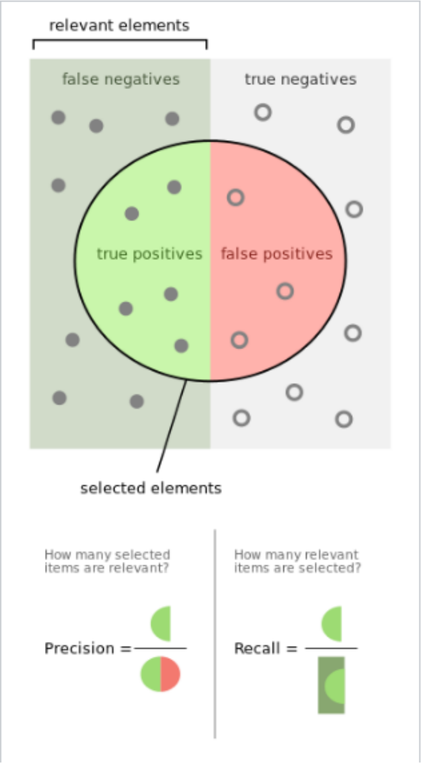


Figure 2.1: Precision and Recall. Source:wikipedia.org

2.2 Machine Learning

2.2.1 Introduction

Machine Learning is a subset of Artificial Intelligence (AI) that provides programs with the ability to learn from experience and generalize in order to make predictions without being explicitly programmed. Machine learning systems use sample data known as “training data” in order to build a mathematical model. Decisions are made based on previous computations that are next applied to unknown samples, referred to as “test data”. The primary aim of scientists is to allow computers to learn for themselves, automatically, so as to perform specific tasks without human intervention.

Machine learning algorithms are used in a wide variety of applications, including computer vision, speech recognition and email filtering. In these areas of study, scientists are not capable of developing traditional algorithms and computational methods to find solutions for the needed tasks. Machine learning is closely related to computational statistics, so it focuses on making predictions using computers. Therefore, such tasks otherwise infeasible to deal with can be handled using statistical models.

In Machine Learning, tasks are generally classified into two distinct categories. These are supervised learning, where training data are accompanied by desired outputs, known as “labels”, and unsupervised learning, where data are not provided with labels. In the first case, models learn the mapping between inputs and outputs, whereas in the second case models have to find themselves structure within the input data.

2.2.2 Machine Learning Methods

Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output. Input data are attributed *labels*, therefore each given sample is an input-output pair. A supervised learning algorithm uses this mapping to generalize from the training data and predict the output of unseen instances in the more accurate way.
Formally, given that X is an input data used for training, it is mapped to the desired output Y using a mapping function F:

$$Y = F(X)$$

Supervised learning models aim to approximate the mapping function so well that given unseen samples, they are capable of predicting the correct output label. The learning procedure stops when this approximation and thus model performance is satisfactory.
Tasks in supervised learning are distinguished into **classification tasks** and **regression tasks**. Classification is the task of identifying to which of a set of categories or classes an unseen observation belongs, given a training procedure of input samples that belong to the set of these classes. On the other hand, regression is the procedure of mapping an input sample to a continuous output value, such as an integer or a floating point value. It is basically a statistical approach to find the relationship between variables.

Unsupervised Learning

In contrast to supervised learning, unsupervised learning allows for modeling of probability densities over inputs, which are not attributed predefined label values. So, the aim of unsupervised learning is to identify undetected patterns in a dataset with no human supervision.
A typical method used for unsupervised tasks is **clustering**. In this case, a similarity metric is measured between input data in order to group them into clusters of similar samples. However, it is often difficult to know how many clusters should exist or how they should look.
Generative models are also a subclass of unsupervised learning models. They are given training data as input and new data are generated from the same distribution. Thus, models have to discover the nature of the input data in order to generate similar samples. This type of learning is unsupervised as there is no human intervention for the generation process. The only observable aspects are the generated samples.

2.2.3 Text Preprocessing

In NLP tasks, applications have to deal with a large amount of text data in general. It is thus crucial to transform text into a shape that can be easily processed by algorithms. Different steps should be followed in this direction.

Tokenization

Given a sequence of characters, tokenization is the procedure of splitting the sequence into subsequences which are part of the general vocabulary, called tokens, while discarding unnecessary characters. In the case of a sentence, tokenization splits the sentence into individual words and removes punctuation characters.

Punctuation and Tags Removal
Having splitted a sequence into words, it is necessary to remove characters that are not part of the tokens. Tags that can also be found in text documents should also be discarded as they do not contribute to the semantic content.

Stop Word Removal
Stop words are common and high frequency words as “I”, “a”, “the”, “an”, “and”, “is” which do not contain salient context. By removing stop words the dimensionality of data is reduced so it is easier to identify key words left in the corpus using feature extraction techniques. Stop word removal can be performed either by defining a list of words to be removed and iterating words in text in the chosen list or by using commonly available lists.

Stemming and Lemmatization
Stemming and Lemmatization are two key methods for text processing. Stemming is the process of reducing a word to its own stem through dropping unnecessary characters, usually a suffix. Related words are usually reduced to the same stem and thus are treated by algorithms as sysnonyms. For example, the words “fishing”, “fisher” and “fished” would be reduced to the word “fish”. The program that performs this procedure is called a stemmer. There are several publicly available stemming models, including Porter and Lancaster supported by the NLTK platform for Python.
Lemmatization, on the other hand, is the process of grouping together the inflected forms of a word, so they can be analysed as a single item. In the case of inflected forms of verbs, lemmatization replaces words with their base form. For instance, verbs “walking”, “walked”, “walks” would be reduced to the base form “walk”, which is called the *lemma* of the word. In order to find the base forms, lemmatization models need to have access to dictionary sources.
These two processed are quite similar to each other. However stemming is performed without knowledge of the context, whereas lemmatization discriminates between words which have different meanings depending on their part of speech.

2.2.4 Feature Engineering

In machine learning, a feature is an attribute that input data have, on which analysis or prediction is to be done. To this end, Feature Engineering is the process of extracting features from raw data using data mining techniques. There is no indicative way of how the extraction should be done, as it depends on the problem we are trying to solve. However, these exist some popular techniques that can be useful in different cases.

Imputation
Missing values is one of the most common problems arising during data preprocessing. These can be due to human errors or privacy concerns and they significantly affect the performance of machine learning models. The most simple solution is to remove the samples with missing values from the dataset. Otherwise, they can be replaced by real values attributed manually to samples, provided that it is considered as a sensible solution, or by the median of the rest available values for the same feature.

Handling Outliers
Outliers are considered to be values that surpass the standard deviation of a variable.

In order to deal with these values, they can be either removed or capped. However, capping can change the data distribution and thus affect the general performance of the model.

One-Hot Encoding
One-hot encoding is one of the most common encoding methods in machine learning and is the representation of categorical variables as binary vectors. The method creates a vector of flag columns for each value and assigns 0 or 1 to them.

Feature Extraction
In machine learning, features can be relevant, weakly relevant or irrelevant to the different tasks. Therefore, feature extraction derives a set of features from a given dataset which are considered to be informative and thus help models learn the input distribution easier and generalize more effectively. This method is also closely related to dimensionality reduction. The process of choosing which dimensions of the input data contain more relevant information is called feature selection and is very efficient especially in the case of a large number of input data, which are difficult to be processed.

2.2.5 Feature-wise transformations

In machine learning, a frequently used method is the integration of external knowledge into the neural architectures. Such approach is applicable to different domains, including visual question-answering [60], image recognition [31] and NLP [50, 12, 15]. There exist various techniques on fusing sources of information. In this work, we focus on a set of approaches known as feature-wise transformations [17]. In this case, the computation carried out by the model is modulated by the information extracted from the auxiliary source and the process is called *conditioning*. Such applications are effective in tasks where different modalities as video, language and audio have to be combined. The auxiliary knowledge can also be features in the form of prior information encoded in linguistic, emotion or sentiment lexica, as in our case. To this end, the raw input is processed in the context of the external information from the auxiliary input.
A typical conditioning method is the concatenation of the word-level external information to the input or to the hidden layers. One popular method is concatenation-based conditioning, as shown in Figure 2.2. In this case, the conditioning representation is concatenated to the input of all layers in the network.
In [50], the conditioning is performed on the network's attention mechanism. In particu-

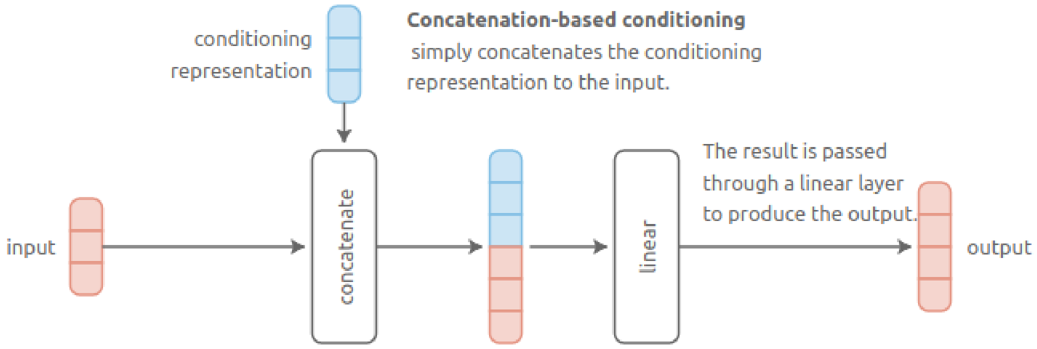


Figure 2.2: Concatenation-based conditioning. Source: [17]

lar, the self-attention mechanism is augmented and the attention weights of each sentence are conditioned on the corresponding word's prior knowledge. To this end, the attention layer is given as input a combination of the word representations, extracted from the word encoder, with the additional information of each word.

2.2.6 Word Embeddings

Word embeddings are a type of word representations that allow words with similar meaning to have a similar representation. Representations are usually numerical vectors containing tens or hundreds of dimensions. These vectors have occured after a learning procedure and are able to capture the context of a word in a document as well as the semantic and syntactic similarity between words. To this end, words with similar meaning are attributed close spatial positions. Word embeddings are largely used in NLP tasks, as they often lead to a better performance.

Word2Vec

Word2Vec [51] is one of the most popular representation techniques of document vocabulary. It is a two-layer neural network that is trained to construct linguistic context of words. It takes as input a text corpus and outputs a vector space, where each word in the corpus is assigned a vector in that space. The aim of Word2Vec is to have words with similar context closely located to each other in the vector space. A simple way of measuring the similarity of vectors is the calculation of cosine similarity. To this end, words with similar linguistic context tend to have a higher cosine similarity score and thus a smaller angle in the vector space. The vectors resulting from the Word2Vec model can be further fed into a deep neural network or used to detect similarities between words.
There exist two different network architectures that can be used for Word2Vec models. In the first way, context is used to predict a target word (CBOW) while in the second, a word is used to predict a target context (skip-gram). Both of the procedures are illustrated in Figure 2.3.

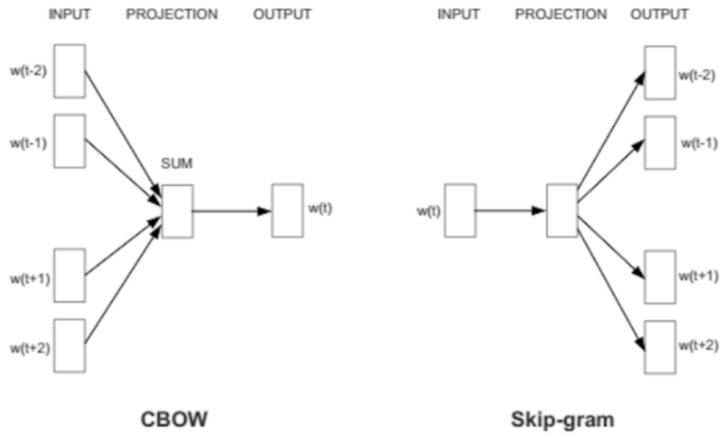


Figure 2.3: CBOW and Skip-gram model architectures.Source:pathmind.com

Common Bag of Words (CBOW)

The CBOW method takes the context of each word in the input and predicts a target word. Imagine having a corpus of N words. We use their one-hot encoding, which is a zero-valued vector of the same length as the vocabulary, except for the index that corresponds to the word we want to represent, which is valued as 1. The one-hot vector is then fed into a neural network. The hidden layer is a dense layer whose weights are the word embeddings (word vectors for the total of words in our corpus) and the output layer produces a probability for the target words in the corpus. CBOW can either take as input a single of multiple words as context vectors.

Skip-gram

Skip-gram looks like the oppposite process of CBOW. In this case, a single target word is fed to the network and the output produces N different probability distributions. It basically predicts the surrounding context words for each input target word. Skip-gram seems to perform better with small amount of data and is found to represent rare words well.

Global Vectors (GloVe)

GloVe is another popular word vector learning technique. While Word2Vec relies only on local context provided by the surroundings of a word, GloVe captures both local and global statistics of a corpus, in order to produce the vector representations. The basic idea is that the model is trained on the co-occurence word statistics, which indicate how frequently each pair of words is used in the given corpus. These frequencies are used as is it considered that they can encode some form of meaning.

2.3 Deep Neural Networks

2.3.1 Introduction to Deep Learning

Deep learning (DL) is part of the broader sector of Machine Learning and is a set of learning methods that imitate the workings of human brain in processing data. The basic elements of deep learning are artificial neural networks which are usually stacked in multiple levels and form different neural architectures. Algorithms in deep learning extract high-level features from raw data by propagating the input through the consecutive levels of such architectures. Each level serves as a function that learns to transform the input data into a representation. Deep learning has been applied to numerous fields of study, including computer vision, speech recognition, natural language processing, bioinformatics and medical image analysis. The capability of DL algorithms to find solutions to complex tasks usually surpasses the human performance.

2.3.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are computational models, inspired by the structure and functioning of neurons in the human brain. The basic component of an ANN are *neurons*, which generally model the neurons of the brain. An overview of a biological neuron is shown in Figure 2.4. Over 10 billion neurons exist in the human body and each of them is connected to several thousands of other neurons. The cell body of the neuron, (*soma*), processes the incoming activations and produces output activations. Neurons also send and receive activation from other neurons through their *axons* and *dendrites*, respectively. Through their *synapses*, they transmit signals to and from the rest of the connected neurons. In parallel to the biological structure, a typical example of an artificial neuron is depicted in Figure 2.5. Connections between artificial neurons are called edges and are responsible for transferring information from one neuron to the other connected ones. Each edge is assigned a weight (ω_1,ω_2,...,ω_m) that changes during the training procedure and thus it alters the strength of the information that needs to be transfered. The output of a neuron is computed by summing the multiplication of each input by the corresponding edge weight. If the final value of the output of the neuron is above a specific threshold, the information is transmitted to the rest neurons. The sum is then passed through a function assigned to the neuron, called activation function and generates an output. Artificial neurons are typically organized in layers. Each layer may have a different number of neurons and perform a different processing on the input data. The structure of an ANN is shown in Figure 2.6. As it can be observed, ANNs consist of hierarchies of layers. Inputs are fed into the network through the **input layer**. Subsequently, they are propagated through the **hidden layers** where they are processed and features are extracted. As the number of hidden layers increases, higher-level representations are constructed. In the case of multiple hidden layers, the network is referred to as a Deep Neural Network (DNN). Finally, the obtained representations pass through the **output layer**, where a decision is being made by the network.

