ΣΤΡΑΤΙΩΤΙΚΗ ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ
Τμήμα Στρατιωτικών Επιστημών

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΑΚΑΔΗΜΑΪΚΟΥ ΕΤΟΥΣ 2017-18

ΣΧΕΔΙΑΣΗ & ΕΠΕΞΕΡΓΑΣΙΑ ΣΥΣΤΗΜΑΤΩΝ
(Systems Engineering)
(ΠΔ 97 /2015/ΦΕΚ 163Α'/20.08.2014)

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Εφαρμοσμένη Τεχνητή Νοημοσύνη σε Μεγάλου Όγκου Δεδομένα για Χρήση Επιλογής Προσωπικού σε Εταιρεία HR

ΒΑΣΙΛΕΙΟΣ ΓΕΩΡΓΙΑΔΗΣ

Α.Μ.:2016018051

Απρίλιος 2022

Η Μεταπτυχιακή Διατριβή του Γεωργιάδη Βασιλείου εγκρίνεται:

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Καραδήμας Νικόλαος (Επιβλέπων),
Επίκουρος Καθηγητής (ΣΣΕ)
Στρατιωτική Σχολή Ευελπίδων ……..

Καρανάσιου Ειρήνη,
Καθηγήτρια (ΣΣΕ)
Στρατιωτική Σχολή Ευελπίδων ……

Τσαφαράκης Στυλιανός,
Αναπληρωτής Καθηγητής,
Πολυτεχνείο Κρήτης

Stelios Tsafarakis
Digitally signed by Stelios Tsafarakis
DN: c=GR, l=Chania, o=Technical University of Crete, ou=Class B - Private Key created and stored in software CSP, sn=Tsafarakis, givenName=Stelios, serialNumber=4644645843, cn=Stelios Tsafarakis, email=tsafarakis@dpem.tuc.gr
Date: 2022.05.09 13:36:48 +03'00'
……

Απρίλιος 2022

Οι αρχές και οι τεχνικές που αναφέρονται στην παρούσα εργασία εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν επίσημες θέσεις της Στρατιωτικής Σχολής Ευελπίδων ή του Πολυτεχνείου Κρήτης. Είναι, οι βέλτιστες διαθέσιμες κατά το χρόνο της συγγραφής αυτής της έρευνας.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την ΣΣΕ καθώς και το Πολυτεχνείο της Κρήτης για το μεταπτυχιακό αυτό πρόγραμμα το οποίο με βοήθησε να διευρύνω τους πνευματικούς μου ορίζοντες με τις πολύτιμες γνώσεις που με προσέφερε.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή Καραδήμα Νικόλαο για την πολύτιμη καθοδήγησή του. Χωρίς αυτή δεν θα ήταν δυνατή η δημιουργία της συγκεκριμένης διατριβής.

Ακόμα, θα ήθελα να ευχαριστήσω την Καθηγήτρια Καρανάσιου Ειρήνη γιατί το μάθημά της ήταν εκείνο που λειτούργησε ως έναυσμα για μένα για την κατανόηση του πεδίου εφαρμογής των γνώσεων που προσφέρει το εν λόγω πρόγραμμα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την στήριξη που μου έδωσε σε αυτό το όμορφο ταξίδι γνώσεων που προσφέρει το συγκεκριμένο αυτό μεταπτυχιακό πρόγραμμα.

Ακρωνύμια

AI	Artificial Intelligence
BPTT	Backpropagation Through Time
CBOW	Continuous Bag of Words
HDFS	Apache Hadoop File System
HR	Human Resource
LSTM	Long Short-Term Memory
KNIME	Konstanz Information Miner
NLP	Natural Language Processing
RNN	Recurrent Neural Network

Περιεχόμενα

Πίνακας Εικόνων

Περίληψη

Τα μεγάλα δεδομένα χρησιμοποιούνται για τη βελτίωση της λήψης αποφάσεων, την παροχή ιδεών, την ανακάλυψη και την υποστήριξη, αλλά πλέον και για βελτιστοποίηση των διαδικασιών σε όλο το εύρος της βιομηχανίας. Η συγκεκριμένη εργασία θα ασχοληθεί με την αναγνώριση τύπου προσωπικότητας χρησιμοποιώντας λεκτικά δεδομένα των χρηστών από τα μέσα κοινωνικής δικτύωσης για ανάγκες εταιρειών ανθρώπινου δυναμικού. Το μοντέλο που θα χρησιμοποιηθεί για την λύση του προβλήματος θα είναι ένα νευρωνικό δίκτυο βαθιάς μάθησης. Η βασική διαφορά του μοντέλου αυτού από τα κλασικά μοντέλα βαθιάς μάθησης είναι ότι στην συγκεκριμένη περίπτωση θα γίνει εκμετάλλευση μια αρχιτεκτονικής μοντέλου, με έναν τύπο μνήμης όπως είναι η Long Short-Term Memory (LSTM) Reccurent Neural Network), η οποία έχει αποδειχθεί ότι είναι πολύ καλύτερη από τις κλασικές αρχιτεκτονικές ειδικά εφόσον θα πρέπει να αντιμετωπιστούν προβλήματα σχετικά με αλληλουχίες (sequential problems) όπου στην συγκεκριμένη περίπτωση η ανάλυση λεκτικού περιεχομένου είναι ένα πρόβλημα αλληλουχίας.

Τέλος ο απώτερος σκοπός αυτής της εργασίας είναι η λύση του προβλήματος το οποίο δημιουργείται στα τμήματα Ανθρώπινου Δυναμικού κατά το στάδιο επιλογής προσωπικού. Συγκεκριμένα οι επιχειρήσεις με μεγάλο όγκο υποψήφιων εργαζομένων, χρησιμοποιώντας τον αλγόριθμο μηχανικής μάθησης, δύνανται να εξοικονομήσουν χρόνο και χρήμα, λαμβάνοντας την εικόνα ψυχογραφήματος των υποψηφίων, η οποία πραγματοποιείται συνήθως από εξωτερικά συνεργαζόμενους ψυχολόγους.

Abstract

Big Data are used for improvement in decision making, but also help in creation support and optimization of new procedures in the whole industry. This paper will show the prediction of personality for candidates of HR firms by using machine learning. Through the feedback of each candidate from the interaction in social media accounts they will be delegated to one of the 20 personalities of potential employees that the science of Human Resources recognizes.

The model that it will be used for solving the problem mentioned above will be a deep learning neural network. The basic difference of this model in contrast to others is that in this case an architecture is used with a Long Short-Term Memory Recurrent Neural Network, which is proven to be much better than the classical ones since we are dealing with sequential problems. In this case the analysis of a verbal content of is a sequential problem.

Finally, the ultimate purpose of this work is to solve the problem that is created in the HR departments of large companies during the selection stage. Specifically, companies with a large volume of potential employees, using the machine learning algorithm, can save time and money by taking the psychogram image of the candidates, which is usually performed by externally affiliated psychologists.

1 Εισαγωγή

Είναι γεγονός ότι την τελευταία 20ετία ο ρόλος του τμήματος προσωπικού διαδραματίζει ολοένα και σημαντικότερο ρόλο στις εταιρείες. Αυτό οφείλεται στην εμπλοκή του σε ολοένα και περισσότερων τομέων των εταιρειών όπου και σε πολλές περιπτώσεις καταλήγει να είναι ο κόμβος επικοινωνίας μεταξύ των διάφορων τομέων. Η εξέλιξη των εταιρειών έχει οδηγήσει και στην εξέλιξη της έννοιας του ανθρώπινου δυναμικού.

Μπορεί να μην φαίνεται η σπουδαιότητα του έργου που παράγει ο συγκεκριμένος κλάδος σε εταιρείες μικρού μεγέθους, όσον αφορά το προσωπικό που απασχολούν, αλλά στις εταιρείες με μεγάλο αριθμό υπαλλήλων, ο ρόλος του τμήματος ανθρώπινου δυναμικού είναι πλέον ζωτικής σημασίας για την ευημερία μιας επιχείρησης.

Μέχρι πρότινος το πρόβλημα που καλείτο να αντιμετωπίσει ο καθένας ήταν η πρόσβαση στην πληροφορία. Όταν αυτό ξεπεράστηκε μέσω του παγκόσμιου ιστότοπου αναδείχθηκε ένα νέο πρόβλημα. Πιο συγκεκριμένα, από την δυσκολία της πρόσβασης στην πληροφορία έγινε η μετάβαση στην διαχείριση των δεδομένων μεγάλης κλίμακας. Αυτός ήταν και ο λόγος της δημιουργίας των Big Data καθώς όμως και των εργαλείων διαχείρισης αυτών όπως αρχικά η Τεχνητή Νοημοσύνη. Η μετεξέλιξη αυτής οδήγησε σε πρώτη φάση στη Μηχανική Μάθηση και μετέπειτα στην Βαθιά Μάθηση. Αναλυτική περιγραφή των προαναφερθέντων όρων, μεταξύ άλλων, ακολουθεί στο κεφάλαιο 2.