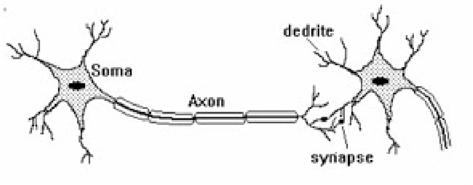


Figure 2.4: Overview of a biological neuron. Source:neuralfuzzy.blogspot.com

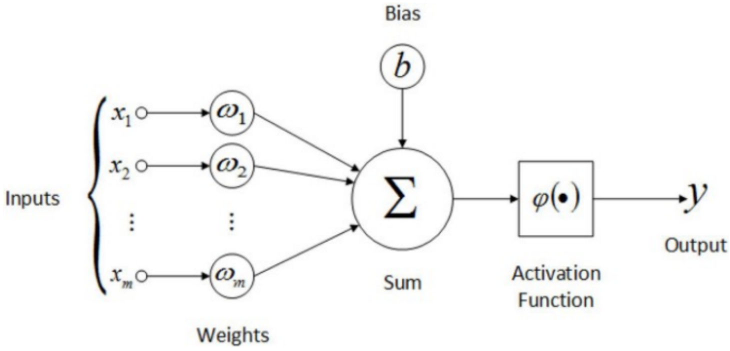


Figure 2.5: Example of an artificial neuron. Source:medium.com

Activation Function

In artificial neurons, the activation function is a mathematical equation that determines the output of the neuron. It serves as a mathematical gate between the neuron's input or

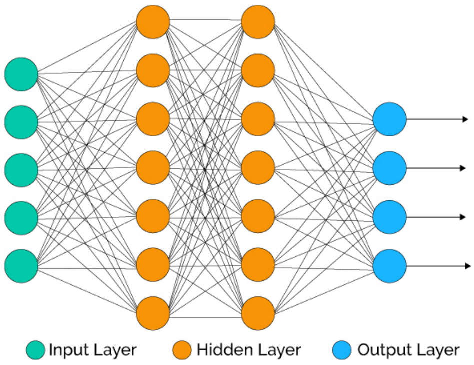


Figure 2.6: Overview of an artificial neuron network (ANN). Source: www.xenonstack.com

set of inputs, and the output that will be transmitted to the next layer. In its simplest form, it can be a binary function that turns the neuron on and off, depending on the input. It can also help normalize the output to a range between -1 and 1, or transform the input signals into output signals. Activation functions can be either linear or non-liner and each neuron in a network can have a different activation function. Some of the most popular functions are presented next.

Sigmoid Function

Sigmoid Function is shown in Figure 2.7. It is a bounded, differentiable, real function which is defined for all real input values by the formula:

$$S = \frac{e^x}{e^x + 1} \tag{2.3}$$

Its formula limits the output in the range between 0 and 1 and is thus used especially when a model has to predict the probability as an output.

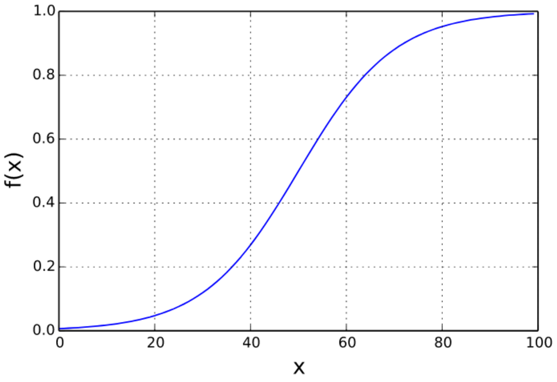


Figure 2.7: The sigmoid function.

Binary Step Function

Binary step function is a threshold-activation function. It is depicted in Figure 2.8. If the input value is above a threshold, the output is set to 1 so the neuron is activated and the signal is sent to the next layer without any transformation. Otherwise, the output is set to 0 and the neuron is deactivated. As the output takes only two values, the function does not support multi-class classification.

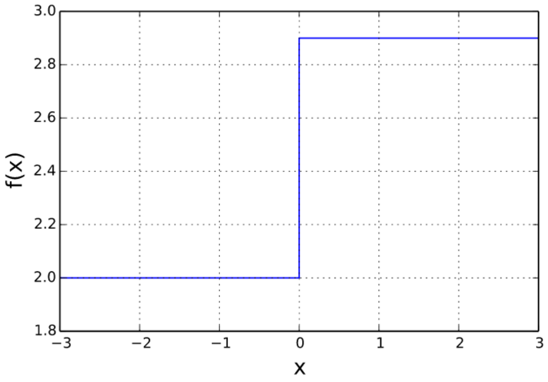


Figure 2.8: The Binary Step function.

Softmax Function

The softmax function takes as input a vector of real numbers and normalizes it into a probability distribution. The distribution consists of a probability value in the range between 0 and 1, for each input number, and the components add up to 1. Softmax function is largely used in neural networks, where it is needed to map the outputs of the networks to a probability distribution over multiple predicted classes. The function for each i-th value in the initial input vector follows this function:

$$f(x)_i = \frac{e^{x_i}}{\sum_{n=1}^{\infty} e^{x_n}} \tag{2.4}$$

The softmax function is illustrated in Figure 2.9.

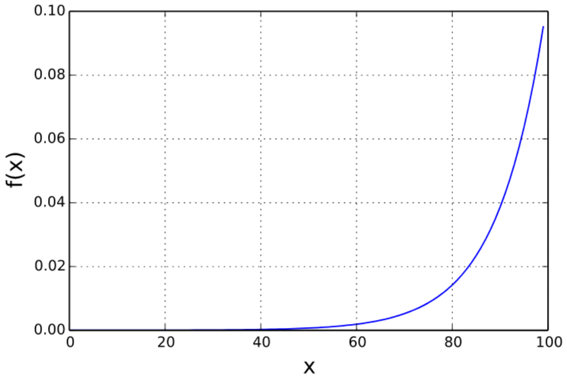


Figure 2.9: The softmax function.

Rectified Linear Unit (ReLU) Function

ReLU is a non-linear function and is the most commonly used activation function in neural networks. It is a simple calculation that returns the value of the input, or 0, if the input value is less than 0. Thus, it can be defined as:

$$f(x) = max(x, 0) \tag{2.5}$$

The finction is shown in Figure 2.10. It is linear for values greater than zero and non-linear for negative values, as they are always output as zero. Because of its functioning, it is also known as ramp function.

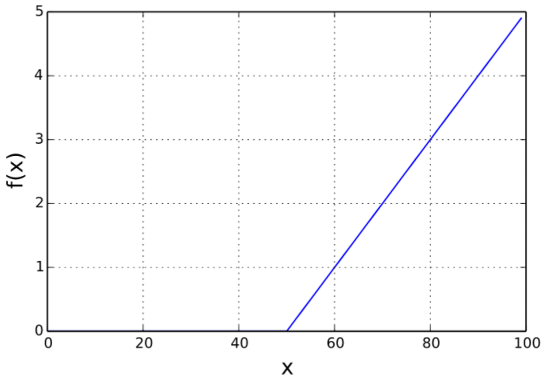


Figure 2.10: The ReLU function.

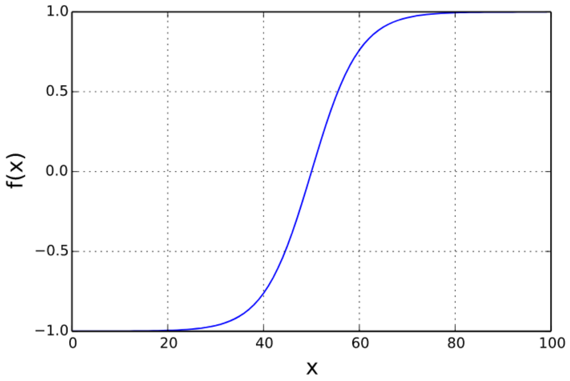


Figure 2.11: The tanh function.

Tanh/Hyperbolic Tangent Function

Tanh or Hyperbolic Tangent function is shown in Figure 2.11 and is defined as the ratio of the hyperbolic sine and hyperbolic cosine functions. Tanh follows the function:

$$tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.6}$$

As an activation function, tanh it is mostly used to model inputs that have strongly negative and positive values as it is zero-centered. Its outputs range between -1 and 1 and is basically a rescaled sigmoid function.

Regularization

When training neural networks, it is crucial to develop an algorithm that will perform well not only on the training data but also on the testing data. When an algorithm performs well on the train set but performs poorly on the test set, we say that the model is *overfitting* and it means that the model has learned too well the details and the noise of the

train set, which results in a poor performance on unseen data. To this end, we need to make modifications to the algorithm so that the model generalizes better.
The collection of methods and techniques used to reduce the error on the test set is known as regularization. These modifications though are not expected to reduce the training error. Regularization is often achieved by introducing constraints to the on the model parameters which serves as a penalty to the weights of nodes. Next, different regularization techniques are presented.

L2 regularization

L2 regularization introduces a new term in the loss function. This technique is also known as *weight decay* as it decreases the weight matrices to small or zero values. The process is based on the assumption that a network with small weights is simpler than a network with large weights. The loss function is therefore defined as:

$$J = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{ij}w_i)^2 + \lambda \sum_{j=1}^{m}((w_j)^2) \tag{2.7}$$

where λ is a model hyperparameter.

Dropout

Dropout training [72, 28] is one of the most popular regularization techniques especially for large neural networks, as it generally produces good results. The main idea is that in every iteration, the algorithm selects randomly some nodes and removes them along with their incoming and outgoing connections, by multiplying the output values by zero. In this way, each iteration consists of a different set of nodes that produce different outputs. The number of excluded nodes is defined by a dropout hyperparameter. Dropout can be applied to both the input and the hidden layers.

Early Stopping

When training models with a large number of training samples where overfitting might occur, it can be observed that although the train set error decreases, the validation set error firstly decreases and then starts to increase steadily. Therefore, it is crucial to stop the training when the validation set error takes its lower value. Models that perform early stopping save the optimal set of parameters obtained at the lowest validation error. A hyperparameter, called *patience*, is defined in models using early stopping so as to denote the number of epochs with no further improvement, after which the training will end.

Data Augmentation

Data Augmentation is another regularization technique mostly used in the case of a small amount of data. The simplest way to reduce overfitting is to increase the training data size. However, is many cases there might be limited labeled data available. To this end, data augmentation creates fake data and adds it to the training set.
This technique is frequently used in object recognition. When training a model on image data, the algorithm can produce new labeled images by shifting, scaling and adding noice to the existing ones. Data augmentation can also be used in the case of dialogue modelling, where larger turns can be splitted into smaller individual sentences.

Optimization

Optimization algorithms are methods used to change the weights of neural networks in order to minimize the output of the loss function.

Gradient Descent (GD) is one of the most popular optimization algorithms used in neural networks and especially in classification and regression tasks. It minimizes a loss function by computing the gradient of the function and by moving in the direction of the steepest gradient, as defined by the negative of the gradient. Through back propagation, the loss is transferred from the last layers to the first ones and thus the weights of each level are updated in a way that minimizes the loss function. Let $J(\theta)$ be the loss function, θ the model parameters and α small enough $\in R$. Then, the new parameters are updated as follows:

$$\theta = \theta - \alpha * \nabla J(\theta) \tag{2.8}$$

Gradient Descent is a widely popular optimization algorithm as it can be easily implemented and computed. However, if the dataset is very large, computing the gradient and changing the weights for the whole dataset can be time consuming. Finally, this algorithm may be also trapped at a local minimum.

Stochastic Gradient Descent [8] is a variant of GD. It replaces the actual calculation of the gradient for the whole dataset by an estimated gradient, computed on a subset of the data. Therefore, it updates the parameters more frequently than GD. This algorithm needs less time to converge to a minimum, which in large datasets is a significant aspect. Memory usage is also reduced compared to GD, as SGD does not store values of loss functions. However, due to the frequent weight alteration, it results in a high variance in model parameters.

Adam [37] is another optimization algorithm used to update the weights of a network. However it differs from the SGD algorithm. In particular, in contrast to the previously described optimizers, where the learning rate remains the same for all weight updates, Adam is an adaptive learning rate method, which means that it keeps a learning rate for every weight (parameter) in the network and separetely adaps them during the training procedure. It uses estimations of first and second moments of gradient to adapt the learning rate for each weight, where the n-th moment of a variable is defined as the expected value of that variable to the power of n. Adam is well suited for problems that are large in parameters, is computationally efficient and has little memory requirement.

Back Propagation

Back propagation [67, 40] is an algorithm used for training feedforward artificial neural networks, for supervised learning.

Every neural network can be illustrated as a directed graph where each neuron corresponds to a node and each weight to an edge. The back propagation algorithm computes the gradient of the loss function for each input-output pair, with respect to each weight individually using the chain rule, while catching intermediate results. The aim is to update the weights of the network in the optimal way after each computation of the loss function. The computation is performed for one layer of the network at a time, starting from the last layer and iterating backwards. The gradients computed can help us understand how quickly the loss function changes when the weights change and thus how well the network is performing. In this way, it is easier to fine tune the weights so as to further minimize the model's loss and improve the overall performance.

2.3.3 Recurrent Neural Networks

A Recurrent Neural Network (RNN) is a type of neural network that have an internal memory. It is recurrent as the output of every step is copied and sent back into the recurrent network and thus it is fed as input to the next step. Thus, the output of the current step depends on the past computations. This "memory" mechanism of RNNs is implemented using an internal **hidden layer** that produces a hidden state, which remembers all information about what has been calculated. The mechanism is very important for tasks as natural language generation and speech recognition, where the model needs to remember the previous words in the sentence so as to understand the context or generate a new word. It makes RNNs applicable to tasks that require to remember the history of previous inputs and outputs.

The unfolded equivalent structure of a recurrent neural network is depicted in Figure 2.12. As it is observed, the RNN takes the first input x_0 from the given sequence and produces an output h_0, which is known as the hidden state. In the next timestep, the network is given both the next input x_1 along with h_0. The procedure is continued for all timesteps in the given sequence. The hidden state at each timestep and the output are calculated as following:

$$h_t = f(W_h h_{t-1} + U_h x_t + b_h) \tag{2.9}$$

$$y_t = f(W_y h_t + b_y) \tag{2.10}$$

where f is the activation function and W_h, W_y, U_t, b_h, b_y the learning parameters of the model.

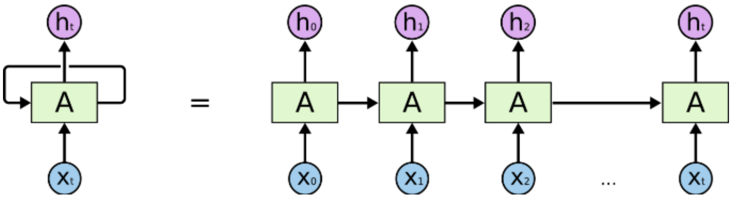


Figure 2.12: An unfolded recurrent neural network. Source: colah.github.io

Bi-directional RNN

Recurrent Neural Networks typically encode the input sentences in a forward manner, so that the hidden state of each timestep includes information about the previous timesteps. In Bi-directional RNNs, it is also possible to capture the information from the last timestep back to the first. Bi-RNNs combine two RNN layers in order to find the hidden state for each timestep. Their structure is illustrated in Figure 2.13. In particular, they compute the hidden state of the forward RNN as well as the corresponding hidden state of the backward RNN. Therefore, the hidden state for a given timestep is the concatenation of the two vectors.

Long Short-Term Memory

The Long Short-Term Memory (LSTM) [29] is a type of a recurrent neural network. This type of network is largely used for dealing with sequential data or data with temporal relationship, as it is capable of holding long term memories. In addition, when back propagating in a typical RNN, the computed gradients may vanish or explode. LSTMs overcome this problem by preserving long-distance dependencies between words and discarding words that are not important.

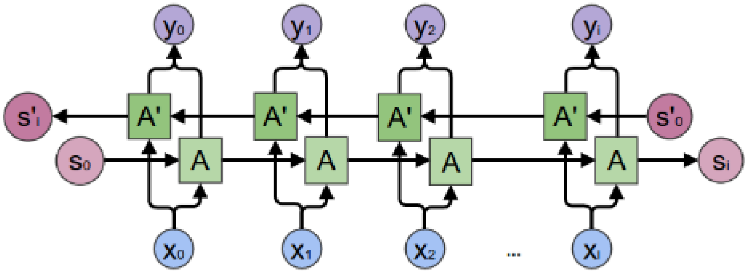


Figure 2.13: The overview of a bi-directional neural network. Source: colah.github.io

The internal architecture of an LSTM is depicted in Figure 2.14. It is composed of a cell state, an input gate, a forget gate and an output gate. These components organize the flow of information through the cell.

- **Input gate**. It controlls the extent to which a new value flows into the cell. The sigmoid function outputs the input in the range 0, 1 and thus decides which inputs to let in. The tanh function also outputs them in the range between -1, 1 so the multiplication of the two functions filters the importance of the current input.
- **Forget gate**. This gate decides which information from the previous timesteps should be kept or discarded. The new input along with the previous hidden state are passed through the sigmoid function. The output is a value between 0 and 1, meaning that it forgets zero-valued inputs and keeps inputs whose sigmoid-output is 1.
- **Output gate**. The output gate decides which information should be passed to the output. The sigmoid function once again stresses the important input information by mapping it between 0 and 1, and the tanh gives weightage to the newly modified cell state. The multiplication of these two outputs generates the new hidden state, which is carried to the next time step.
- **Cell state**. The existing cell state gets multiplied by the forget vector. Therefore, it is possible to drop values of the forget gate which are close to 0. The output vector of this multiplication is then added to the vector of the input gate. The cell state is now updated and contains the new values that the network considers as relevant.

The mathematical equations for the forward pass of an LSTM unit for a given input of vectors x_1, x_2, ... , x_n are the following:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2.11}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2.12}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{2.13}$$

$$u_t = tanh(W_u x_t + U_u h_{t-1} + b_u) \tag{2.14}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t) \tag{2.15}$$

$$h_t = o_t \odot tanh(c_t) \tag{2.16}$$

where matrices W and U contain the weights of the input and recurrent connections of forget, input and output gate (subscripts f, i, o respectively).

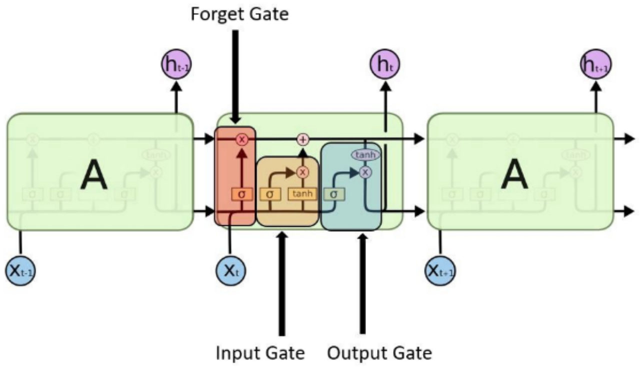


Figure 2.14: The LSTM cell. Source: towardsdatascience.com

Gated Recurrent Unit

The Gated Recurrent Unit (GRU) [10] is a variant of an LSTM, in the sense that it includes fewer parameters and a forget gate [23] yet it lacks an output gate and a cell state. The performance of the two types of networks is found to be generally equivalent. GRUs effectively solve the vanishing gradient problem using the update and reset gates, which decide the information that should pass to the output.

- **Update gate**. The update gate acts similar to the forget gate and input gate of an LSTM. The input x_t at timestep t is added to the hidden state from the previous timesteps and the result is passed through a sigmoid activation function. The result is thus squashed between 0 and 1. To this end, the update gate determines how much of the previous information needs to be passed along to the future.
- **Reset gate**. The reset gate determines how much of the past information should the model forget. The procedure is the same as the update gate, in the sense that the current input is added to the previous hidden state.

The equations that describe the function of a GRU are the following:

$$z_t = \sigma(W_z x_t + U_z x_t + b_z) \tag{2.17}$$

$$r_t = \sigma(W_r x_t + U_r x_t + b_r) \tag{2.18}$$

$$h'_t = tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{2.19}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h'_t \tag{2.20}$$

where matrices W and U contain the weights of the input and recurrent connections of the network.

2.3.4 Attention Mechanisms

Attention [7] is a mechanism that was developed to improve the performance of encoder-decoder RNN models in machine translation. However, it is currently used in a wide range

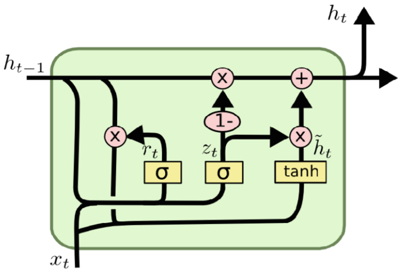


Figure 2.15: The GRU cell. Source: github.com/roomylee/

of applications as image captioning and dialogue modeling. The key contribution of the mechanism is that it helps models to direct their focus and pay greater attention to certain factors when processing the data. It is proposed as a solution to any sequence model that handles input with temporal dependencies, as long sentences, where the model has to look back at previous states. Therefore, instead of discarding the intermediate states of an encoder and using only its final states for the decoder as in a traditional Seq2Seq model [39], attention mechanism develops a context vector by utilizing all the intermediate encoder states.

Attention mechanism produces a context vector following a specific procedure. First of all, the encoder states h_1, h_2, ..., h_N pass through an attention function f, which learns to assign higher scores to the states of the encoder that need to be paid most of the attention and outputs the scores s_1, s_2, ..., s_N. Next, the softmax function is applied so as to get a probabilistic interpretation of the attention weights. The resulting scores e_1, e_2, ..., e_N define how much of each hidden state should be considered for each output and they must sum up to 1.Finally, the context vector is computed as the sum of the hidden states of the input sequence weighted by the attention scores. The corresponding equations are:

$$s_t = f(h_t, h_i) \tag{2.21}$$

$$e_t = softmax(s_t) \tag{2.22}$$

$$c_t = \sum_t e_{ti} h_i \tag{2.23}$$

where the softmax function is given by equation 2.4

Self-Attention

Self-attention [41], also known as intra-attention, is an attention mechanism Attention can be also applied in a single sentence when there is no additional information, by allowing it to attend to itself using self-attention. In this case, we use a feed forward neural network with tanh activation which is described by the equation:

$$f(h_i) = v_a^T tanh(W_a h_i) \tag{2.24}$$

where W_a and v_a are learnable attention parameters. The vector representation is given by equation 2.23. The self-attention mechaninsm is illustrated in Figure 2.16.

Multi-Head Attention

Multi-head attention is an another attention architecture, which consists of several attention layers running in parallel. This mechanism allows the model to jointly attend

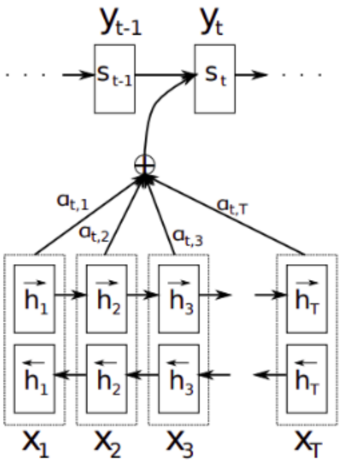

y_{t-1}
y_t
s_{t-1}
s_t
$a_{t,1}$
$a_{t,2}$
$a_{t,3}$
$a_{t,T}$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_T$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_T$
X_1 X_2 X_3 X_T

Figure 2.16: The self-attention mechanism. Source: [83]

to information from different representation subspaces at different positions. It is found that multi-head attention works better than single-head, as it applies the usual attention mechanism to multiple chucks in parallel, and then concatenates the results.

Hierarchical Attention
A variant of the classic attention mechanism is the hierarchical attention, which can be effectively applied on various levels of a network. A typical such network is depicted in Figure 2.17. Supposing that we are performing document classification, the input has to be text data and the network consists of two levels. Leveraging the hierarchical nature of documents, the first level of the network processes the words of the sentences and the second level processes the sentences which form the document. The attention mechanism at the two stages states which sentences are important for classifying the document and which words are salient in each sentence.

2.4 Classification Models

2.4.1 Introduction

Classification is the process of identifying to which of a set of categories a new observation belongs, as far as training sets containing observations are concerned. In machine learning, this procedure refers to a predictive modeling task where a label is predicted for each observation in the dataset. The algorithm that implements classification is known as a *classifier*.
There exist various classification algorithms for modeling classification tasks. Yet, not all classifiers can be successfully applied to all tasks, and vice versa. To this end, it is recommended to conduct experiments and discover which method results in the best performance for a given task.
One of the most popular metric for evaluating the performance of a model is *accuracy*. Classification accuracy calculates the percentage of the samples with accurately predicted labels in the total number of samples.

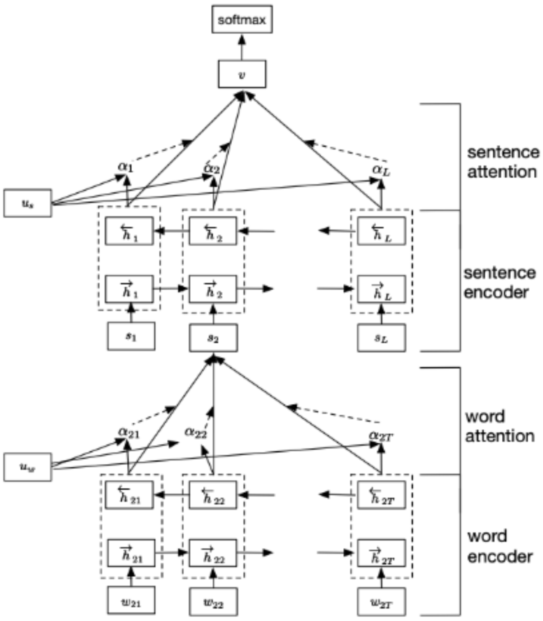


Figure 2.17: A hierarchical attention network. Source: [83]

2.4.2 Support Vector Machines

Support Vector Machines (SVM) are supervised learning models that serve as binary linear classifiers. Suppose given some data points in the N-dimensional space belonging to one of two different classes. An SVM model has to find a (N-1)-dimensional hyperplane that can separate the data points of each class by a clear gap. This hyperplane is referred to as a classifier. Different hyperplanes may separate the two classes for a given set of examples. The objective is to find the decision plane that has the maximum margin between the samples of each category, which means that it has the maximum distance from the nearest data point on each side. In this way, it is likely that future examples will be classified in the correct class.

Suppose we have a dataset with M input vectors x_1, x_2, ..., x_M with the corresponding labels y_1, y_2, ..., y_M. Let f be the function that classifies samples:

$$f(x) = w^T x + b \tag{2.25}$$

Then the result of f should be $f(x_i) > 0$ for x_i belonging to the first class and $f(x_i) < 0$ for x_i belonging to the second class. As shown in Figure 2.18, taking as example the 2-dimensional space, we want to find the separating line for which the minimum distance between the two classes is as wide as possible. The distance between a point x_i and the separator is $\frac{f(x_i)}{\|w\|}$. In order to maximize the distance $\frac{1}{\|w\|}$, we need to minimize the norm $\|w\|$ subject to $y_i(wx_i - b_i) > 1$, for i between 1 and M. In addition, it often happens that the given data points are not linearly separable in the space. For this reason, SVM models can also perform non-linear classification by mapping the input vectors to a higher-dimensional space, where it is more likely that they can be linearly separable. This mapping is implemented using a kernel function $k(x, y)$. Some common kernel functions are:

- Polynomial kernel: $k(x_i, x_j) = (x_i \cdot x_j + 1)^d$

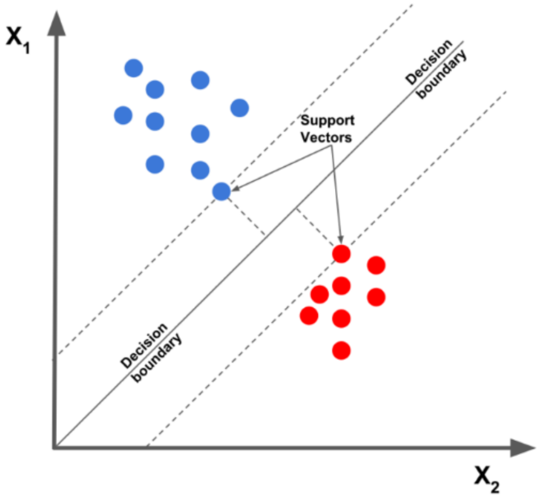


Figure 2.18: The maximal margin classifier. Source: www.learnopencv.com

- Gaussian radial basis kernel: $k(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$, for γ>0
- Hyperbolic tangent kernel: $k(x_i, x_j) = tanh(kx_i \cdot x_j + c)^d$

2.4.3 Logistic Regression

In machine learning, logistic regression (LR) is an analysis conducted to describe data and examine the relationship between one dependent binary variable and one or more independent. The dependent variable can take only 0 or 1 as values, representing its participation in one of two possible classes. LR is thus used to model the probability that an observation belongs to a specific class or not. The activation function of LR for a given vector x is a sigmoid function and is defined as follows:

$$\sigma(w^T x) = \frac{1}{1 + e^{-w^T x}} \tag{2.26}$$

Mathematically, LR computes the linear regression function:

$$l = log(\frac{p}{1-p}) \tag{2.27}$$

where p is the probability P(Y=1) of the event Y=1.
Logistic regression is frequently used in tasks as a solid baseline, especially for NLP tasks. An indicative task where LR is applied is the classification of emails as spam (1) or not spam (0). In the case of multi-class classification, LR can be also performed to each pair of classes and treat them as individual classification tasks.