Ο όρος Ανθρώπινο Δυναμικό (Human Resource - HR) είναι ένας γενικός όρος ο οποίος καλύπτει ένα ευρύ φάσμα εργασιών. Αρκεί να αναλογιστεί κανείς ότι ορισμένα από τα καθήκοντα των εργαζομένων που καλύπτει αυτό το τμήμα είναι η πρόσληψη και απόλυση εργαζομένων, η δημιουργία εταιρικής ταυτότητας, η διευθέτηση συγκρούσεων στον χώρο εργασίας, η συμμόρφωση της εταιρίας με την εργασιακή νομοθεσία, η διαχείριση καταγγελιών από τους εργαζομένους, η συμμόρφωση της εταιρίας με τις προδιαγραφές ασφαλείας, η πληρωμή, επιβράβευση και εκπαίδευση των εργαζομένων καθώς την εύρεση των κατάλληλων υποψηφίων για την στελέχωση της εκάστοτε θέσης που μπορεί να προκύψει. Ειδικά το τελευταίο κομμάτι της στελέχωσης είναι και αυτό στο οποίο θα δοθεί έμφαση στη συγκεκριμένη διατριβή. Με άλλα λόγια το Ανθρώπινο Δυναμικό δεν αποτελείται από μια δραστηριότητα ή λειτουργία. Βασικά ο όρος HR αναφέρεται σε όλα όσα έχουν να κάνουν με την σχέση εργοδότη-εργαζόμενου, τόσο άμεσα όσο και έμμεσα.

Αν και για πολλά χρόνια θεωρείτο υποστηρικτική λειτουργία, τα τελευταία χρόνια το HR παίρνει μια πιο στρατηγική μορφή στον εταιρικό κόσμο καθώς τα διευθυντικά στελέχη

αναγνώρισαν τους υπαλλήλους σαν πηγή συγκριτικού πλεονεκτήματος. Εταιρίες που εφάρμοσαν πρακτικές HR οι οποίες κατάφεραν να δημιουργήσουν ισχυρή εταιρική κουλτούρα και καλύτερο εργασιακό περιβάλουν πέτυχαν εμφανή καλύτερα οικονομικά αποτελέσματα καθώς είχαν αυξημένη καινοτομία, επίτευξη στόχων, και υψηλή αποδοτικότητα από πλευράς υπαλλήλων. Ταυτόχρονα η παγκοσμιοποίηση έκανε πιο περίπλοκο τον ρόλο του HR για εταιρίες που δραστηριοποιούνται σε περισσότερες από μια χώρες, ενώ η τεχνολογία δημιούργησε μια γκάμα ευκαιριών για αυτοματοποίηση της διοικητικής και πρακτικής εφαρμογής του HR σε όλους τους τομείς. Όπως είναι αναμενόμενο οι ευθύνες και δραστηριότητες των ανθρώπων του HR διαφέρουν ανάλογα με το μέγεθος της εταιρίας. Σε μία μικρή επιχείρηση ο υπεύθυνος ανθρώπινου δυναμικού έχει γενικά καθήκοντα ενώ σε μεγάλες εταιρίες θα συναντήσει κανείς πολλούς ρόλους HR, τόσο ειδικούς όσο και γενικούς. Πολλές μεσαίες και μικρές επιχειρήσεις καταφεύγουν στην συνεργασία με εξωτερικούς συνεργάτες για κάποιες λειτουργίες ή το σύνολο του τμήματος HR έχοντας δημιουργήσει έτσι μία βιομηχανία πολλών εκατομμυρίων ευρώ.

Όσο η τεχνολογία εξελίσσεται, τόσο επεκτείνονται και οι εφαρμογές της σε πολλούς κλάδους της βιομηχανίας όπου κύριο στόχο αποτελεί η ελαχιστοποίηση του κόστους σε συνδυασμό με την αναβάθμιση των υπηρεσιών που προσφέρονται. Το ίδιο ισχύει και για τον τομέα του ανθρώπινου δυναμικού και ειδικά την τελευταία δεκαετία. Είναι από τα τμήματα σε μια εταιρεία που καλείται να διαχειριστεί πραγματικά τεράστιο όγκο δεδομένων.

Αυτός ο τεράστιος όγκος δεδομένων έφτασε να καθιστά την λειτουργία αυτού του τμήματος απαγορευτική ως προς την προσέγγιση του κόστους συντήρησής του, σε τέτοιο βαθμό όπου η αγορά ανάγκασε την δημιουργία ενός νέου κλάδου, αυτού του HR, γιατί η σημασία του παρέμενε και παραμένει αν μη τι άλλο το ίδιο. Πλέον στρέφονται οι επιχειρήσεις σε εταιρείες HR να την επίλυση του κυριότερου προβλήματός τους, δηλαδή την εξεύρεση των κατάλληλων υποψηφίων για την επάνδρωση των θέσεων εργασίας που προσφέρονται. Είναι μία αντικειμενικά δύσκολη διαδικασία που καλούνται οι εταιρείες HR πλέον να φέρουν εις πέρας για λογαριασμό τόσο των πελατών τους όσο και την υποψηφίων για την εκάστοτε θέση.

Τόσο η ζήτηση όσο και η προσφορά όμως σε αυτόν τον κλάδο δημιουργούν με την σειρά τους νέα προβλήματα για τις εταιρείες ανθρώπινου δυναμικού. Σε αυτό το σημείο θα πρέπει να αναφερθεί πως αρκετά μεγάλο εύρος των υπηρεσιών τους πλέον οι εταιρείες HR το αναθέτουν και αυτές με την σειρά τους σε ένα πλήθος εξωτερικών συνεργατών τους μιας και ο όγκος δουλειάς δεν είναι διαχειρίσιμος. Προσεγγίζοντας το προαναφερόμενο πρόβλημα χρηματοοικονομικά, γίνεται άμεσα αντιληπτό για το ποιο το μέγεθος του κόστους το οποίο δημιουργείται σε όλες τις εμπλεκόμενες εταιρείες. Μόνο οι εργατοώρες που χρειάζονται για

να προκύψει το εκάστοτε αποτέλεσμα φτάνουν να είναι απαγορευτικές από πλευράς κόστους. Αυτό το πρόβλημα πρόκειται να προσεγγιστεί-αναλυθεί και τελικά να επιλυθεί στην συγκεκριμένη διατριβή, όπου με την χρήση των δεδομένων μεγάλης κλίμακας και αξιοποιώντας την μηχανική μάθηση, θα παρουσιαστεί η πλεόν γρήγορη κατηγοριοποίηση υποψηφίων στα προφίλ των προσωπικοτήτων που χρησιμοποιεί πλέον η επιστήμη του HR για την ταξινόμηση και τον έλεγχο καταλληλόλητας αυτών σε πιθανές θέσεις εργασίας. Αισιόδοξο φαίνεται να είναι το χαμηλό εύρος αστοχίας που παρουσιάζει ο αλγόριθμος που θα αναλυθεί στα κεφάλαια 4 και 5. Στην συγκεκριμένη διατριβή θα φανεί πώς μπορεί να βοηθηθεί μια εταιρεία HR στην λήψη αποφάσεων αξιοποιώντας προς όφελός της το μέχρι τώρα πρόβλημά της και πιο συγκεκριμένα τα δεδομένα μεγάλης κλίμακας με την βοήθεια της μηχανικής μάθησης, αποκτώντας έτσι σημαντικό πλεονέκτημα έναντι των ανταγωνιστών της.

Αρκεί να σκεφτεί κανείς ότι η επιστήμη του ανθρώπινου δυναμικού ξεκίνησε έχοντας 4 προσωπικότητες εργαζομένων αλλά ο συνδυασμός της εξέλιξής της αγοράς και των διευρυμένων αναγκών της οδήγησε το HR στην δημιουργία παραπάνω από 16 επιπλέον προσωπικοτήτων οι οποίες θα συνεχίσουν να αυξάνονται, ακόμα και στο άμεσο μέλλον. Επομένως είναι άμεσα αντιληπτή και η πολυπλοκότητα του θέματος. Για αυτό και θα παρουσιαστεί ο τρόπος επίλυσης αυτού του προβλήματος με την χρήση ενός ανατροφοδοτούμενου αλγορίθμου ο οποίος σαν είσοδο θα χρησιμοποιεί τα σχόλια των υποψηφίων από τους λογαριασμούς τους στα μέσα μαζικής δικτύωσης για μία θέση, τα οποία στην πραγματικότητα θα είναι και τα δεδομένα μεγάλης κλίμακας. Για τις ανάγκες της εργασίας θεωρούμε σαν δεδομένο ότι υπάρχει η συναίνεσή των υποψηφίων για την χρήση των δεδομένων από τους λογαριασμούς τους στα μέσα κοινωνικής τους δικτύωσης, από οπού και με την βοήθεια της μηχανικής μάθησης θα δίνεται η δυνατότητα στον αλγόριθμο να αξιολογεί τα παραπάνω στοιχεία, και μέσω αυτών να κατατάσσει τους υποψηφίους σε μία από τις προσωπικότητες της επιστήμης του ανθρώπινου δυναμικού.

Συμπερασματικά, στο επόμενο κεφάλαιο θα γίνει αναλυτική περιγραφή και διεξοδική ανάλυση των όρων που θα χρησιμοποιηθούν ώστε να είναι εμφανές το σκεπτικό που ακολουθήθηκε στην αναγνώριση, προσέγγιση και λύση του προβλήματος. Επιπρόσθετα, στο κεφάλαιο 3 θα γίνει η παρουσίαση του προβλήματος καθώς και η ανάλυση αυτού. Στο κεφάλαιο 4 ακολουθεί η λύση και ο σχεδιασμός της υλοποίησης της και στη συνέχεια, στο κεφάλαιο 5 θα γίνει συζήτηση για τα αποτελέσματα όπου θα φανούν και τα συμπεράσματα της έρευνας. Τέλος, στο κεφάλαιο 6 ακολουθούν τα συμπεράσματα της διπλωματικής εργασίας.