2.4.4 Loss Function

A loss function or cost function is a function that maps a value of one or more variables onto a real number. In other words, it maps decisions to their associated costs. In machine

learning, it is used to estimate how well the algorithms can model the given data. When optimization problems are concerned, algorithms try to minimize the output of the loss function. Especially in classification tasks, the meaning of the loss function is the penalty for an incorrectly classified observation.
There exist several popular loss functions. Not all of them fit to the total of possible tasks. Choosing the right loss function may be subject to the type of the machine learning task or the ease of calculating the derivatives. Generally, they are classified into two groups, depending on the type of task we are facing (regression losses/ classification losses). In this work we are dealing with the problem of depression detection which is a classification task, so we will present next some popular loss functions used for classification.

Cross Entropy Loss

Cross entropy loss (CEL) measures the performance of a classification model, whose output is a probability with value between 0 and 1. It takes the true distribution and the estimated distribution and computes the cross entropy between the two as:

$$CE = -\sum_{i}^{c} x_i log(y_i) \tag{2.28}$$

where x is the true label (0 or 1), y is the predicted probability that an observation belongs to a specific class and c in the number of classes. An activation function has been applied to the scores before computing the CE Loss.
For binary classification problems, each observation in the given data set has a known class label with probability 1 and probability 0 for the rest of the labels. A model is used to estimate the probability that the observation belongs to each of these classes. Therefore, the CELoss increases as the estimated probability diverges from the actual label. Hence, the CE Loss is computed as:

$$CE = -xlog(y) - (1-x)log(1-y) \tag{2.29}$$

For example, as shown in Figure 2.19, predicting a probability near 0 when the actual label is 1 results in a high loss value. However, as the predicted probability approaches 1, log loss decreases.

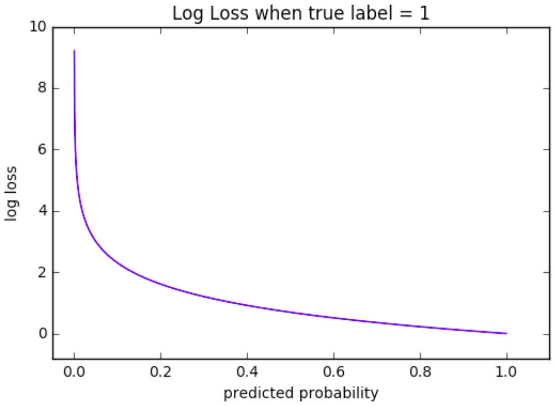


Figure 2.19: Log loss for predicted probabilities when true label is 1.

Binary Cross Entropy Loss

It is a sigmoid activation with a Cross Entropy Loss. It is used for multi-label classification as the output loss computed for class is independent of the rest classes. To this end, the probability of an observation belonging to a specific class is independent of the probability of belonging to another as well. The mathematical formula for binary classification is thus computed as:

$$CE = -xlog(f(y)) - (1-x)log(1-f(y)) \tag{2.30}$$

where f is the sigmoid function $f(y) = \frac{1}{1+e^{-y}}$.

2.5 Document Classification

2.5.1 Introduction

Document Classification is an example of a Machine Learning task relevant to Natural Language Processing. It is the act of asigning one or more labels or categories to a text document according to its content or according to their attributes, such as document type or author. It is used so as to easily sort and manage text-based documents as emails, articles or survey responses. This process can be either done manually, as in library science, or automatically, using classification algorithms. Both methods have their advantages and disadvantages. On the one hand, humans can decide which categories to use and have a greater control over the process. On the other hand, when dealing with a large amount of documents, automatic classification methods are definitely much faster and do not change their criteria over time.

Text classification is often confused with document classification, yet they are slightly different terms. It refers to performing an analysis on text-based documents. However, it can be applied to a subset of the total document as to a single paragraph or a sentence. There are different learning approaches to document classification.

- **Supervised**. In this method, humans have to manually assign labels to documents first before asking the model to do so. Based on these examples, the model will subsequently learn to make associations between text and expected tags in unseen documents.
- **Unsupervised**. In unsupervised learning, models learn to classify documents into categories with no human intervention. The classification is conducted without reference to external information and thus classifiers group together documents with similar words or sentences.
- **Rule-based**. The method is based on linguistic, morphologic semantic or syntactic rules that define each classification category and give instructions to models. Following these patterns, models then automatically tag the texts.

2.5.2 Feature Engineering

Given a dataset of text documents, we need to transform the raw data into feature vectors which will be used for classification. Next, some popular methods for feature extraction are presented.

TF-IDF
Term frequency-Inverse document frequency (TF-IDF) is a term-weighting scheme that reflects how important role plays a word in a document. The result is obtained through multiplying the number of times a word appears in a document (TF) and the inverse document frequency (IDF) of the word, across a corpus of documents. TF-IDF increases proportionally to the number of occurencies of a word in a document and is offset by the number of documents that contain the word. In this way, words that are popular across documents, even if they appear many times within a document, are not important to the content of the text. On the other hand, words that appear in some documents only, indicate that they might contain salient information. TF and IDF are presented next.

- **Term Frequency (TF)**. It refers to the frequency of a word. It can be calculated by counting the number of its occurencies in a document.
- **Inverse Document Frequency (IDF)**. This metric shows how common or rare is a word across documents. It is computed by dividing the total number of documents in a corpus with the number of documents that contain a specific word, and then calculating the algorithm of the result. Its value actually "penalizes" a word for appearing in many documents.

Overall, the TF-IDF metric for a word (t) in the document (d) which is part of a set of documents (D), is computed using the following formula:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{2.31}$$

However, there exist variant techniques for applying TF-IDF method that use weighting schemes in the computation of TF and IDF.

1. TF grows linearly as the number of word occurences in documents increases and thus it is difficult to measure the relevance between documents. For this reason, TF is often replaced by the value of 1+Log(TF), as Log function imposes a logarithmic growth and thus thresholds the value of TF.
2. Another variant is to compute the cosine similarity of the document vector and result in a unit vector. In this way, the different length of each document vector will not affect the final computation.
3. In other cases, we can also take into consideration the document length and thus standardize TF value by dividing with the maximum TF in that document. By doing so, we incorporate the difference between short and long documents, as the frequency of words in a corpus is more discriminating than just their number of occurences.

Word Embeddings
As presented in section 2.3.5, word embeddings are a form of representing words using a dense vector representation. The position of representations in the vector space is either learned by training a model on a corpus of documents, or can be obtained through pre-trained word embeddings. In the second case, sentences can be tokenized and split into individual words using a tokenizer, and then each word in the document can be represented by the predefined word embedding. These vectors form the final document representation

that can be fed into a classifier.

NLP based features
In document classification, it is also possible to extract text based features in order to improve the performance of classification models. These features can be statistics regarding:

- the total number of characters, words or sentences in a document
- the number of words as far as their part-of-speech is concerned (verbs, adjectives, nouns, pronouns)
- the punctuation count in the document
- the average word length of the title attributed to a document

Topic Modeling
Topic Modeling is a statistical model that discovers abstract topics that occur in a set of documents. It is frequently used in NLP tasks for identifying hidden structures in text. Each document may cover more than one topics, which are conceived as clusters. Thus, each document is represented as a distribution over topics. One popular algorithm for topic modeling is Latend Dirichlet Allocation (LDA), which allows a set of observations to be explained by undefined groups. The distribution over topics is a significant document-level feature and can be further used in document classification tasks.

2.5.3 Classification Methods

Once the raw data are transformed into feature representations, the next step is to feed the vectors into a classifier. There exist various classification algorithms, some of which have already been presented in section 2.5. Among the classification techniques, the most popular used for document classification are Naive Bayes Classifier, Support Vector Machines, Linear Classifier (Logistic Regression) and Decision Trees.

2.6 Depression Detection

2.6.1 Introduction to Mood Disorders

Mood disorders, also known as mood affective disorders, *is a group of conditions where a disturbance in the person's mood is the main underlying feature* [68]. It is a broad mental health class that health professionals use to describe all types of bipolar and depression disorders. People who have mood disorders experience an unstable emotional state and their mood often disrupts their everyday life and interferes with their ability to function. However, mood disorders are generally distinguished in depressive, manic and bipolar disorder. In the first case, people suffer from an overall depressed mood and they may experience periods of feeling extremely sad. On the other hand, maniac episodes are characterized by highly elevated mood, while people with bipolar disorder typically cycle between both.
The causes of a mood disorder vary and they are considered to be both biological and environmental. In particular, people whose family history includes mental disorder are more likely to experience themselves too. Apart from heredity factors, mood disorders

can be also caused by brain structure and functioning. Depression may be caused due to an an imbalance in brain chemicals while bipolar disorder is found to be highly related to mitochondrial dysfunction. In addition, a stressful and traumatic life event, such as a divorce, a job dismissal or the death of a relative, may also lead people to a depressed mood.
People normally experience changes in their mood over periods of time. However, mood disorders are generally more difficult to handle, as the symptoms may vary in intensity and duration from a normal change in emotional state. It is interesting pointing out that women are considered to be twice more prone to developing mood disorders than men. Fortunately, there is available treatment for mood disorders. In the case of depression, treatment involves a combination of self-help, clinical sessions, medication and exercise. The recommended treatment will be based on the type and severity of depression that the individual is experiencing. According to the Diagnostic and Statistical Manual of Mental Disorders [1], mood disorders are classified in depressive disorders, bipolar disorders, substance-induced and not-otherwise specified disorders. As we are dealing with depression detection, we focus on depressive disorders and describe next their subcategories along with their common symptoms.

Bipolar Disorders
Bipolar disorders, previously known as manic-depressive disorders, include bipolar I, bipolar II, cyclothymic substance/medication-induced bipolar bipolar due to another medical condition and other specified and unspecified bipolar disorder. They are characterized by periods of normal feelings, depression and elevated mood. If elevated periods are severe, they are associated with mania. Therefore, people usually swing between being abnormally happy and energetic, followed by feeling depressed and sad. The suicide risk in bipolar disorders is increased and reaches near 6% of affected individuals.

2.6.2 Depressive Disorders

As stated in DSM-5, depressive disorders include major depressive disorder, disruptive mood dysregulation disorder, persistent depressive disorder, premenstrual dysphoric disorder substance/medication-inducted depressive disorder, other specified and unspecified depressive disorder. The features among them are common and concern the presence of sad or irritable mood accompanied by physical changes that affect the way people function in everyday life. What differs among them is the duration, timing, severity and etiology. Moreover, depressive episodes can be appear along with other features, like anxious distress and/or psychotic symptoms (delusions, hallucinations). Next, we provide more information for each depressive category.

Major Depressive Disorder (MDD)
Major depressive disorder, known also as depression, is a recurrent mental disorder and is characterized by episodes of at least 2 weeks duration. Those experiencing MDD generally feel sad, hopeless, discouraged not interested in and not enjoying everyday activities that used to enjoy. They tend to speak slower, with a lower volume and generally experience a retardation in movement and cognitive functioning as thinking and decision making. In order to diagnose an individual with Major Depressive Disorder, they have to experience 5 or more of the following symptoms:

- depressed mood throughout the day
- anhedonia, meaning an inability to derive pleasure from activities or hobbies
- eating disorders (loss or increase in appetite)
- sleeping disorders (hypersomnia or insomnia)
- low self-esteem and feelings of guilt
- suicide ideation

Disruptive Mood Dysregulation Disorder
Disruptive Mood Dysregulation Disoder is a disorder that concerns children and adolescents. It is usually caused before the age of 10 and can be diagnosed in childen over 6 years old. It typically includes severe temper outbursts, manifested verbally of behaviorally which are inconsistent with the current situation. These episodes occur about 3 times per week and are usually observed by the family or teachers. DSM states that children with chronic irritability are at risk of developing anxiety or unipolar depressive disorders when they grow up.

Persistent Depressive Disorder (Dysthymia)
Persistent Depressive Disorder is a more chronical form of depression. It concerns adults who experience a mood disturbance for over 2 years, for most of the day. To diagnose the disorder, 2 or more of the following symptoms have to be present:

- sleep disturbance
- eating disturbance
- low self-esteem
- fatigue or low energy
- difficulty concentating
- hopelessness

As in MDD, these symptoms have to cause significant distress in social or other area functioning so as to be considered as severe. In addition, it is important to state that clinicians, before diagnosing the disorder, have to be sure that symptoms are not attributed to the side effects of a medication or to substance abuse.

Premenstrual Dysphoric Disorder
To diagnose Premenstrual Dysphoric Disorder, at least five of the following symptoms have to be present in the majority of menstrual cycles during the final week and improve afterwards:

- irritability or anger
- mood swings
- depressed mood, hopelessness and/or self-deprication

- poor concentation
- lack of energy, lethargy, hypersomnia
- physical symptoms as muscle pain and weight gain

Substance/Medication-induced or due to another medical condition Depressive Disorder
In this case, depression-like phenomena are caused by prescribed medications, substances of abuse and other medical conditions. The symptoms are common with those in Major Depressive Disorder and last longer than the medication withdrawal period. They typically cause impairment in everyday functioning and include a disturbance in mood or diminished interest in activities.

Other Specified Depressive Disorder
This category applies to people who experience symptoms of a depressive disorder but do not meet the full criteria of any of the above presented categories. In this case, clinicians have to specify the exact reason why the occasion does not fall under the rest of the categories. Such reasons may be due to short-duration episodes or episodes with insufficient symptoms.

Unspecified Depressive Disorder
As in the Other Specified Depressive Disorder, symptoms once again do not meet the requirements of the rest depressive disorders. However, in this case, the clinician chooses not to specify the reason, as there might be not enough information to do so, like in emergency conditions.

2.6.3 Analysis of therapy sessions

Therapy sessions provide a valuable insight into the cognitive and behavioral functioning of clients. They also enlighten the relationship between therapist and client and aspects of the therapeutic activity during the course of the treatment. Due to their complex and multifactorial nature, therapy sessions require a range of perspectives in order to be analyzed.