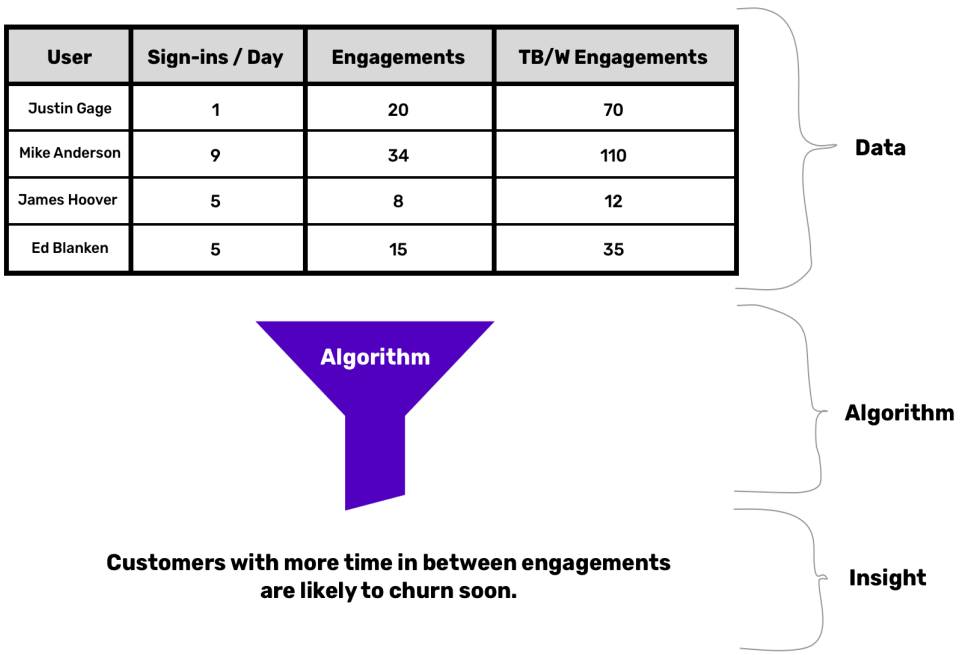
2 Ανάλυση Εννοιών

2.1 Δομή Δεδομένων/Αλγορίθμου

Οποιοδήποτε πρόβλημα Machine Learning μπορεί να αναλυθεί σε 2 βασικά μέρη:

- Τα δεδομένα
- Ο αλγόριθμος

Οποιεσδήποτε άλλες εκδοχές που μπορεί να ακούσει κάποιος όπως η βαθιά εκμάθηση, η κατάβαση λόφων (Gradient Descent) και η ενισχυτική μάθηση - είναι απλώς παραλλαγές σε αυτά τα δύο βασικά κομμάτια. Αν ποτέ υπάρξει σύγχυση ή χαθεί κανείς μέσα στο χάος των όρων που υπάρχουν, απλά αρκεί να αναρωτηθεί αν το πρόβλημα που συναντιέται έχει να κάνει με τα δεδομένα ή τον αλγόριθμό.



Εικόνα 2. 1 Η διάρθρωση του συστήματος μας

2.1.1 Δεδομένα

Το κομμάτι της διαχείρισης δεδομένων της μηχανικής μάθησης αναφέρεται στην προσπάθεια πρόβλεψης και με στον τρόπο τον οποίο σκοπεύει κανείς να επεξεργαστεί τα δεδομένα ώστε να εκπαιδευτεί ο υπολογιστής. Ένα παράδειγμα που μπορεί να βοηθήσει στην καλύτερη κατανόηση είναι αυτό του διαχωρισμού. Σε αυτή την περίπτωση τα δεδομένα

ενδέχεται να είναι παρελθούσες αλληλεπιδράσεις χρηστών. Τα δεδομένα είναι συνήθως οργανωμένα ως σειρές, με στήλες που αντιπροσωπεύουν χαρακτηριστικά των δεδομένων μας (βλ.Εικόνα 2.1).

Στην ουσία οι μέθοδοι εκμάθησης θα προσπαθήσουν να βρουν σχέσεις ανάμεσα στα δεδομένα. Ενδέχεται να εντοπιστούν μερικές διαφορετικές σχέσεις:

1. Οι χρήστες με μεγάλες χρονικές περιόδους μεταξύ των δεσμεύσεων είναι πιθανό να παρουσιάσουν μεταβολές σε σχέση με τις αρχικές τους προτιμήσεις.
2. Οι χρήστες με μεγάλο αριθμό δεσμεύσεων είναι απίθανο να παρουσιάσουν μεταβολές.

2.1.2 Ο Αλγόριθμος

Τα δεδομένα θα μπορούσε να πει κανείς ότι είναι συνθετικά, και ο αριθμός των πραγματικών σχέσεων μπορεί πραγματικά να υπάρχει μέσα στα ίδια τα δεδομένα. Στη Μηχανική Μάθηση, ο αλγόριθμος είναι στην πραγματικότητα η μέθοδος που χρησιμοποιείται για να βρεθούν αυτές σχέσεις μέσα στα δεδομένα. Οι αλγόριθμοι μπορούν να είναι σύνθετοι ή απλοί, μεγάλοι ή μικροί, ή μπορεί να παρουσιάζουν η οποιαδήποτε μεταβολή των χαρακτηριστικών, αλλά αναφερόμενοι στον πυρήνα, είναι απλά τρόποι να υπολογιστεί το "τι". Όσον αφορά τη βαθιά μάθηση, είναι στην ουσία ένας τύπος αλγορίθμου. Όπως ο αλγόριθμος MergeSort είναι αποτελεσματικός στην ταξινόμηση πινάκων, οι αλγόριθμοι Machine Learning είναι αποτελεσματικοί στην εμφάνιση σχέσεων και συσχετίσεων.

Διαφορετικοί τύποι αλγορίθμων μπορούν να βοηθήσουν να επιτευχθούν διαφορετικοί στόχοι. Εάν θελήσει κανείς, είναι δυνατόν να εξηγήσει τις σχέσεις που βρίσκονται στην ανθρώπινη ομιλία. Ένας απλός αλγόριθμος όπως η γραμμική παλινδρόμηση είναι ίσως μια καλή επιλογή. Εάν υπάρχει ενδιαφέρον περισσότερο για την ακρίβεια και όχι τόσο στο να δίνεται έμφαση στην ερμηνεία, τα νευρωνικά δίκτυα μπορούν να επιτύχουν υψηλότερα ποσοστά ακρίβειας. Αυτό καλείται να λύσει η ανταλλαγή μεταξύ ακρίβειας και εξήγησης, που οδηγεί σε μια σημαντική απόφαση προϊόντος που πολλοί επιστήμονες δεδομένων πρέπει να κάνουν. Οτιδήποτε άλλο προκύπτει στον κόσμο του ML έχει να κάνει με ένα από αυτά τα δύο πράγματα δηλαδή κλιμάκωση χαρακτηριστικών και τροποποίηση των δεδομένων. Συνεχίζοντας ένας τύπος αλγορίθμου είναι η Διασταυρωμένη Επικύρωση (Cross Validation). Λειτουργεί σαν ένας τρόπος βελτίωσης του αλγορίθμου.

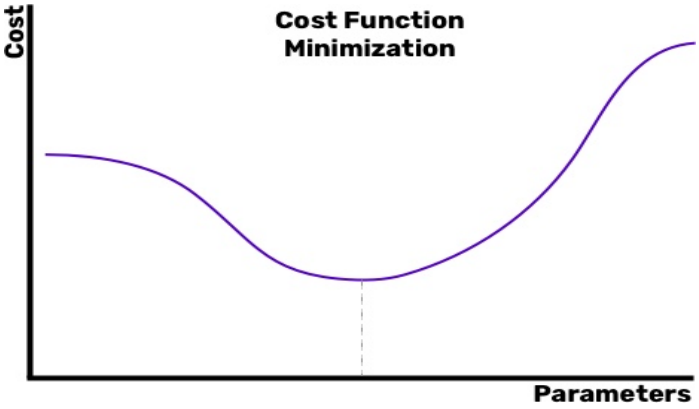
2.1.3 Πως Λειτουργεί?

Οι έννοιες πίσω από το πώς λειτουργεί η Μηχανική Μάθηση είναι στην πραγματικότητα πολύ απλές, ακόμα κι αν οι υποκείμενοι αλγόριθμοι μπορούν να γίνουν πολύπλοκοι. Η πλειοψηφία των αλγορίθμων Μηχανικής Μάθησης χρησιμοποιεί μια μέθοδο για να βρει εκείνες τις σχέσεις για τις οποίες έγινε αναφορά και κυριολεκτικά ψάχνει γύρω στο σκοτάδι. Είναι τεχνικά το αποκαλούμενο Gradient Descent (γραμμική παλινδρόμηση). Η μέθοδος είναι απλή – γίνεται εκκίνηση σε κάποιο τυχαίο σημείο και προσπαθούμε να βελτιώσουμε τις προβλέψεις μας.

Αρχικά χρησιμοποιείται με μια συνάρτηση κόστους γιατί είναι ένας τρόπος για να μετρηθεί πόσο καλά εξελίσσεται η εύρεση των πιθανών σχέσεων. Με άλλα λόγια, βοηθά να υπολογιστεί πόσο μακριά είναι οι προβλέψεις που έχουν γίνει μεταξύ τους, ώστε να μπορέσουν να γίνουν οι απαραίτητες ενέργειες για τη βελτίωσή τους. Η συνάρτηση κόστους διαφέρει ανάλογα με τον αλγόριθμο που χρησιμοποιείται.

Αυτό μπορεί να φαίνεται περίπλοκο στην αρχή, αλλά αν δοθεί εστίαση στη λειτουργία κόστους φαίνεται ότι στην πραγματικότητα είναι η ποσοτικοποίηση της διαφοράς μεταξύ του τι προβλέπει ο αλγόριθμός και ποια είναι η πραγματική τιμή της μεταβλητής στόχου. Ο στόχος πρέπει να είναι η ελαχιστοποίηση αυτής της συνάρτησης κόστους, επειδή το ζητούμενο είναι οι προβλεπόμενες τιμές να είναι όσο το δυνατόν πιο κοντά στις πραγματικές τιμές μεταβλητών στόχων.

Πώς λοιπόν επιτυγχάνεται αυτό; Ξεκινώντας με μια τυχαία σειρά προβλέψεων γίνεται προσπάθεια βελτίωσης. Εάν οι προβλέψεις είναι πολύ υψηλές, θα γίνει προσαρμογή. Αν είναι πολύ χαμηλές, θα προσαρμοστεί και πάλι. Ο αλγόριθμος κινείται στο σκοτάδι μέχρι να βρει αυτό που ψάχνει. Το ενδιαφέρον είναι ότι η Gradient Descent είναι στην πραγματικότητα ένας αλγόριθμος που εφαρμόζεται στον εαυτό του. Πιο συγκεκριμένα, χρησιμοποιεί έναν αλγόριθμο για να βελτιώσει τον ίδιο αλγόριθμο. Περισσότερες λεπτομέρειες σχετικά με τον τρόπο με τον οποίο εφαρμόζεται ο αλγόριθμος Gradient Descent εκτός από τις άλλες λύσεις για τη βελτιστοποίηση της συνάρτησης κόστους, ακολουθούν σε επόμενο κεφάλαιο.



Εικόνα 2. 2 Ελαχιστοποίηση συνάρτησης κόστους

Συνοψίζοντας χρησιμοποιείται ο Gradient Descent για την βελτιστοποίηση της συνάρτησης κόστους μιας και είναι ο τρόπος με τον οποίο ο αλγόριθμος βρίσκει την υποκείμενη σχέση στα δεδομένα. Αυτό γίνεται ξεκινώντας από μερικές τυχαίες προβλέψεις και βαθμολογώντας αργά την προσέγγιση μέχρι να φτάσει όσο το δυνατόν πιο κοντά στις πραγματικές αξίες που αναζητούνται.

2.2 Επιβλεπόμενη Μάθηση και Μάθηση Χωρίς Επίβλεψη

Υπάρχουν δύο βασικοί τύποι μηχανικής μάθησης στην πράξη που χρησιμοποιούν διαφορετικές σχέσεις δεδομένων-αλγορίθμου: Η επιβλεπόμενη μάθηση (supervised learning) και η μάθηση χωρίς επίβλεψη (unsupervised learning). Η διάκριση μεταξύ αυτών των δύο είναι λεπτή αλλά σημαντική. Η κύρια διαφορά έγκειται στο πώς παρουσιάζονται τα δεδομένα στον αλγόριθμο.

Στην επιβλεπόμενη μάθηση, καθορίζονται τα πιθανά αποτελέσματα. Στο παράδειγμα του καιρού, μπορεί να ειπωθεί ότι μια δεδομένη ημέρα μπορεί να είναι ζεστή, ψυχρή ή με μέση θερμοκρασία. Στη συνέχεια, μεταφέρονται τα δεδομένα στον αλγόριθμο ο οποίος και θα καταλάβει τι οδηγεί σε ζεστές μέρες, τι σε κρύες αλλά και τι οδηγεί σε μέτριες ημέρες. Από τότε που ορίζονται ποιες είναι οι επιλογές - ζεστό, κρύο και μέτριο - ο αλγόριθμος θα εξάγει προβλέψεις μέσα σε αυτό το πλαίσιο.

Αλλά μερικές φορές, δεν είναι πραγματικά γνωστό ποιες είναι οι επιλογές ή ακόμα και ποιες είναι οι επιθυμητές επιλογές. Η μη επιβλεπόμενη μάθηση λαμβάνει τα δεδομένα και προσπαθεί να καταλάβει ποιες είναι οι διάφορες πιθανές ομαδοποιήσεις. Ένας αλγόριθμος μάθησης χωρίς επίβλεψη μπορεί να διαπιστώσει ότι οι ομάδες με τις πιο ξεχωριστές λειτουργίες είναι ημέρες πολύ ζεστές, μέτριες και συννεφιασμένες. Αντί να χρειάζεται να

ληφθεί η απόφαση εκ των προτέρων για τις ομάδες αποτελεσμάτων και να γίνεται προσπάθεια χαρτογράφησης των σχέσεών τους, αφήνουμε τον αλγόριθμο να βρει εκείνες που αισθάνεται ότι είναι οι πιο φυσικές.

Ουσιαστικά, η διαφορά μεταξύ αυτών των δύο τύπων Machine Learning είναι το μοντέλο εξόδου-εισόδου. Στην επιβλεπόμενη μάθηση, παρέχονται στον αλγόριθμο τα X και Y και υπολογίζουν τις σχέσεις μεταξύ των δύο. Σε μάθηση χωρίς επίβλεψη, δεν υπάρχει Y - απλώς γίνεται προσπάθεια κατανόησης για την υποκείμενη οργάνωση των δεδομένων.

2.2.1 Εισαγωγή στην Επιβλεπόμενη Μάθηση

Ένα μοντέλο, είναι ένας επιβλεπόμενος αλγόριθμος (supervised algorithm) αν βασίζεται σε δεδομένα εκπαίδευσης που περιέχουν ήδη τη σωστή ετικέτα (label) για κάθε είσοδο (features). Στην ουσία αυτό που κάνει ο αλγόρυθμός είναι να εξάγει συμπεράσματα βάσει αυτής της σχέσης με απώτερο σκοπό την πρόβλεψη νέων ή μη γνωστών δεδομένων.

Οι επιβλεπόμενοι αλγόριθμοι χρησιμοποιούνται συχνά για προβλήματα ταξινόμησης (classification), όπως επισήμανση αισθήσεων, ανίχνευση αντικειμένων σε εικόνες, ανίχνευση απάτης με πιστωτικές κάρτες και φιλτράρισμα ανεπιθύμητων μηνυμάτων.

Οι δύο κύριοι τύποι επιβλεπόμενης μηχανικής μάθησης είναι η παλινδρόμηση (Regression) και η ταξινόμηση (Classification). Για παράδειγμα, χρησιμοποιείται ένα μοντέλο παλινδρόμησης για την πρόβλεψη συνεχών δεδομένων (συνεχείς μεταβλητές), όπως η πρόβλεψη των τιμών κατοικιών βάσει ιστορικών δεδομένων με συνεχή μεταβλητή τον χρόνο και τάσεων της αγοράς. Χρησιμοποιείται ένα μοντέλο ταξινόμησης για την πρόβλεψη δεδομένων με διακριτές μεταβλητές, όπως για παράδειγμα την ανάθεση διακριτών ετικετών σε ένα μοντέλο ταξινόμησης εικόνας που χαρακτηρίζει την εικόνα ως πρόσωπο ή τοπίο.

Εν ολίγοις υπάρχουν πολλοί τύποι επιβλεπόμενων αλγορίθμων και ένας από τους πιο δημοφιλείς είναι το μοντέλο Naive Bayes όπως θα φανεί στην συνέχεια με λεπτομέρεια, το οποίο αποτελεί συχνά ένα καλό σημείο εκκίνησης για τους προγραμματιστές, καθώς είναι αρκετά εύκολο να κατανοηθεί το υπόκρυφο πιθανοτικό μοντέλο και να εκτελεστεί εύκολα.

2.2.2 Δέντρα Απόφασης

Τα δέντρα απόφασης (Decision Trees) είναι επίσης ένα μοντέλο πρόβλεψης. Συναντώνται δύο τύποι δέντρων:

Α) παλινδρόμηση που παίρνει συνεχείς τιμές και

Β) μοντέλα ταξινόμησης (που παίρνουν πεπερασμένες τιμές). Χρησιμοποιούν την γνωστή στρατηγική "διαίρει και βασίλευε" (divide and conquer) που χωρίζει αναδρομικά τα δεδομένα για να δημιουργήσει το δέντρο.

Επίσης τα δέντρα απόφασης είναι από τους πιο παλιούς στον τομέα της μηχανικής μάθησης. Μάλιστα μπορούν να διαχωρίσουν εύκολα πολλαπλές κλάσεις χωρίς να μάθουν απέξω τα δεδομένα εκπαίδευσης (overfit).

2.2.3 Νευρωνικά Δίκτυα

Προχωρώντας στα νευρωνικά δίκτυα (Neural Networks) τα οποία είναι ένα μοντέλο εμπνευσμένο από τον τρόπο που τα βιολογικά νευρωνικά δίκτυα λύνουν προβλήματα και μπορούν ανεξάρτητα από το αν να επιβλέπονται ή δεν επιβλέπονται. Σημασία έχει να αναφερθεί ότι τα νευρωνικά δίκτυα που επιβλέπονται, έχουν μια γνωστή έξοδο και είναι χτισμένα σε στρώματα διασυνδεδεμένων σταθμισμένων κόμβων με ένα στρώμα εξόδου που μας δίνει μια γνωστή έξοδο, όπως μια ετικέτα εικόνας (classification). Η ταξινόμηση Naïve Bayes είναι ένας αλγόριθμος που επιχειρεί να κάνει προβλέψεις με βάση τα δεδομένα που είχαν προηγουμένως επισημανθεί χρησιμοποιώντας ένα πιθανοτικό μοντέλο (Λι, Πίτερ 2012). Τα χαρακτηριστικά γνωρίσματα είναι ανεξάρτητα το ένα από το άλλο, πράγμα που σημαίνει ότι ένα χαρακτηριστικό δεν επηρεάζει την αξία ενός άλλου χαρακτηριστικού και ότι ένα σύνολο ετικετών εξετάζεται και εκχωρείται εκ των προτέρων.