Through this analysis we can be examine verbal and paraverbal aspects, as well as nonverbal features. Verbal features typically refer to aspects of communication and concern semantic or syntantic characteristics. Non-verbal aspects are mostly visual and include the analysis of gestures, head position or global posture. Paraverbal aspects have to do with the quality of voice and refer to the rate, pitch and volume of speech. Due to the reliability of such insights for therapy analysis, there exists a plethora of studies where authors focus on the body movements [63, 64], facial expressions [35], tone of voice [80], silences [21] and speech disruptions [30].

In the case of the therapist, the analysis of therapy sessions can investigate and evaluate the skills and approaches. In particular, through this analysis, we can examine the communicative skills of the therapist and the adherence to professional requirements. We can also observe the impact on the client and evaluate the overall counseling approach and its effectiveness, during the course. In order to do so, both verbal and nonverbal features can provide significant cues. The therapist's nonverbal behaviour is considered to play an important role in the development of a good clinical relationship [27] and of a good

therapeutic alliance [62].
Analysis of the sessions is also very informative for the state of the clients. Through the dialogue, we can explore the way they interract and participate in a conversation as well as their capability to develop a solid relationship with the therapist. In [36], it is stated that the most important place to search for relationship is the nonverbal behavior of the interactions between therapists and clients. Through sentiment analysis, it is also possible to identify the emotional state of the client and the change during the course of the treatment.

Chapter 3

Depression Detection using Recurrent Neural Networks

3.1 Introduction

Depression detection is the problem of identifying signs of depression in individuals. These signs might be identified in peoples' speech, facial expressions and in the use of language. In our task, we consider the binary classification task of detecting depression in transcribed clinical sessions between a therapist and a client. These sessions provide valuable insights of the cognitive and behavioral functioning of clients. Therefore, we leverage behavioral and psycholinguistic cues of the client and therapist language to enhance our models.

Computational methods can help extract such insights and help psychologists diagnose disorders. To this end, although interpersonal relations between therapists and clients can not be replaced, human decision making can be further augmented by cognitive algorithms. In this way, psychologists can leverage the ability of artificial intelligent systems that process massive amounts of data and thus consult these methods before coming to conclusions.

In this work we focus on the problem of depression detection in psychotherapy sessions. We employ a two-staged hierarchical network functioning at word and turn-level. Each level is equipped with an attention mechanism to extract important content from different parts of the session. To leverage the affective context of depressive language we employ a conditioning method [50] using affective lexica and fuse them in the word-level attention network. We also incorporate the summary attributed to each session into the proposed architectures. Our key contribution is that we integrate existing affective information which improves the results of our hierarchical neural network for depression detection, especially in the case we have small amount of data. This fact results in high performing models and improved robustness across two corpora.

3.2 Related Work

Previous studies have shown that depression affects the language use of depressed individuals. They tend to use more absolutist words [4], negatively valenced-words and the pronoun "I" [66] and mention pharmaceutical treatment for depressive disorder [22, 65].

People in distress also make less use of first person plural pronouns [2] and become more self-focused [53]. In [61], linguistic metadata features are employed across with external knowledge including domain-adapted lexica while in [44], Losada et al. propose evaluation methods of existing depression lexica and create sub-lexica based on part-of-speech tagging. Moreover, for the General Psychotherapy Corpus [1], Malandrakis et al. [47] have explored differences in language between therapist and client using psycholinguistic norms and Imel et al. [33] have identified semantic topics discussed in therapy sessions. Other studies based on therapy sessions have also predicted empathy through motivational interviews [25] and have explored behavioral coding learning models for different psychotherapy approaches [24].

Hierarchical models have been proposed for document classification tasks, in order to leverage the hierarchies existing in the document structure and construct a document-level representation based on turn-level and word-level representations [73]. These models have been augmented with attention mechanisms [7, 77] to identify salient words and sentences in the document [83]. In addition, affective lexica have been published [74, 32, 86, 52, 38, 79] which can effectively contribute in sentiment analysis. As a useful external linguistic knowledge, they can be incorporated into neural architectures [76]. In [50], attentional conditioning methods were proven to enhance model performance for sentiment classification tasks.

The abundance of available data has motivated researchers to investigate depressive language in the context of social media. Orabi et al. [54] examined the performance of different convolutional and recurrent neural networks on Twitter posts of both affected and unaffected individuals. Jamil et al. [34] also used tweets to built user-level and tweet-level classifiers. Schwartz et al. [69] proposed a user-level regression model to predict one's degree of depression based on their posts on Facebook.

3.3 Proposed Systems

Our task is a document classification task, where the input to the model is the transcription of the therapy session and the output is a prediction of the subject's depression status. Hierarchical Neural Networks are a natural fit for document classification, since sessions are composed of turns, which consist of words, forming a hierarchical textual structure.

3.3.1 Model Architecture

Hierarchical Model

In this model, the input sequence of words are embedded into a low-dimensional vector space. In document classification, we want to extract the hierarchies existing in documents in a bottom-up manner. To this end, we use a two-stage hierarchical network that operates at word and turn-level, as we can see in Fig. 3.1. Both the word-level and the turn-level encoders are implemented using Recurrent Neural Networks (RNN). Since not all words or turns contribute equally to the final session representation, we augment both encoders with an attetion mechanism [7]. At the first level of the hierarchy, a word-level encoder produces turn-level representations. We feed the words of each turn to the encoder and then combine them to a single representation using an attention mechanism. Let h_{ki} be

[1] http://alexanderstreet.com

the annotation of the i-th word in the k-th turn obtained through the word-level encoder. The k-th turn representation results as follows:

$$\begin{aligned} \gamma_{ki} &= g(h_{ki}), \\ \alpha_{ki} &= \frac{e^{\gamma_{ki}}}{\sum_i e^{\gamma_{ki}}}, \\ t_k &= \sum_i \alpha_{ki} \cdot h_{ki} \end{aligned} \tag{3.1}$$

where g is a learnable mapping, a_{ki} are the attention weights for each word and t_k is the k-th turn representation.

The session representations are extracted in a similar manner. The turn representations t_k are fed into the turn-level encoder and then the attention weights are calculated. The final representations are the weighted sum of the turn-level encoder hidden states with the attention weights.

$$\begin{aligned} \beta_k &= f(t_k), \\ \tau_k &= \frac{e^{\beta_k}}{\sum_i e^{\beta_k}}, \\ r &= \sum_k \tau_k \cdot \beta_k \end{aligned} \tag{3.2}$$

where f is a learnable mapping, τ_k are the attention weights and r is the session-level representation.

3.3.2 Summary Incorporation

In the General Psychotherapy Corpus, sessions are accompanied with a summary given by an expert. This summary can be seen as a high-level overview of the topics discussed during the session and is denoted as "title" in the dataset. Similar to [5], we extract the summary's vector representation through the word-level encoder and concatenate it directly with the final session representation, before feeding it to the classifier. Let o_t be the summary representation obtained through the word-level encoder. Concatenating it with the session representation (3.2) produces the final vector $(o_t||r)$ that is fed to the classifier.

3.3.3 External Knowledge Conditioning

According to [66, 44], the affective content can be a distinguishing factor between depressed and not-depressed language. Based on this observation, we employ external linguistic knowledge about the affective content of words. These features can be obtained by sources created by human experts. We consider emotion, sentiment, valence and psycho-linguistic annotations for words. Specifically, we construct a context vector c_{ki} for each word i in turn k, where each dimension corresponds to an annotation from existing affective lexica.

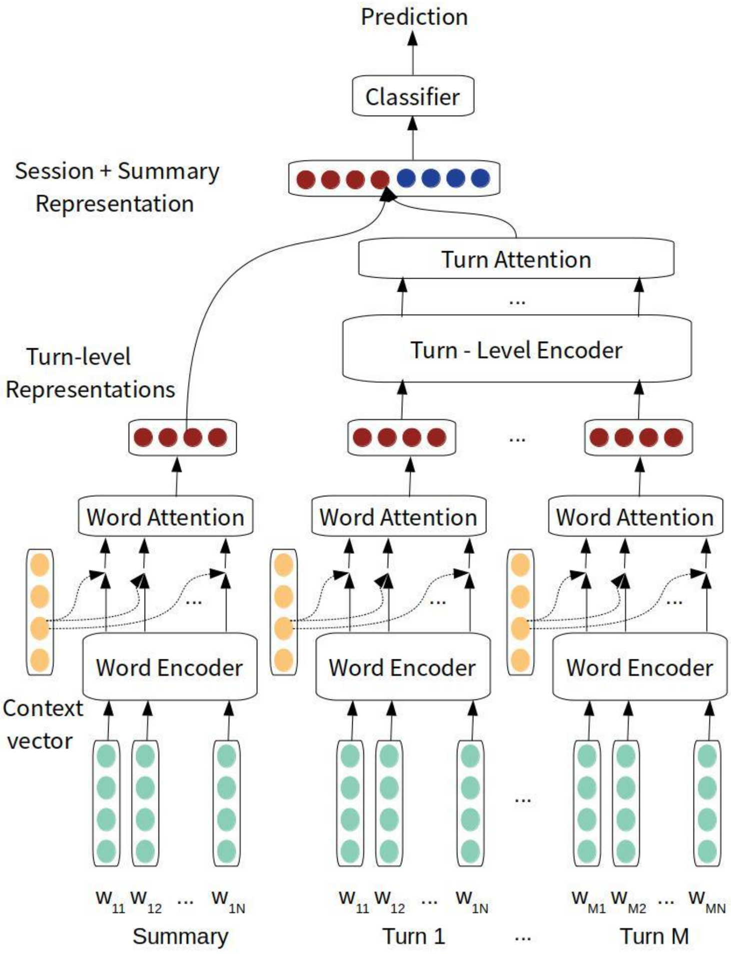


Figure 3.1: Hierarchical Model with Attentional Conditioning

We set missing dimensions to zero and we integrate the context vector in the attention mechanism of the word-level encoder. Specifically, we concatentate, $||$, the context vector to the hidden representation of each word h_{ki}, modifying Eq. **??**:

$$\begin{aligned}\gamma_{ki} &= g(h_{ki}||c_{ki}),\\ \alpha_{ki} &= \frac{e^{\gamma_{ki}}}{\sum_i e^{\gamma_{ki}}},\\ t_k &= \sum_i \alpha_{ki} \cdot (h_{ki}||c_{ki})\end{aligned} \tag{3.3}$$

Eq. 3.3 shows that we compute the intermediate representations γ_{ki} using both the word hidden states and the context vector. The softmax function is then applied to γ_{ki} to create the attention weights distribution α_{ki}. The incorporation of external information at this level can force higher values for attention weights corresponding to salient affective words. We also use the concatenated $h_{ki}||c_{ki}$ to create the turn representations t_k to propagate the affective features to the turn-level encoder.

3.4 Corpora Overview

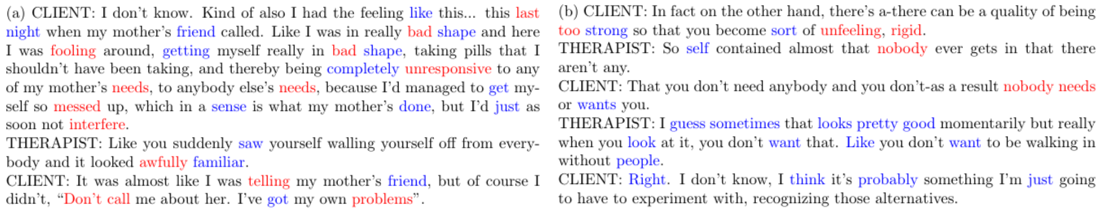
In this work, we use two datasets for addressing the problem of detecting depression. These are the General Psychotherapy Corpus and the DAIC-WoZ datasets. As they provide depression-based annotations, they are a good choice for training models for the task of depression classification. Following is a presentation of the two.

General Psychotherapy Corpus

The General Psychotherapy Corpus (GPC) is a dataset by the "Alexander Street Press". It is a collection of over 1300 transcribed therapy sessions conducted between a therapist and a client and it covers a variety of clinical approaches. For each session, the dialogue is split into individual turns of the two speakers which are annotated as patient-side or therapist-side turns. Metadata are also provided at session level and include demographic information for both the therapist and the client, the symptoms that the clients are experiencing and a summary attributed to each session, labeled as "title". It also provides professional information about the therapist, as the psychological approach and the years of experience. Some of the sessions are conducted between more than two speakers, so we extract a subset of 1262 sessions, which consist of one therapist and one client and will be used as the training dataset. Among the 1262 sessions, 881 of them are annotated as "not-depressed" samples whereas the rest 381 are annotated as "depressed".

DAIC-WoZ

The DAIC-WoZ dataset is part of the DAIC corpus [26]. It contains a set of clinical interviews which were carried out so as to assist the task of detecting distress disorders. Each interview is conducted between a client and a virtual agent serving as therapist, called Ellie, which is controlled by a human interviewer placed in another location [13]. The dataset contains audio and video recordings as well as transcripts of the clinical interviews. Data are split into train, development and test set, consisting of 107, 35 and 47 samples respectively. Depression is evaluated on the PHQ-8 depression scale.



(a) CLIENT: I don't know. Kind of also I had the feeling like this... this last night when my mother's friend called. Like I was in really bad shape and here I was fooling around, getting myself really in bad shape, taking pills that I shouldn't have been taking, and thereby being completely unresponsive to any of my mother's needs, to anybody else's needs, because I'd managed to get myself so messed up, which in a sense is what my mother's done, but I'd just as soon not interfere.
THERAPIST: Like you suddenly saw yourself walling yourself off from everybody and it looked awfully familiar.
CLIENT: It was almost like I was telling my mother's friend, but of course I didn't, "Don't call me about her. I've got my own problems".

(b) CLIENT: In fact on the other hand, there's a-there can be a quality of being too strong so that you become sort of unfeeling, rigid.
THERAPIST: So self contained almost that nobody ever gets in that there aren't any.
CLIENT: That you don't need anybody and you don't-as a result nobody needs or wants you.
THERAPIST: I guess sometimes that looks pretty good momentarily but really when you look at it, you don't want that. Like you don't want to be walking in without people.
CLIENT: Right. I don't know, I think it's probably something I'm just going to have to experiment with, recognizing those alternatives.

Figure 3.2: Example of sessions for (a) a depressed client and (b) a not-depressed client. Blue: positive words, Red: negative words, as found in LIWC, AFINN and Bing Liu Opinion Lexicon

3.4.1 Data Analysis

Analysis on the General Psychotherapy Corpus

In this section, we analyze and present the structure of data in the GPC dataset and explore statistics regarding the language used by depressed and not-depressed individuals. An indicative example of the form of a transcribed session is shown in Figure 3.2. It can be seen that sessions are provided as consecutive turns which have a therapist or client label. We also used the three lexica mentioned in the caption which have psycholinguistic annotations and coloured the words that express positive and negative affective content. As it is shown in the Figure, the depressed client's language contains more negative affective content.