Εντελώς αφηρημένα, το Naive Bayes είναι ένα πιθανοτικό μοντέλο:

$p(C_k \mid x_1, \ldots, x_n)$ για κάθε ένα από τα K πιθανά αποτελέσματα ή κλάσεις. Ορισμένα παραδείγματα ετικετών που χρησιμοποιούνται στους ταξινομητές είναι βαθμολογίες συναισθημάτων. Αυτά μπορεί να είναι λεκτικά δεδομένα, ακέραιοι αριθμοί ή κλάσματα. Βέβαια ο αλγόριθμος δέχεται μόνο αριθμητικά δεδομένα. Για ανίχνευση αντικειμένων θα μπορούσαν να χρησιμοποιούνται ετικέτες όπως καρέκλα, τραπέζι ή γραφείο για να περιγράψετε αντικείμενα στις εικόνες. Η ανίχνευση χαρακτηριστικών αποφασίζεται εκ των προτέρων, όπως η εμφάνιση λέξεων-κλειδιών ή μήκους email στην ανίχνευση ανεπιθύμητων μηνυμάτων.

2.3 Αξιολόγηση Μοντέλου (Model Evaluation)

Όπως γίνεται αντιληπτό σε αυτό το σημείο, το να εκπαιδευτεί ένας αλγόριθμος πάνω σε δεδομένα, στην ουσία δεν αποτελεί απόδειξη για την συμπεριφορά του σε δεδομένα τα οποία δεν θα έχει ξαναδεί.

Ένας απλός τρόπος για να εξεταστεί η συμπεριφορά του αλγορίθμου σε άγνωστα δεδομένα είναι να περαστεί σε αυτόν ένα σετ δεδομένων το οποίο δεν έχει ξαναδεί, το επονομαζόμενο testing set.

Στην ουσία το testing set δεν είναι τίποτε άλλο από ένα κομμάτι των αρχικών δεδομένων (συνήθως γύρω στο 20-30% του αρχικού σετ δεδομένων, οπότε απομένει ένα ποσοστό μεταξύ 70% έως 80% για training και 20% με 30% για testing. Τονίζεται ότι ενώ μπορεί να είναι εξαρχής γνωστές οι ετικέτες του testing set ο αλγόριθμος όμως δεν τις γνωρίζει και για αυτόν τον λόγο μπαίνει στην διαδικασία να τις προβλέψει αφού πρώτα έχει εκπαιδευτεί με το training set.

Τέλος, από την φάση του model evaluation εξάγεται ένα αποτέλεσμα το οποίο περιγράφει ποσοστιαία σε πόσα δείγματα προβλέφθηκε σωστά η ετικέτα τους. Συνήθως, αυτός ο τρόπος δεν είναι αρκετά αντιπροσωπευτικός για την αξιολόγηση του μοντέλου ειδικά όταν το δείγμα των δεδομένων είναι μικρό, όμως είναι ο πιο απλός και γρήγορος.

Παρακάτω ακολουθεί η περιγραφή εκπαίδευσης ενός αλγορίθμου σε βήματα:

1. Πρώτο Βήμα: Διαχωρισμός των εκάστοτε δεδομένων σε **training** (δεδομένα εκπαίδευσης του αλγορίθμου) και **testing** (δεδομένα για την αξιολόγηση του αλγορίθμου) set
2. Δεύτερο Βήμα: Εκπαίδευση του Αλγορίθμου
3. Τρίτο Βήμα: Αξιολόγηση του αλγορίθμου χρησιμοποιώντας το testing set

Σε περίπτωση που τα δεδομένα έχουν πολύ μικρό όγκο τότε υπάρχουν άλλες μέθοδοι για την αξιολόγηση του μοντέλου. Η παραπάνω πιο πολύ θα χαρακτηριζόταν ως η πιο συμβατική μέθοδος. Μια τέτοια μέθοδος και ίσως και η πιο γνωστή ονομάζεται διασταυρωμένη επικύρωση (cross validation)

Για παράδειγμα, ο αλγόριθμος της διασταύρωσης K-fold ακολουθεί τα εξής βήματα:

1) Το σύνολο των δεδομένων εκπαίδευσης χωρίζεται σε υποσύνολα δεδομένων, ένα ως δεδομένα αξιολόγησης και τα υπόλοιπα σύνολα δεδομένων προορίζονται για εκπαίδευση. Με αυτόν τον τρόπο χρησιμοποιείται το ίδιο test σετ σε κάθε υποσύνολο που χρησιμοποιείται για την εκπαίδευση δεδομένων
2) Υπολογίζεται η τυπική απόκλιση για κάθε σετ test / training.
3) Τέλος, μετράται το ποσοστό σφάλματος του μέσου όρου σε κύκλους για την εκτίμηση της απόδοσης του μοντέλου.

2.4 Εισαγωγή στους Βελτιστοποιητές (Optimizers)

Η εύρεση της καλύτερης αριθμητικής λύσης σε ένα συγκεκριμένο πρόβλημα είναι ένα σημαντικό μέρος πολλών κλάδων στα μαθηματικά και το Machine Learning δεν αποτελεί εξαίρεση. Οι βελτιστοποιητές, σε συνδυασμό με τη συνάρτηση κόστους, είναι τα βασικά κομμάτια που επιτρέπουν στο Machine Learning να δουλέψει για τα δεδομένα που χρησιμοποιούνται.

Σε αυτό το υποκεφάλαιο θα αναφερθεί η διαδικασία βελτιστοποίησης και πώς αυτή εφαρμόζεται στο Machine Learning, καθώς και πώς οι συναρτήσεις κόστους ταιριάζουν στην εξίσωση και μερικές δημοφιλείς προσεγγίσεις.

2.4.1 Τι είναι ένας Βελτιστοποιητής

Αναλύθηκε προηγουμένως με τη συνάρτηση κόστους πόσο λανθασμένες είναι οι προβλέψεις μας. Κατά τη διάρκεια της διαδικασίας κατάρτισης, γίνεται τροποποίηση και αλλαγή στις παραμέτρους (βάρη) του μοντέλου μας για να προσπαθήσουμε να ελαχιστοποιήσουμε αυτή τη λειτουργία απώλειας και να γίνουν οι προβλέψεις όσο το δυνατόν πιο σωστές. Αλλά πώς ακριβώς γίνεται αυτό; Πώς αλλάζουν οι παράμετροι του μοντέλου μας, τόσο ποσοτικά όσο και χρονικά;

Σε αυτό το σημείο επεμβαίνουν οι βελτιστοποιητές. Πιο συγκεκριμένα, συνδέουν τη συνάρτηση κόστους και τις παραμέτρους του μοντέλου. Αυτό επιτυγχάνεται με την ενημέρωση του μοντέλου σαν απάντηση στην έξοδο της συνάρτησης κόστους. Με απλούστερους όρους, οι βελτιστοποιητές διαμορφώνουν και μορφοποιούν το μοντέλο στην ακριβέστερη δυνατή μορφή του, αλλάζοντας τα βάρη. Συγκεκριμένα, η συνάρτηση κόστους λειτουργεί σαν ο οδηγός για το έδαφος, λέγοντας στο βελτιστοποιητή αν κινείται προς τη σωστή ή λάθος κατεύθυνση, αν αναγάγουμε την έννοια τις κατάβασης λόφου.

Για ένα χρήσιμο συνθετικό μοντέλο, αρκεί να σκεφτεί κανείς έναν πεζοπόρο που προσπαθεί να κατεβεί ένα βουνό με τα μάτια του κλειστά. Είναι αδύνατο να γνωρίζει σε ποια κατεύθυνση θα κινηθεί, αλλά υπάρχει ένα πράγμα που μπορεί να ξέρει: εάν πέφτει (προχωρώντας) ή αν ανεβαίνει. Τελικά, αν συνεχίσει να κάνει βήματα που τον οδηγούν προς τα κάτω, θα φτάσει στη βάση.

Ομοίως, είναι αδύνατο να γνωρίζει κανείς ποια είναι τα βάρη του μοντέλου από την αρχή. Αλλά με κάποια trial and error practice που βασίζονται στη συνάρτηση κόστους αν για παράδειγμα η πεζοπορία είναι φθίνουσα, μπορεί να καταλήξει στον στόχο που θα θεωρηθεί η επιθυμητή βάση.

2.4.2 Παραδείγματα Βελτιστοποιητών

Είναι δύσκολο να εκτιμήσει κανείς πόσο δημοφιλής είναι η κάθοδος λόφων. Χρησιμοποιείται ακόμη και σε πολύπλοκες αρχιτεκτονικές νευρωνικών δικτυών. Η backpropagation είναι βασικά η κάθοδος λόφων που εφαρμόζεται σε ένα νευρωνικό δίκτυο.

2.4.2.1 Adagrad

Ένα χαρακτηριστικό παράδειγμα μπορεί να θεωρηθεί ο βελτιστοποιητής Adagrad (Duchi,Singer 2011) που προσαρμόζει τον ρυθμό εκμάθησης ειδικά σε μεμονωμένα χαρακτηριστικά. Αυτό σημαίνει ότι ορισμένα από τα βάρη του συνόλου δεδομένων μας θα έχουν διαφορετικούς ρυθμούς εκμάθησης από άλλους. Λειτουργεί πολύ καλά για αραιά σύνολα δεδομένων όπου λείπουν πολλά παραδείγματα εισροών. Ωστόσο, το Adagrad έχει ένα σημαντικό ζήτημα και συγκεκριμένα τον ρυθμό προσαρμοστικής μάθησης που τείνει να είναι πολύ μικρός με την πάροδο του χρόνου. Κάποιοι άλλοι βελτιστοποιητές παρακάτω προσπαθούν να εξαλείψουν αυτό το πρόβλημα.