Next, we examine the language use of the two speakers. In Table 3.1, we present the average number of tokens in the turns of clients and therapists. We notice that the clients speak twice as much as the therapists on average.

Table 3.1: Dialogue turns statistics for therapists and clients

Features	**Sum**
Average number of turns/session	196
Average number of tokens in turns	32.3
Average number of tokens in client turns	42.9
Average number of tokens in therapist turns	20.7

The next step is to compare the language use between depressed and not-depressed clients. In particular, we are interested in the use of words that express positive and negative sentiment, sadness and anxiety. To this end, we employ the LIWC lexicon [74] which provides psycho-linguistic annotations for $18,504$ words, for 73 different word categories. In Table 3.2 we compare the vocabularies of depressed and not-depressed people and specifically show the vocabulary sizes and the percentage of their words which are associated with each of these four affective categories, in LIWC. We see that depressed subjects use a more consice vocabulary, but include a higher percentage of affective words in it. This hints to the importance of incorporating knowledge about affective language in depression detection.

Table 3.2: Vocabulary use statistics between the two classes

Features	**Depressed**	**Not-depressed**
Samples	381	881
Total turns	41589	88191
Vocabulary size	16166	23201
Number of affective words	1672	2036
Percentage of affective words	10.34%	8.77%

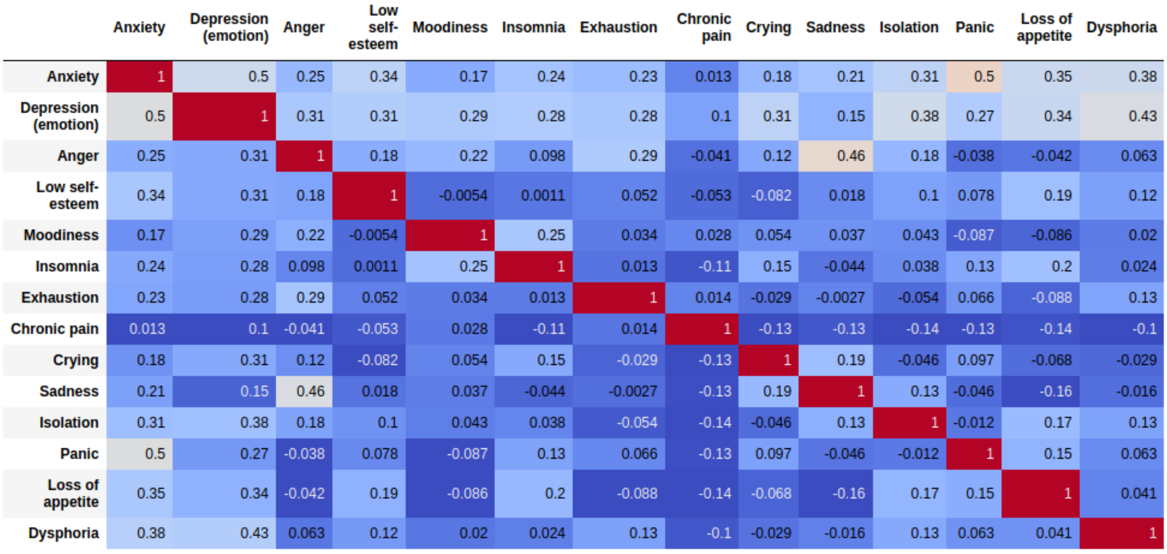
Moreover, we further examine the statistics of the four word categories in the language of clients. In Table 3.3 we present the percentages of each of these four categories. Specifically, we split the dataset into the samples of depressed and not-depressed people. Subsequently, we use the LIWC lexicon and count the number of the occurences of words that belong to each of these affective categories. Finally, we compute their percentages, in the total words of the samples of depressed and not-depressed people. The results indicate that depressed individuals tend to use more negatively-valenced words, which stands in agreement with the related literature [66].

Table 3.3: Percentage of occurences of affective word categories in client language across the two classes

Categories	Depressed	Not-depressed
Positive sentiment	2.17%	2.26%
Negative sentiment	1.38%	1.30%
Sadness	0.32%	0.32%
Anxiety	0.30%	0.22%

Correlation Map

For the General Psychotherapy Corpus, there are more than one symptoms attributed to each session in the provided metadata that cover a wide range of emotional states. Therefore, it is important to examine their their co-occurences and the correlation between the symptoms. In Figure 3.3, we present the symptoms' correlation map. Due to the large number of different symptoms (83 in total) and the lack of available space for illustrating all the categories, there are shown only the ones which have the higher correlation with depression.



	Anxiety	Depression (emotion)	Anger	Low self-esteem	Moodiness	Insomnia	Exhaustion	Chronic pain	Crying	Sadness	Isolation	Panic	Loss of appetite	Dysphoria
Anxiety	1	0.5	0.25	0.34	0.17	0.24	0.23	0.013	0.18	0.21	0.31	0.5	0.35	0.38
Depression (emotion)	0.5	1	0.31	0.31	0.29	0.28	0.28	0.1	0.31	0.15	0.38	0.27	0.34	0.43
Anger	0.25	0.31	1	0.18	0.22	0.098	0.29	-0.041	0.12	0.46	0.18	-0.038	-0.042	0.063
Low self-esteem	0.34	0.31	0.18	1	-0.0054	0.0011	0.052	-0.053	-0.082	0.018	0.1	0.078	0.19	0.12
Moodiness	0.17	0.29	0.22	-0.0054	1	0.25	0.034	0.028	0.054	0.037	0.043	-0.087	-0.086	0.02
Insomnia	0.24	0.28	0.098	0.0011	0.25	1	0.013	-0.11	0.15	-0.044	0.038	0.13	0.2	0.024
Exhaustion	0.23	0.28	0.29	0.052	0.034	0.013	1	0.014	-0.029	-0.0027	-0.054	0.066	-0.088	0.13
Chronic pain	0.013	0.1	-0.041	-0.053	0.028	-0.11	0.014	1	-0.13	-0.13	-0.14	-0.13	-0.14	-0.1
Crying	0.18	0.31	0.12	-0.082	0.054	0.15	-0.029	-0.13	1	0.19	-0.046	0.097	-0.068	-0.029
Sadness	0.21	0.15	0.46	0.018	0.037	-0.044	-0.0027	-0.13	0.19	1	0.13	-0.046	-0.16	-0.016
Isolation	0.31	0.38	0.18	0.1	0.043	0.038	-0.054	-0.14	-0.046	0.13	1	-0.012	0.17	0.13
Panic	0.5	0.27	-0.038	0.078	-0.087	0.13	0.066	-0.13	0.097	-0.046	-0.012	1	0.15	0.063
Loss of appetite	0.35	0.34	-0.042	0.19	-0.086	0.2	-0.088	-0.14	-0.068	-0.16	0.17	0.15	1	0.041
Dysphoria	0.38	0.43	0.063	0.12	0.02	0.024	0.13	-0.1	-0.029	-0.016	0.13	0.063	0.041	1

Figure 3.3: Correlation map of psychotherapy symptoms.

3.5 Experiments

3.5.1 Lexica

In this work, we employ lexica and derive lexical representations for words. These lexica provide psycho-linguistic, sentiment and emotional annotations for a different range of words. The lexica used are namely LIWC [74], Bing Liu Opinion Lexicon [32], AFINN [86], Subjectivity Lexicon (MPQA) [79], SemEval 2015 English Twitter Lexicon (Semeval) [38] and NRC Emotion Lexicon (Emolex) [52].

AFINN, Semeval15 and Bing Liu provide $1D$ positive/negative sentiment annotations for 6,786, 1,515 and 2,477 words respectively. MPQA provides $4D$ sentiment ratings for 6,886 words. Emolex provides $19D$ emotion ratings for 14,182 words. LIWC provides $73D$ psycho-linguistic annotations for 18,504 words. The combined six lexica provide a vocabulary coverage of 25,534 words. These features are concatenated into a 99-dimensional context vector.

3.5.2 Experimental Setup

We conduct a set of different experiments and employ five network architectures. As weak baseline models we employ Tf-Idf for feature extraction combined with an SVM classifier with linear kernel (SVM), or with Logistic Regression with L2 regularization, for classification (LR). Moreover, we develop a Hierarchical Attention-based Network with no external knowledge, which is referred to as HAN. Subsequently, we augment this model with affective conditioning at the attention mechanism, where lexicon annotations are concatenated with word hidden states before the word-level attention layer (HAN+L). We also utilize the session summaries that are provided with the GPC dataset and extend the HAN model with the integration of the summary's representation before the classification layer (HAN+S). Our last model results from the combination of the two previous network architectures (HAN+L+S). As there is no summary assigned to the sessions of the DAIC-WoZ corpus, we evaluate the HAN and HAN+L models on this dataset. For our experiments on GPC we report macro-averaged F1 score due to the class imbalance present in the datasets. This score is calculated using 5-fold cross-validation, where each fold contains an $80\% - 20\%$ train-validation split of the original data. For the DAIC-WoZ corpus we follow the experimental procedure of [48], thus we additionally measure the Unweighted Average Recall (UAR) and present results on the development and test set.

3.6 Implementation Details

Our model consist ofs two encoder layers, where a Bi-directional Gated Recurrent Unit (GRU) is implemented on the first stage and two more on the second. All encoders use 300 hidden size and 0.2 dropout rate. Model parameters are optimized using Adam with 10^{-3} learning rate. The model is trained for a maximum of 40 epochs and we use early stopping to select the model with the lowest validation loss. For the system implementation, we use Pytorch framework [57] and Scikit-Learn [58].

3.7 Results & Discussion

We compare the performance of the proposed models when given as input the client turns (Client), the therapist turns (Therapist) or the whole dialogue (Client+Therapist). In Table 3.4 we present the results for the GPC dataset. We see that the integration of external affective and psycholinguistic features improves model performance for all model configurations over the HAN and SVM baselines. Furthermore, we notice that when we add the summary information we also gain a performance boost, sometimes greater that the external affective information. Summary and lexica integration leads to a performance increase when we provide only the client data. In addition, we see that the client turns are more important for depression detection, as expected, and incorporation of the therapist turns contributes little to the overall model performance. Based on our results, the best performance can be achieved by the HAN+L+S model, while HAN+L model performs best if such annotation is not available.

Table 3.4: Results of different architectures on the GPC

Experiment	Client	Therapist	Client+Therapist
LR	0.476	0.447	0.468
SVM	0.478	0.464	0.484
HAN	0.681	0.647	0.695
HAN+S	0.698	0.641	**0.718**
HAN+L	0.693	**0.659**	0.706
HAN+L+S	**0.715**	0.640	**0.716**

We further use a confusion matrix to visualize the model performance when attentional conditioning is applied. In Figure 3.4, we depict the confusion matrix of the HAN+L model when trained using the client data. The number of correctly predicted labels correspond to the True Positive and True Negative values. We notice that these values are higher than the false predictions, as expected.

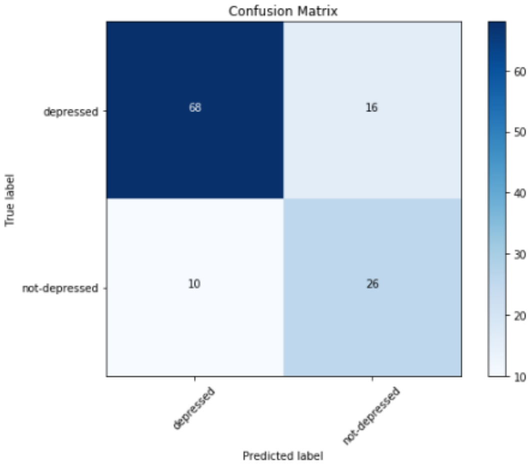


Figure 3.4: Confusion Matrix for the HAN+L (Client) model.

In Table 3.5 we present results for the DAIC-WoZ dataset. We observe that affective conditioning significantly improves the performance over the baseline model (HAN). Our HAN+L model also shows improved F1 and UAR scores over the models proposed in [48]. Overall, we see that conditioning of external psycho-linguistic knowledge in this small dataset (189 samples) enhances the performance and the results are comparable to these of the GPC corpus.

Table 3.5: Results of the DAIC-WoZ corpus

Method	**Devel. Set**		**Test Set**	
	F1-macro	**UAR**	**F1-macro**	**UAR**
[48]HCAN	0.51	0.54	0.63	0.66
[48]HLGAN	0.60	0.60	0.35	0.33
HAN	0.46	0.48	0.62	0.63
HAN+L	**0.62**	**0.63**	**0.70**	**0.70**

3.8 Conclusions

We propose a novel model for depression detection with integrated external affective and psycho-linguistic information. Our model is a Hierarchical Attention Network that encodes words and dialogue turns in different levels of the architecture. The external features are integrated into the attention mechanism, forcing the attention weights to focus on salient affective information. The external knowledge integration leads to high performing models and increased robustness for both the small datasets (1262 and 189 samples respectively) we explore. Finally, a future plan would be to incorporate more elaborate information sources, e.g. expert knowledge bases from psychologists.

Chapter 4

Dialogue Modeling for Depression Detection

4.1 Introduction

In this section, we explore the task of dialogue modeling and propose network architectures applied to the task of depression detection from transcribed therapy sessions.

Dialogue modeling is the task of automatically detecting and understanding discourse structure. It concerns the incorporation of dialogue level information into the implemented models, by considering dialogue interactions between the speakers of a conversation. Especially in the case of transcribed conversations, it is crucial not to treat utterances as a sequence of inputs, as common models implemented for NLP tasks, but to capture the inter-speaker dependencies.

In the case of depression detection, apart from the analysis of posts on social media and forums, salient information can be also obtained through the analysis of therapy sessions. These provide a more thorough insight into the emotional state of the client, as the dialogue between the therapist and the client can significantly help to investigate the way depressed people interact and participate in a discussion. Therefore, modeling the dialogue of therapy sessions can identify communicative cues in a context-aware manner and provide valuable insights of the cognitive functioning of clients.

4.2 Related Work

In sequential data with a hierarchical structure, there often exist complex dependences between subsequences. In the case of a conversation, these subsequences are represented by utterances. Dialogue models have been proposed in order to incorporate the discourse information in the network architectures, instead of treating the utterances as a sequence of inputs, as they may not well capture the pauses, turn-talking and grounding phenomena in a dialogue [11]. In [9], Hori et al. propose a network of role-dependent RNNs where the context vector is provided along with the input of the next timestep. In [43], Liu et al. also encode the context vector by employing a separate RNN. The results show that role-dependent systems perform better than role-independent for dialogue modeling. Speaker role is also incorporated in language models implemented in [45] and RNNLMs are adjoined for all turns to model the whole conversation. Moreover, for conversation modeling based on data collected from online forums, in [85], Zayats et al. propose a novel approach for modeling threaded discussions on social media, using graph-structured bidirectional LSTM. They capture discussion dynamics by representing the conversation in a both hierarchical and temporal structure. Xu et al. also address the same task [82] by proposing a deep neural network whch incorporates background knowledge for conversation modeling. They employ a special Recall gate which cooperates with the local memory of the LSTM cell and thus decides whether an utterance is related to the dialogue history or not.