2.4.2.2 RMSprop

Το RMSprop είναι ένας άλλος βελτιστοποιητής που στην πραγματικότητα μια ειδική έκδοση του Adagrad που ανέπτυξε ο καθηγητής Geoffrey Hinton στην τάξη των νευρικών δικτυών του. Το RMSprop είναι παρόμοιο με το Adaprop, δηλαδή άλλο ένα εργαλείο βελτιστοποίησης που επιδιώκει να λύσει μερικά από τα θέματα που αφήνει ανοιχτό το Adagrad.

2.4.2.3 Adam

Ο Adam αντιπροσωπεύει την εκτίμηση της προσαρμοστικής ροπής και είναι ένας άλλος τρόπος για να χρησιμοποιηθούν παρελθόντες βαθμίδες για τον υπολογισμό των βαθμίδων. Ο συγκεκριμένος βελτιστοποιητής χρησιμοποιεί επίσης την έννοια της ορμής προσθέτοντας κλάσματα προηγούμενων κλίσεων στην τρέχουσα. Για αυτόν το λόγο έχει γίνει αρκετά διαδεδομένος και είναι πρακτικά αποδεκτός για χρήση στην εκπαίδευση νευρωνικών δικτύων.

Είναι εύκολο να χαθεί κανείς στην πολυπλοκότητα ορισμένων από αυτούς τους νέους βελτιστοποιητές. Έμφαση πρέπει να δοθεί ότι όλοι έχουν τον ίδιο στόχο και αυτός είναι η ελαχιστοποίηση της λειτουργίας της απώλειας.

2.4.2.4 Κατάβαση Λόφων

Οποιαδήποτε συζήτηση σχετικά με τους βελτιστοποιητές πρέπει να ξεκινήσει με τον πιο δημοφιλή, και ονομάζεται κατάβαση λόφων (Gradient Descent). Αυτός ο αλγόριθμος χρησιμοποιείται σε όλους τους τύπους Μηχανικής Μάθησης καθώς και άλλων μαθηματικών

προβλημάτων που έχουν να κάνουν με βελτιστοποίηση. Είναι γρήγορος και ευέλικτος. Η λειτουργία του μπορεί να περιγραφεί ως εξής:

1. Υπολογίζεται τι θα έκανε μια μικρή αλλαγή σε κάθε μεμονωμένο βάρος στη συνάρτηση κόστους, δηλαδή ποια κατεύθυνση πρέπει να περπατήσει ο πεζοπόρος.

2. Ρυθμίζεται κάθε μεμονωμένο βάρος με βάση την κλίση του. Πιο συγκεκριμένα γίνεται ένα μικρό βήμα στην καθορισμένη κατεύθυνση.

Συνεχίζοντας γίνονται τα βήματα 1 και 2 έως ότου το κόστος γίνει το χαμηλότερο δυνατό. Το δύσκολο μέρος αυτού του αλγορίθμου και των βελτιστοποιητών εν γένει είναι η κατανόηση των διαβαθμίσεων. Οι διαβαθμίσεις είναι μαθηματικά μερικά παράγωγα και είναι ένα μέτρο αλλαγής. Συνδέουν τη λειτουργία απώλειας και τα βάρη. Δίνεται έτσι η πληροφορία ποια συγκεκριμένη λειτουργία θα πρέπει να κάνουν με τα βάρη - προσθέστε 5, αφαιρέστε το .07 ή οτιδήποτε άλλο - για να μειωθεί η απόδοση της συνάρτησης κόστους και έτσι να γίνει το μοντέλο πιο ακριβές. Ένα πρόβλημα που μπορεί να εμφανιστεί κατά τη διάρκεια της βελτιστοποίησης είναι να σταματήσει στα τοπικά ελάχιστα. Όταν πρέπει να αντιμετωπιστούν σύνολα δεδομένων μεγάλης διαστάσεως με πολλές μεταβλητές δηλαδή, είναι πιθανό να βρεθεί μια περιοχή όπου φαίνεται σαν να έχει επιτευχθεί η χαμηλότερη δυνατή τιμή για την συνάρτηση κόστους μας, αλλά στην πραγματικότητα δεν είναι κάτι άλλο από ένα τοπικό ελάχιστο. Στη φλέβα της αναλογίας των πεζοπόρων, αυτό είναι σαν να βρίσκεται σε μια μικρή κοιλάδα μέσα στο βουνό κατά την κάθοδό του. Φαίνεται ότι έχει φθάσει κανείς στους πρόποδες - η έξοδος από τους πρόποδες όμως απαιτεί αντιθέτως την αναρρίχηση . Για να αποφευχθεί να υπάρξει τελμάτωση σε τοπικά ελάχιστα, θα πρέπει να βεβαιωθεί κανείς ότι έγινε χρήση του κατάλληλου ρυθμού εκμάθησης. Υπάρχουν μερικά άλλα στοιχεία που συνθέτουν τη κατάβαση λόφων και επίσης γενικεύονται σε άλλους βελτιστοποιητές.

2.4.2.5 Στοχαστική Κατάβαση Λόφων (Stochastic Gradient Descent)

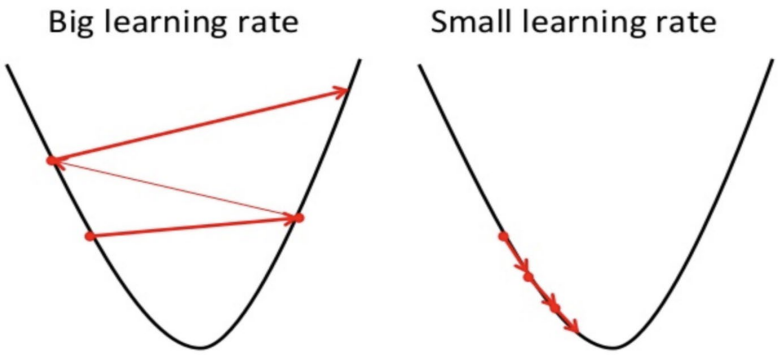
Αντί να υπολογίζει κάποιος τις διαβαθμίσεις για όλα τα παραδείγματα εκπαίδευσης του σε κάθε πέρασμα κλίσης (Ellis, 1995), είναι μερικές φορές πιο αποτελεσματικό να χρησιμοποιεί μόνο ένα υποσύνολο των εκπαιδευτικών παραδειγμάτων κάθε φορά. Η κατάκτηση στοχαστικής κλίσης είναι μια εφαρμογή που είτε χρησιμοποιεί παρτίδες παραδειγμάτων σε χρόνο είτε τυχαία παραδείγματα σε κάθε πέρασμα.

2.5 Ρυθμός Εκμάθησης (Learning Rate)

Η απότομη αλλαγή των βαρών προσθέτοντας ή αφαιρώντας δηλαδή λαμβάνοντας ακραίες τιμές μπορεί να εμποδίσει την ικανότητα να ελαχιστοποιηθεί η συνάρτηση κόστους.

Δεν είναι επιθυμητό να γίνει ένα άλμα τόσο μεγάλο ώστε να παραλειφθεί η βέλτιστη τιμή για ένα δεδομένο βάρος. Για να μην συμβεί αυτό, χρησιμοποιείται μια μεταβλητή που ονομάζεται που λειτουγεί ως ο ρυθμός εκμάθησης. Για την ακρίβεια είναι ένας πολύ μικρός αριθμός, συνήθως κάτι σαν 0.001, που πολλαπλασιάζει τις διαβαθμίσεις, για να τις κλιμακώσει. Αυτό εξασφαλίζει ότι οι τυχόν αλλαγές που μπορεί να γίνουν στα βάρη είναι σχεδόν αμελητέα μικρές.

Ταυτόχρονα, δεν θέλουμε να λάβουμε μέτρα που είναι πολύ μικρά, διότι τότε δεν μπορούμε να καταλήξουμε ποτέ με τις σωστές τιμές για τα βάρη μας. Συνοψίζοντας, ο ρυθμός εκμάθησης διασφαλίζει ότι επιτυγχάνονται αλλαγές στα βάρη με το σωστό ρυθμό, χωρίς να γίνονται αλλαγές που είναι θα πολύ μεγάλες ή πολύ μικρές.



Εικόνα 2. 3 Γραφική περιγραφή του ρυθμού εκμάθησης.

2.6 Κανονικοποίηση (Regularization)

Στη μηχανική μάθηση, υπάρχει ο φόβος της υπερφόρτωση (overfit). Η υπερφόρτωση σημαίνει απλώς ότι το μοντέλο προβλέπει καλά τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή του, αλλά εκτελεί άσχημα τα δεδομένα που έρχονται από τον πραγματικό κόσμο και δεν έχει ξαναδεί. Αυτό μπορεί να συμβεί εάν μια παράμετρος ζυγίζεται πολύ βαριά και καταλήγει να κυριαρχεί στη φόρμουλα. Η κανονικοποίηση είναι ένας όρος που προστίθεται στη διαδικασία βελτιστοποίησης που βοηθά στην αποφυγή αυτού.

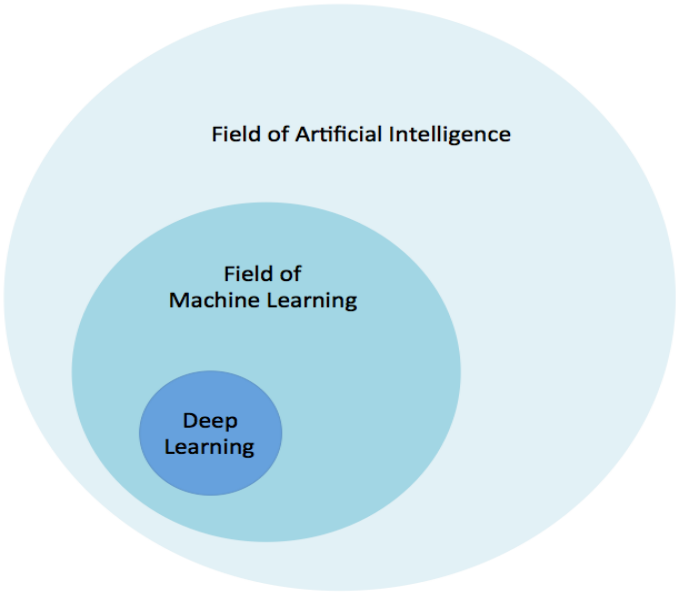
2.7 Εισαγωγή στην Βαθιά Μάθηση (Deep Learning)

Η Βαθιά Εκμάθηση είναι η αιχμή της Μηχανικής Μάθησης (Goodfellow, Bengio and Courville, 2016), και τόσο οι προγραμματιστές όσο και τα ηγετικά στελέχη των επιχειρήσεων σίγουρα πρέπει να καταλάβουν τι είναι και πώς λειτουργεί. Αυτός ο μοναδικός τύπος

αλγορίθμου έχει ξεπεράσει τα προηγούμενα σημεία αναφοράς για την ταξινόμηση εικόνων, κειμένου και φωνής. Καλύπτει επίσης μερικές από τις πιο ενδιαφέρουσες εφαρμογές στον κόσμο, όπως αυτόνομα οχήματα και μετάφραση σε πραγματικό χρόνο.

2.7.1 Τι είναι η Βαθιά Μάθηση

Για να γίνει κατανοητό το τι είναι η βαθιά εκμάθηση, πρέπει πρώτα να αποσαφηνιστεί η σχέση που έχει η βαθιά μάθηση με τη μηχανική μάθηση, τα νευρωνικά δίκτυα και την τεχνητή νοημοσύνη. Ο καλύτερος τρόπος να απεικονιστεί αυτή τη σχέση είναι με ομόκεντρους κύκλους όπως φαίνεται στην εικόνα 2.4 παρακάτω:



Εικόνα 2. 4 Σχέση Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και Βαθιάς Μάθησης

Η βαθιά μάθηση είναι ένα συγκεκριμένο υποσύνολο της Μηχανικής Μάθησης, όπου η Μηχανική Μάθηση με την σειρά της είναι ένα συγκεκριμένο υποσύνολο της Τεχνητής Νοημοσύνης. Πιο συγκεκριμένα την τελευταία δεκαετία είμαστε μάρτυρες της τρομακτικής μετεξέλιξης της Τεχνητής Νοημοσύνης. Εδώ και αρκετά χρόνια το πρόβλημα που είχε δημιουργηθεί και προσπαθούσε να ξεπεραστεί ήταν η διαχείριση και η εξαγωγή της χρήσιμης πληροφορίας απο το μεγάλο εύρος δεδομένων που υπήρχε μετά το 2010 .

Πέρα από τις κλασικές μεθόδους Μηχανικής Μάθησης που έγιναν γνωστές μέχρι το 2011, η επιστήμη της Μηχανικής μάθησης εξελίχθηκε ακόμα περισσότερο με την εισαγωγή

της βαθιάς μάθησης το 2012. Η εισαγωγή αυτού του νέου υποτομέα εξέλιξε της δυνατότητες στην ικανότητα διαχείρισης μεγάλων δεδομένων σε πολύ μεγάλο βαθμό και συνεισέφερε στην δημιουργία μοντέλων με πολύ μεγαλύτερη ακρίβεια στις προβλέψεις τους.

Ο συνδυασμός του εύρους δεδομένων μαζί με την τεχνητή νοημοσύνη δημιούργησαν τον υποτομέα της Μηχανικής Μάθησης ο οποίος έλυσε και το παραπάνω πρόβλημα. Η Μηχανική Μάθηση μπορεί να κατανοηθεί καλύτερα μέσω τεσσάρων προοδευτικών προσεγγίσεων :

1. Η ευρεία: Μηχανική μάθηση είναι η διαδικασία της πρόβλεψης των πραγμάτων, συνήθως με βάση αυτά που έχουν γίνει στο παρελθόν.
2. Η Πρακτική: Η Μηχανική Μάθηση προσπαθεί να βρει σχέσεις στα δεδομένα μας που μπορούν να μας βοηθήσουν να προβλέψουμε τι θα συμβεί στη συνέχεια.
3. Η Τεχνική: Μηχανική Μάθηση χρησιμοποιεί στατιστικές μεθόδους για την πρόβλεψη της τιμής μιας μεταβλητής στόχου χρησιμοποιώντας ένα σύνολο δεδομένων εισόδου.
4. Το Μαθηματικό: Η Μηχανική Μάθηση προσπαθεί να προβλέψει την τιμή μίας μεταβλητής Υ δεδομένης της εισαγωγής ενός συνόλου χαρακτηριστικών Χ.

Επιπρόσθετα, για να κατανοηθεί η σημασία της βαθιάς μάθησης, η υπολογιστική όραση είναι ένα καλό παράδειγμα ενός έργου που η βαθιά μάθηση έχει μετατρέψει σε κάτι ρεαλιστικό για επιχειρηματικές εφαρμογές. Η χρήση βαθιάς μάθησης για την ταξινόμηση και την ετικέτα των εικόνων δεν είναι μόνο καλύτερη από οποιονδήποτε άλλο παραδοσιακό αλγόριθμο αλλά αρχίζει να είναι αποδοτικότερη και από τους πραγματικούς ανθρώπους.

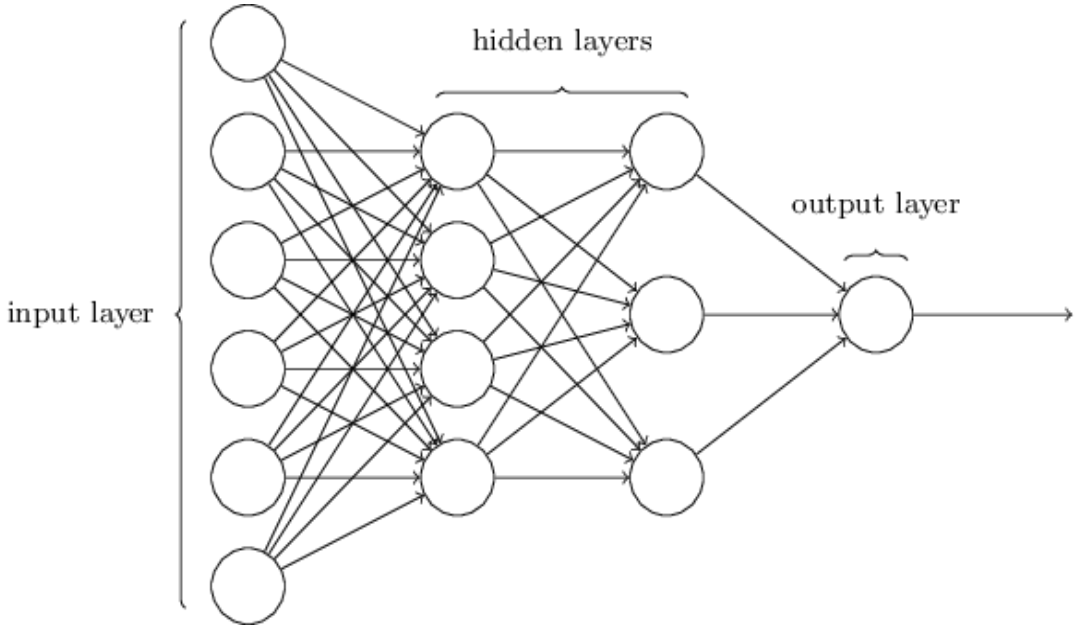
Το Facebook, που θα γίνει εκτεταμένη αναφορά σε επόμενο κεφάλαιο, είχε μεγάλη επιτυχία με την αναγνώριση προσώπων σε φωτογραφίες με τη βοήθεια της βαθιάς μάθησης. Δεν είναι μόνο μια οριακή βελτίωση, αλλά μια ριζική καινοτομία στον τομέα: "Αν ερωτηθεί κανείς αν δύο άγνωστες φωτογραφίες κάποιων ατόμων δείχνουν το ίδιο πρόσωπο, ένας άνθρωπος θα το πετύχει σωστά με ποσοστό 97,53% . Το νέο λογισμικό που αναπτύχθηκε από τους ερευνητές στο Facebook μπορεί να πετύχει 97,25 τοις εκατό για την ίδια πρόκληση, ανεξάρτητα από τις διακυμάνσεις του φωτισμού ή αν το άτομο στην εικόνα βρίσκεται απέναντι στην κάμερα.

Η αναγνώριση ομιλίας είναι ένας άλλος τομέας που έχει αισθανθεί την επίδραση της. Οι ομιλούμενες γλώσσες είναι τόσο τεράστιες και διφορούμενες. Η Baidu για παράδειγμα - μια από τις κορυφαίες μηχανές αναζήτησης της Κίνας - έχει αναπτύξει ένα σύστημα αναγνώρισης φωνής που είναι πιο γρήγορο και ακριβέστερο από τον άνθρωπο για την παραγωγή κειμένου σε κινητό τηλέφωνο, τόσο στα αγγλικά όσο και στα μανδαρινικά.