Dialogue modeling is also applied to emotion recognition tasks based on conversations, as it is crucial for developing empathetic machines. In [46], Majumder et al. keep track and use the individual party states for emotion classification. In [14], they train a unigram model for each emotion category and perform utterance-level emotion prediction. Moreover, in [42], they show that augmenting the standard lexical and prosodic features with contextual features improves the model performance. In the case of identifying mood disorders, conversational features can be effective for detecting the severity of mood symptoms in patients [16, 84]. Lastly, in [56], Park et al. classify client utterances in counseling dialogues by augmenting a pretrained conversation model [78] with task-specific layers.

4.3 Proposed Systems

Our task is a document classification task where the input is a transcribed therapy session and the output is the prediction of the client's depression status. To leverage the hierarchical structure of documents, we employ a hierarchical network and adapt it so as to incorporate the speaker role of each utterance.

4.3.1 HAN with Speaker-Role Discrimination

In this model, we use a hierarchical network that functions at word-level and sentence-level in order to extract information that exists in smaller and larger parts of the document. The overall system architecture is shown in Figure 4.1. For both stages of the architecture, we employ recurrent neural networks. In addition, not all words and all sentences in the document contain important information. To this end, we augment each level of the network with an attention mechanism so as to focus on the salient parts. At the first level of the network, a word-level encoder produces turn-level representations. The input words of each turn are fed into the encoder and the attention mechanism outputs a single turn-representation. Supposing that h_{ki} is the representation of the i-th word in the k-th turn, then the representation of the k-th turn results as follows:

$$\begin{aligned} \gamma_{ki} &= g(h_{ki}), \\ \alpha_{ki} &= \frac{e^{\gamma_{ki}}}{\sum_i e^{\gamma_{ki}}}, \\ t_k &= \sum_i \alpha_{ki} \cdot h_{ki} \end{aligned} \tag{4.1}$$

where g is a learnable mapping, a_{ki} are the attention weights for each word and t_k is the k-th turn representation.

Subsequently, turn-level representations are extracted through the second stage of the network. In particular, we implement two encoders, attributed to each speaker. Having acquired the speaker role of each turn from the dataset, the model feeds the turn into the corresponding speaker encoder. The client-based encoder produces through the attention mechanism a final representation for the client turns, while the therapist-based encoder produces the respective representation for the therapist turns. Let t_{kc} be the k-th client turn and t_{kt} be the k-th therapist turn. Then, client-turns representation is given by:

$$\begin{aligned} \beta_k &= f_c(t_{kc}), \\ \tau_k &= \frac{e^{\beta_k}}{\sum_i e^{\beta_k}}, \\ r_c &= \sum_k \tau_k \cdot \beta_k \end{aligned} \tag{4.2}$$

while the therapist-turns representation is given by:

$$\begin{aligned} \beta_k &= f_t(t_t), \\ \tau_k &= \frac{e^{\beta_k}}{\sum_i e^{\beta_k}}, \\ r_t &= \sum_k \tau_k \cdot \beta_k \end{aligned} \tag{4.3}$$

where f_c and f_t are learnable mappings, τ_k are the attention weights and r_c, r_t are the client-turns and therapist-turns total representations, respectively.
Finally, the two representations are concatenated and fed to a classifier for prediction.

4.3.2 3 HAN encoders and Late Fusion

The overall proposed architecture is shown in Figure 4.2. In this case, we use 3 distinct HAN encoders, for the turns of client, therapist and the whole dialogue, respectively. We train each HAN model separately as we subsequently fuse the output representations before feeding them to the classifier. As a fusion method we use concatenation. To this end, the weights of the HAN systems are freezed and so we train only the upper linear layer that is used for classification.

4.4 Experiments

4.4.1 Data

The proposed architecture can be applied to different datasets with transcribed conversations, for emotional recognition tasks. In this work, we use the General Pshychotherapy Corpus. A thorough presentation and analysis of the dataset is provided in section 3.4.1.

4.4.2 Experimental Setup

After input data tokenization, we derive $300D$ word-level representations using GloVe [59] pretrained word embeddings, trained on the Common Crawl corpus. Our model consists

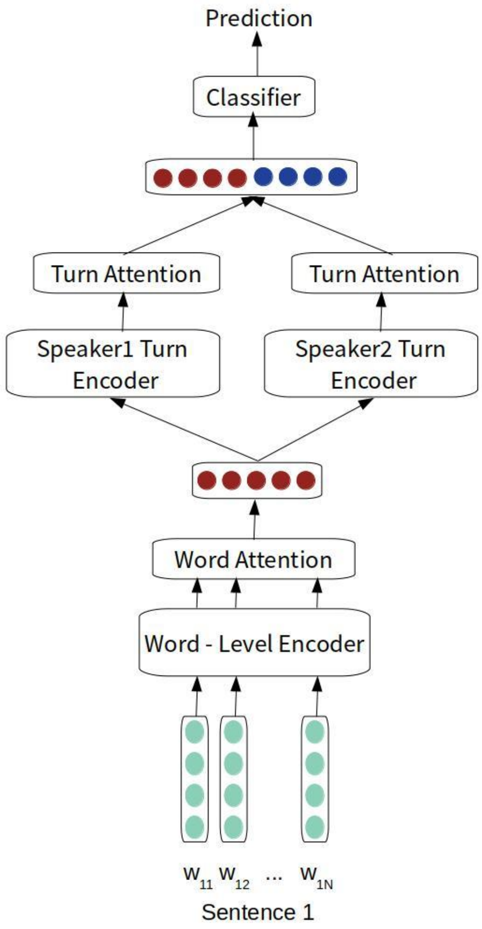


Figure 4.1: Hierarchical Model with Speaker Role Discrimination.

of two encoder layers, where a Bi-directional Gated Recurrent Unit (GRU) is implemented on each stage. All encoders use 300 hidden size and 0.2 dropout rate. Model parameters are optimized using Adam with 10^{-3} learning rate. Model is trained for a maximum of 40 epochs and we use early stopping to select the model with the lowest validation loss. For model implementation, Pytorch framework [57] is used.

4.4.3 Baselines

In Table 4.1, we present the result of the proposed system in 4.3.1 when trained using the GPC dataset. We also present the corresponding score for the HAN baseline, for comparison reasons.

Dataset	HAN	Speaker-Dependent HAN
General Psychotherapy Corpus	69.5	68.9

Table 4.1: F1-scores for model architectures for the GPC dataset.

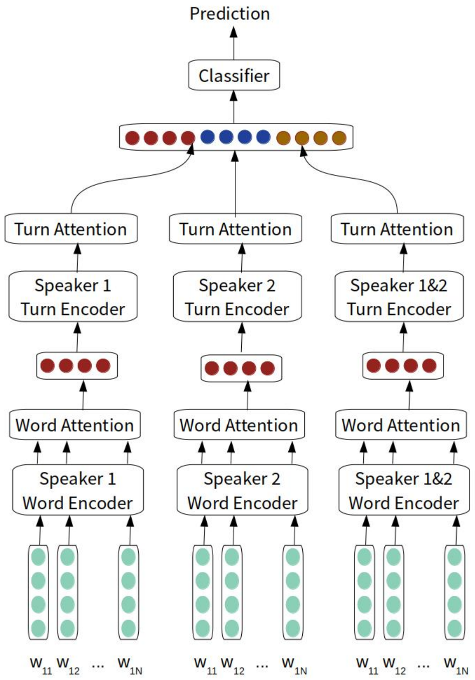


Figure 4.2: 3 Encoders with Late Fusion

4.5 Discussion

As shown in Table 4.1, the proposed Speaker-Dependent HAN performs generally equally to the HAN baseline. This hints to the importance of introducing connections between the role-dependent encoders, so as to be informed each time not only from the previous turns of the same speaker but also from the previous turns in the conversation.

4.6 Conclusions

We introduce a model with hierarchical speaker-dependent structure for depression classification and compare its efficacy with the hierarchical baseline model. The method is simple yet efficient achieving comparable performance to the baseline.
In the future, we aim to introduce connections between the role-dependent encoders and condition the attention mechanisms of each encoder on the context vectors derived from previous conversation turns.

Chapter 5

Conclusions

5.1 Discussion

In this work, we investigate methods and architectures in order to tackle the task of depression detection, from transcribed clinical interviews. We perform text-based classification and explore the valuable insights into the subject's emotional state that can be provided through the language used during the therapy. Our results show that the analysis of therapy sessions contributes majorly to the identification of depression signs in the language of the client. The proposed methods introduce a novel way to examine the use of language and also incorporate the speaker role during a conversation. Our approaches enhance the performance over the baseline models by a significant margin.

Firstly, we propose a novel model for depression detection with integrated external affective and psycho-linguistic information. Therapy sessions consist of consecutive turns between the therapist and the client in a document structure. Therefore, we develop models for document classification in the context of detecting the depression severity in the language of the client. In particular, our basic system consists of a two-staged hierarchical network equipped with attention at both stages, which treats the dialogue as a sequence of turns and performs at both word-level and turn-level. Moreover, as there is a summary attributed to each session available in the corpus used, we experiment with incorporating the summaries into the model architecture. To leverage the affective content of the input data, we further augment our system by integrating external psycholinguistic features into the attention mechanism, forcing the attention weights to focus on salient affective information in the corpus. Our models are trained on two small datasets and their performance is compared when given as input the client turns, the therapist turns or the whole dialogue. Our resutls show that the integration of external affective and psycho-linguistic features improves model performance for all model configurations over the baselines. Furthermore, we notice that when we add the summary information we also gain a performance boost, sometimes greater that the external affective information. In addition, we see that the client turns are more important for depression detection, as expected, and the incorporation of the therapist turns contributes little to the overall model performance. Overall, we show that the external knowledge integration leads to high performing models and increased robustness for both the small datasets (1262 and 189 samples respectively) we explore.

Secondly, we experiment with dialogue modeling in order to leverage the existing dependences in the conversation. In particular, we propose a variant of the previously described hierarchical structure in which we incorporate the speaker role into the turn encoding process. This is implemented by developing role-dependent turn-level encoders, which encode the turns based on the corresponding speaker. The method is simple yet efficient and the performance of the system is comparable to the hierarchical baseline model, which implies that the introduction of inter-speaker dependencies could possibly increase further the performance of our models. We also propose other architectures that are considered to be effective for the task in hand.

5.2 Future Work

In the future, we aim to augment the first work by incorporating into the model architectures more elaborate information sources, compared to the psycho-linguistic annotations provided. These could be expert knowledge bases from psychologists. Another possible auxiliary source used could be a task-specific lexicon. This type of lexicon could be created manually with depression severity annotations for a subset of words and then expanded for the whole corpus. Moreover, since our data concern dialogue interactions, non-verbal features may also contain salient information. These features could be derived from the therapy session and concern the duration or the frequency of turns, in order to help the models make the diagnosing process of depression more discriminative. Sessions' summaries could also be used alone for classification. In addition, another future direction would be to replace the RNN word-level encoder with a BERT model. The outputs of BERT would be combined in the same way to obtain a final turn representation and then fed into the turn-level encoder. Lastly, topic modeling is a method applied to text documents or conversations and it could also be applied to the nature of our data. It groups the turns into semantic sections and thus helps algorithms determine the different conversation topics in each session.

A limitation of the second work is the fact that the proposed model encodes the turns of each speaker separately, without taking into consideration the speaker dependencies. In particular, it considers the self-speaker yet not the inter-speaker influences. To this end, a possible idea could be to create recurrent connections between speaker-dependent encoders, in order to provide the output of the last speaker to the input (or to the hidden state) of the next speaker. Another possible direction could be to make use of the attentional conditioning method implemented, in order to condition the attention weights of the turn encoder on the context vector of the last turn in the conversation. In this case, we would provide the knowledge of the prior conversation turns to the distribution of the attention weights. The proposed models are further described in subsection 5.2.1.

In addition, another limitation of the task is the nature of data. Given that the sessions include the original transcribed conversation between therapist and client, there exists a heterogeneity in the nature of turns, meaning that there are a lot backchannels and turns that do not contribute to the overall session content. The number of tokens in each turn also varies and this fact makes the conversation modeling a difficult task. To this end, we could possibly discriminate backchannels by separating turns based on their token length, or by using a pretrained dialogue act classification model to make predictions for each turn. These models are further presented in subsection 5.2.2.

Finally, a significant constraint of the task in hand is the small amount of labeled data, due to medical confidentiality. Therefore, we aim to alleviate this limitation by performing transfer learning from a conversation model. This should be preferably trained on a corpus with sentiment or emotion annotations, in order to use the shared feature representations for depression detection.

5.2.1 Proposed Conversation Models

In this section, we present models that incorporate the speaker role and act as conversation models. By doing so, we leverage the structure of conversations and thus consider each turn based not only on the previous turns of the specific speaker, but also the previous turns in the whole conversation.

Turn Encoders with Recurrent Connections

The system overview is depicted in Figure 5.1. It is a variant of the system shown in Figure 4.1 in the sense that we introduce recurrent connections between turn encoders. To this end, we consider not only the self-speaker but also the inter-speaker dependences. To this end, we provide each turn encoder the context of the conversation, by feeding the hidden state of the interlocutor's previous turn, as proposed in [9]. The hidden state of the previous turn is thus fused with the current turn representation and passed through the corresponding speaker-turn encoder.

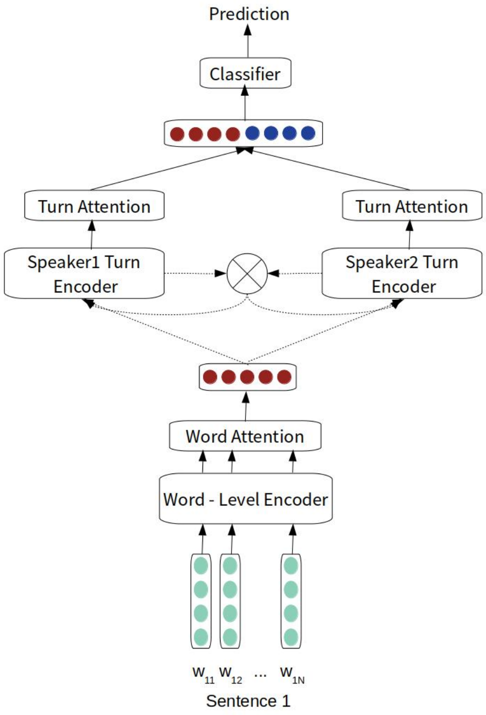


Figure 5.1: System Overview with Recurrent Connections between Encoders

Turn Encoders with Context-vector Conditioning

The proposed system is a variant of the previous model, and is shown in Figure 5.2. In particular, we use a shared word-level encoder to extract turn-level representations. We use two distinct turn-level encoders that correspond to the two speaker roles. Then, we use the hidden state of the last timestep as a context vector and condition the attention weights of the next turn representation on this vector. To this end, we force the attention weights to focus on information relevant to the information contained in the previous turn.

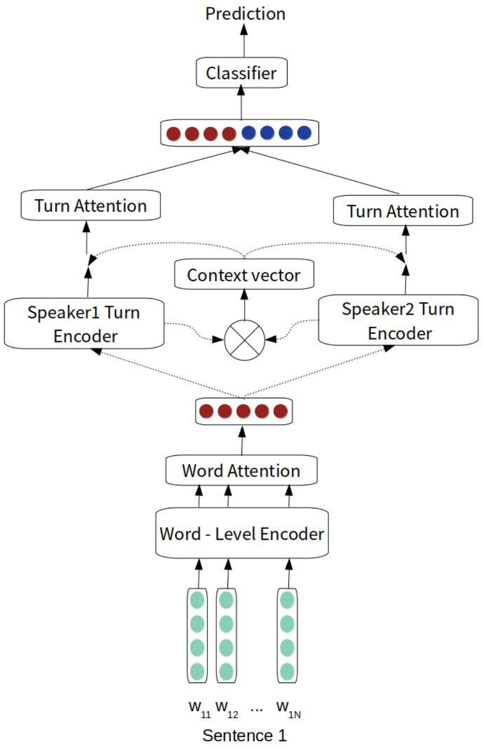


Figure 5.2: System Overview with context vector conditioning

5.2.2 Proposed Turn-Distinguished Models

In this section we present models that distinguish turns depending on their nature. Our dataset consists of consecutive turns, for which the speaker role is available. However, we have no information regarding the nature of turns. That is, we do not know whether each turn contributes to the content of the session or just consists a non-verbal response. To this end, the token length of turns differs significantly across the corpus and this factor makes the generalization process of the models difficult. We thus need to distinguish turns into backchannels and salient turns. As backchannels, we define sentences and responses that include sentence completions, requests for clarification, brief statements and non-verbal responses. Next, we propose two models to address this limitation.

HAN based on DA Classification System

The general architecture is shown in Figure 5.3. Since we have no information available for the nature of turns, we could possibly use a pretrained Dialogue-Act Classification system which is trained to recognize the content of turns. We further augment the hierarchical model by introducing two separate turn encoders, for salient turns and backchannels respectively. The output of the DA system acts as a gate that decides which turn encoder to feed with each input turn. The resulting representations derived from the turn encoders are fused and then fed to the classifier.

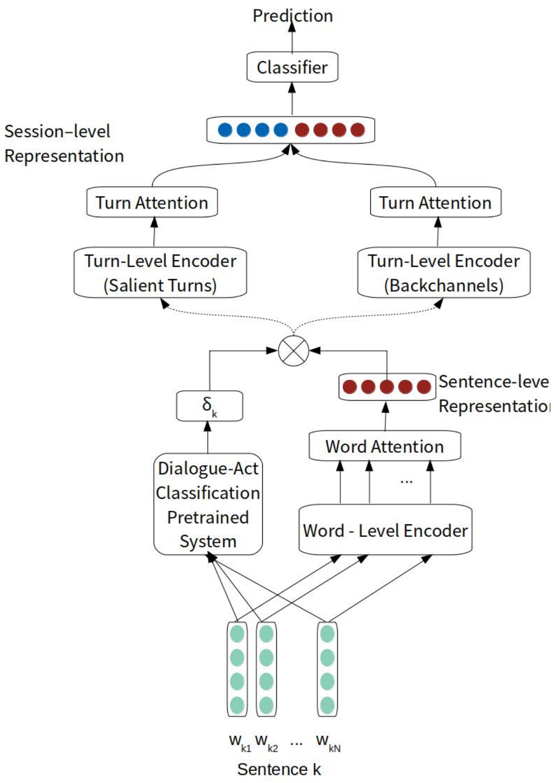

Prediction
Classifier
Session–level Representation
Turn Attention
Turn Attention
Turn-Level Encoder (Salient Turns)
Turn-Level Encoder (Backchannels)
δ_k
Sentence-leve Representatio
Word Attention
...
Dialogue-Act Classification Pretrained System
Word - Level Encoder
w_{k1} w_{k2} ... w_{kN}
Sentence k

Figure 5.3: HAN based on Pretrained Dialogue Act Classification System

HAN with Turn Length-based Encoders

In this case, we distinguish between turns without using a pretrained DA classification system. We consider that turns with small token length do not typically contain much salient information. Therefore, we define a token length threshold and regard turns with less tokens than the threshold as backchannels. Subsequently, we also use a shared word-level encoder and two distinct turn-level encoders as proposed in Figure 5.3. Classification is finally performed on the fused representations derived from turn encoders.

Overall, depression detection is a very significant task which can majorly contribute to the medical community. The development of intelligent systems aims to help both clinicians and patients detect signs of depression at an early stage. Therapy sessions also play an important role in providing salient cues of the clients' emotional state and thus help to distinguish depression. In this work we leverage these cues and augment the proposed architectures with external affective information in order to make the detection process more discriminative. Our results show that the proposed method significantly increases the models' performance. We consider that the integration of the speaker role will further enhance the results and thus we aim to maintain the hierarchical structure yet discriminate the encoding scheme based on the speaker role and introduce connections between the individual speaker-dependent subsystems.

Appendix A

Abbreviations

(AI): Artificial Intelligence
(ML): Machine Learning
(DNN): Deep Neural Network
(ANN): Artificial Neural Network
(LSTM): Long Short-Term Memory
(RNN): Recurrent Neural Network
(NLP): Natural Language Processing
(BiLSTM): Bidirectional LSTM
(GRU): Gated Recurrent Unit
(BiGRU): Bidirectional GRU
(BiRNN): Bidirectional Recurrent Neural Network
(SVM): Support Vector Machine
(BOW): Bag-Of-Words
(CBOW): Continuous Bag-Of-Words
(ReLU): Rectified Linear Unit
(GD): Gradient Descent
(SGD): Stochastic Gradient Descent
(HAN): Hierarchical Attention Network
(LIWC): Linguistic Inquiry Word Count
(TF-IDF): Term-Frequency-Inverse Document Frequency

Bibliography

[1] *Diagnostic and statistical manual of mental disorders (5th ed.).* American Psychiatric Association, 2013.

[2] The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of Research in Personality*, 47(3):218 – 225, 2013.

[3] N. Aida Osman and S. Azman Mohd Noah. Sentiment-based model for recommender systems. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6, 2018.

[4] Mohammed Al-Mosaiwi and Tom Johnstone. *In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation.* 2018.

[5] Fergadis Aris, Christos Baziotis, Dimitris Pappas, Haris Papageorgiou, and Alexandros Potamianos. Hierarchical bidirectional attention-based rnn in biocreative vi precision medicine track, document triage task. 2017.

[6] Richard Bagozzi, Mahesh Gopinath, and Prashanth Nyer. The role of emotions in marketing. *Journal of the Academy of Marketing Science*, 27:184–206, 04 1999.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural machine translation by jointly learning to align and translate.* 2015.

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. *Proc. of COMPSTAT*, 01 2010.

[9] S Watanabe JR Hershey C Hori, T Hori. *Context-Sensitive and Role-Dependent Spoken Language Understanding using Bidirectional and Attention LSTMs.* Interspeech 2016, 2016.

[10] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[11] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on socially shared cognition*, 1991.

[12] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks, 2016.

[13] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, page 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[14] Laurence Devillers, Lori Lamel, and Ioana Vasilescu. Emotion detection in task-oriented spoken dialogues. pages III– 549, 08 2003.

[15] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension, 2016.

[16] Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 307–310, New York, NY, USA, 2015. Association for Computing Machinery.

[17] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm Vries, Aaron Courville, and Y. Bengio. Feature-wise transformations. *Distill*, 3, 07 2018.

[18] Anastasia Pampouchidou et al. *Depression Assessment by Fusing High and Low Level Features from Audio, Video and Text.* Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, http://dx.doi.org/10.1145/2988257.2988266, 2016.

[19] J. R. Williamson et al. *Detecting Depression using Vocal, Facial and Semantic Communication Cues.* Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, https://dl.acm.org/citation.cfm?doid=2988257.2988263, 2016.

[20] Le Yang et al. *Decision Tree Based Depression Classification from Audio, Video and Language Information.* Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, http://dx.doi.org/10.1145/2988257.2988269, 2016.

[21] Ze'ev Frankel and Heidi Levitt. Clients' experiences of disengaged moments in psychotherapy: A grounded theory analysis. *Journal of Contemporary Psychotherapy*, 39:171–186, 09 2008.

[22] Michael Gamon, Munmun Choudhury, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Association for the Advancement of Artificial Intelligence*, 2013.

[23] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with lstm. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 2, pages 850–855 vol.2, 1999.

[24] J. Gibson, D. Atkins, T. Creed, Z. Imel, P. Georgiou, and S. Narayanan. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, 2019.

[25] James Gibson, Nikolaos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, pages 1947–1951.

[26] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

[27] Judith Hall, Jinni Harrigan, and Robert Rosenthal. Nonverbal behavior in clinician—patient interaction. *Applied and Preventive Psychology*, 4:21–37, 12 1995.

[28] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[30] Leonard M. Horowitz, Harold Sampson, Ellen Y. Siegelman, Aaron Wolfson, and Jakob Weiss. On the identification of warded-off mental contents: an empirical and methodological contribution. *Journal of abnormal psychology*, 84 5:545–58, 1975.

[31] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017.

[32] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. pages 168–177, 08 2004.

[33] Zac E. Imel, Mark Steyvers, and David C. Atkins. Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52 1:19–30, 2015.

[34] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. *Monitoring Tweets for Depression to Detect At-risk Users.* Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, 2017.

[35] Angela Jeffrey and Terene McMah. Counsellor facial expression and client-perceived rapport. *Counselling Psychology Quarterly*, 19:343–356, 12 2006.

[36] Donald J. Kiesler. An interpersonal communication analysis of relationship in psychotherapy. *Psychiatry*, 42(4):299–311, 1979. PMID: 504511.

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[38] Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50, 2014.

[39] Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50, 08 2014.

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[41] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding, 2017.

[42] Jackson Liscombe, Giuseppe Riccardi, and Dilek Hakkani-Tur. Using context to improve emotion detection in spoken dialog systems. pages 1845–1848, 01 2005.

[43] Bing Liu and Ian Lane. *Dialog Context Language Modeling with Recurrent Neural Networks*. 2017.

[44] David Losada and Pablo Gamallo. Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, pages 1–24, 08 2018.

[45] Yi Luan, Yangfeng Ji, and Mari Ostendorf. LSTM based conversation models. *CoRR*, abs/1603.09457, 2016.

[46] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN for emotion detection in conversations. *CoRR*, abs/1811.00405, 2018.

[47] Nikos Malandrakis and Shrikanth S. Narayanan. Therapy language analysis using automatically generated psycholinguistic norms. In *INTERSPEECH*, pages 1952–1956, 2015.

[48] Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Proc. Interspeech 2019*, pages 221–225, 2019.

[49] Lena Mallon and Jerker Hetta. Detecting depression in questionnaire studies: Comparison of a single question and interview data in community sample of older adults. *European Journal of Psychiatry*, 16:135–144, 07 2002.

[50] Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. Attention-based conditioning methods for external knowledge integration. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3944–3951. Association for Computational Linguistics, 2019.

[51] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[52] Saif Mohammad and Peter Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29, 2013.

[53] Nilly Mor and Jennifer Winquist. Self-focused attention and negative affect: A meta-analysis. *Psychological bulletin*, 128:638–62, 2002.

[54] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. *Deep Learning for Depression Detection of Twitter Users*. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, Association for Computational Linguistics, 2018.

[55] World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates.* 2017.

[56] Sungjoon Park, Donghyun Kim, and Alice Oh. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1448–1459, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[57] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. *Automatic differentiation in pytorch.* 2017.

[58] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[59] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[60] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.

[61] Andrija Perušić, Denis Kustura, and Ivan Matak. Using linguistic metadata for early depression detection in social media. 2018.

[62] P. Philippot, R. S. Feldman, and E. J. Coats. The role of nonverbal behavior in clinical settings: Introduction and overview. *Series in affective science. Nonverbal behavior in clinical settings*, pages 3–13, 2003.

[63] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79:284–95, 06 2011.

[64] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony of head- and body-movement in psychotherapy: Different signals have different associations with outcome. *Frontiers in Psychology*, 5:979, 09 2014.

[65] Nairán Ramírez-Esparza, Cindy Chung, Ewa Kacewicz, and James Pennebaker. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. 01 2008.

[66] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. *Language use of depressed and depression-vulnerable college students*. Journal Cognition and Emotion, Pages 1121-1133, 2004.

[67] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.

[68] Benjamin J. Sadock and Virginia A. Sadock. *Kaplan and Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry (9th ed.).* Wolters Kluwer, 2002.

[69] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. *Towards Assessing Changes in Degree of Depression through Facebook.* Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, 2014.

[70] Chenlin Shen, Changlong Sun, Jingjing Wang, Yangyang Kang, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. Sentiment classification towards question-answering with hierarchical matching network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3654–3663. Association for Computational Linguistics, October-November 2018.

[71] Abhinav Dhall Shubham Dham, Anirudh Sharma. *Depression Scale Recognition from Audio, Visual and Text Analysis.* https://arxiv.org/pdf/1709.05865.pdf.

[72] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[73] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432. Association for Computational Linguistics, 2015.

[74] Yla Tausczik and James Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54, 2010.

[75] Brett Thombs, Roy Ziegelstein, and Mary Whooley. Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: Data from the heart and soul study. *Journal of general internal medicine*, 23:2014–7, 09 2008.

[76] Marcel Trotzek, Sven Koitka, and Christoph Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32:588–601, 2018.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. *Attention is all you need.* 2017.

[78] Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.

[79] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[80] Hadas Wiseman and Laura Rice. Sequential analyses of therapist-client interaction during change events: A task-focused approach. journal of consulting and clinical psychology, 57, 281-286. *Journal of consulting and clinical psychology*, 57:281–6, 05 1989.

[81] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. Narayanan. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. 2020.

[82] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm, 2016.

[83] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics, 2016.

[84] Zhou Yu, Stefen Scherer, David DeVault, Jonathan Gratch, Giota Stratou, Louis-Philippe Morency, and Justine Cassell. Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs. 2013.

[85] Victoria Zayats and Mari Ostendorf. Conversation modeling on reddit using a graph-structured lstm. *Transactions of the Association for Computational Linguistics*, 6:121–132, 2018.

[86] Finn Årup Nielsen. A new anew: evaluation of a word list for sentiment analysis in microblogs. pages 93–98, 2011.