Η Google χρησιμοποιεί βαθιά μάθηση για τη διαχείριση της ενέργειας στα κέντρα δεδομένων της εταιρείας. Έχουν μειώσει τις ενεργειακές τους ανάγκες για ψύξη κατά 40%. Αυτό μεταφράζεται σε περίπου 15% βελτίωση της αποδοτικότητας της χρήσης ενέργειας για την εταιρεία και εξοικονόμηση εκατοντάδων εκατομμυρίων δολαρίων.

Η βαθιά εκμάθηση είναι σημαντική γιατί τελικά καθιστά αυτές τις εργασίες εφικτές και φέρνει στο προσκήνιο παρελθοντικά προβλήματα τα οποία φαίνονταν άλυτα στο πεδίο του Machine Learning. Ακόμα, η Βαθιά Εκμάθηση είναι ένας τρόπος όπου η Μηχανική μάθηση χρησιμοποιεί έναν συγκεκριμένο αλγόριθμο που ονομάζεται Νευρωνικό Δίκτυο. Για να αποφευχθεί το μπέρδεμα των παραπάνω ορισμών, αρκεί να θυμάται κανείς ότι η Βαθιά Μάθηση είναι απλά ένας τύπος αλγορίθμου που φαίνεται να λειτουργεί πολύ καλά για πρόβλεψη. Τα νευρωνικά δίκτυα εμπνέονται από τη δομή του εγκεφαλικού φλοιού. Στο βασικό επίπεδο είναι το perceptron, η μαθηματική αναπαράσταση ενός βιολογικού νευρώνα. Όπως και στον εγκεφαλικό φλοιό, μπορεί να υπάρχουν πολλά στρώματα διασυνδεδεμένων αναγνωριστών (perceptron's). Οι τιμές εισόδου, ή με άλλα λόγια τα υποκείμενα δεδομένα, περνούν μέσα από αυτό το "δίκτυο" κρυφών επιπέδων μέχρι τελικά να συγκλίνουν στο στρώμα εξόδου. Το στρώμα εξόδου είναι η πρόβλεψη. Μπορεί να είναι ένας κόμβος εάν το μοντέλο εξάγει μόνο έναν αριθμό ή μερικούς κόμβους εάν είναι ένα πρόβλημα πολυκατηγορίας ταξινόμησης.

Τα κρυμμένα στρώματα ενός νευρωνικού δικτύου πραγματοποιούν τροποποιήσεις στα δεδομένα για να βρουν τελικά τι σχέση έχει με τη μεταβλητή στόχου. Κάθε κόμβος έχει βάρος και πολλαπλασιάζει την τιμή εισόδου του με το βάρος αυτό. Για να γίνει αντιληπτό τι πρέπει να είναι αυτά τα μικρά βάρη, συνήθως χρησιμοποιούμε έναν αλγόριθμο που ονομάζεται Backpropagation (Russell, Norvig 1995).



Εικόνα 2. 5 Backpropagation

Η μεγάλη αποκάλυψη για τα νευρωνικά δίκτυα και οι περισσότεροι αλγόριθμοι μάθησης, στην πραγματικότητα είναι ότι δεν είναι όλοι τόσο έξυπνοι. Βασικά, μέσω συνεχόμενων δοκιμών και σφαλμάτων καταφέρνουν να εντοπίζουν τις σχέσεις μεταξύ των δεδομένων.

Στη δεκαετία του 1980, τα περισσότερα νευρωνικά δίκτυα αποτελούσαν ένα μόνο στρώμα λόγω του κόστους υπολογισμού και διαθεσιμότητας δεδομένων. Σήμερα μπορούμε να διαθέσουμε περισσότερα κρυμμένα στρώματα στα Νευρωνικά μας Δίκτυα και έτσι προέκυψε ο όρος "Deep" Learning.

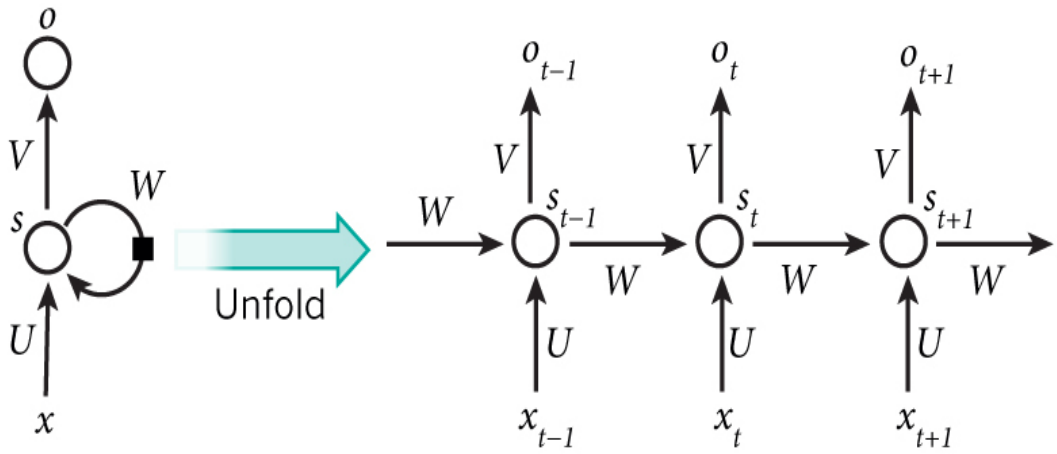
2.8 Αναδρομικά Νευρωνικά Δίκτυα

Σε αυτό το σημείο θα γίνει μια εισαγωγή στα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks ή RNN's) καθώς είναι η αρχιτεκτονική νευρωνικού δικτύου που θα χρησιμοποιηθεί για την λύση του προβλήματος μας.

Τα αναδρομικά νευρωνικά δίκτυα (RNN) είναι δημοφιλή μοντέλα που έχουν δείξει αρκετά υποσχόμενα σε προβλήματα σχετικά με NLP (Natural Language Processing).

Η ιδέα πίσω από τα RNNs είναι να κάνουν χρήση διαδοχικών πληροφοριών. Σε ένα παραδοσιακό νευρωνικό δίκτυο γίνεται η υπόθεση ότι όλες οι είσοδοι και οι έξοδοι είναι ανεξάρτητες η μια απο την άλλη. Αλλά για πολλά προβλήματα αυτό είναι μια πολύ κακή ιδέα. Τα RNN ονομάζονται αναδρομικά επειδή εκτελούν την ίδια εργασία για κάθε στοιχείο μιας

ακολουθίας, με την έξοδο να εξαρτάται από τους προηγούμενους υπολογισμούς. Ένας άλλος τρόπος να σκεφτεί τα RNNs είναι ότι έχουν μια «μνήμη» που συλλαμβάνει πληροφορίες για το τι έχει υπολογιστεί μέχρι τώρα. Θεωρητικά, τα RNNs μπορούν να χρησιμοποιήσουν πληροφορίες σε αυθαίρετα μακρές ακολουθίες, αλλά στην πράξη περιορίζονται στην επιστροφή μόνο μερικών βημάτων. Παρακάτω φαίνεται ένα τυπικό RNN δίκτυο :



Εικόνα 2. 6 Διάγραμμα Αναδρομικών Νευρωνικών Δικτύων

Το παραπάνω διάγραμμα δείχνει ένα RNN που ξεδιπλώνεται σε ένα πλήρες δίκτυο. Για παράδειγμα, εάν η ακολουθία που μας ενδιαφέρει είναι μια πρόταση των 5 λέξεων, το δίκτυο θα ξεδιπλωθεί σε ένα 5-στρώματα νευρωνικό δίκτυο, ήτοι ένα στρώμα για κάθε λέξη. Οι τύποι που διέπουν τον υπολογισμό που συμβαίνει σε ένα RNN είναι οι εξής:

- x_t είναι η είσοδος το χρονικό βήμα t. Για Παράδειγμα, x_1 θα μπορούσε να είναι ένα διάνυσμα το οποίο εκπροσωπεί την δεύτερη λέξη μιας πρότασης.
- s_t είναι η κρυφή κατάσταση στο χρονικό βήμα t. Είναι η "μνήμη" του δικτύου. s_t υπολογίζεται με βάση την προηγούμενη κρυφή κατάσταση και την είσοδο στο τρέχον βήμα: $s_t = f(Ux_t + Ws_{t-1})$. Η συνάρτηση f είναι συνήθως μια μη γραμμικότητα όπως tanh ή ReLU. s_{-1}, που απαιτείται για τον υπολογισμό της πρώτης κρυφής κατάστασης, τυπικά αρχικοποιείται σε όλα τα μηδενικά.
- o_t είναι η έξοδος στο βήμα t.Για παράδειγμα, αν θέλαμε να προβλέψουμε την επόμενη λέξη σε μια πρόταση, θα ήταν ένας φορέας πιθανών λέξεων από όλο το λεξιλόγιό μας. $o_t = \text{softmax}(Vs_t)$.

Τα RNN επέδειξαν μεγάλη επιτυχία σε πολλά προβλήματα NLP. Σε αυτό το σημείο πρέπει να αναφερθεί ότι ο πιο συχνά χρησιμοποιούμενος τύπος RNNs είναι τα LSTMs (Drenth 2017), τα οποία είναι πολύ καλύτερα στην απόκτηση μακροπρόθεσμων εξαρτήσεων από τα πρωταρχικά RNN. Αλλά τα LSTMs είναι ουσιαστικά τα ίδια με τα RNN που θα αναπτυχθούν. Έχουν απλώς έναν διαφορετικό τρόπο υπολογισμού του κρυμμένου κόστους. Κατανόηση της Οπισθοδρόμησης μέσα στον χρόνο. Παρά την μεγάλη επιτυχία της μεθόδου της οποσθοδιάδοσης, εν τούτοις υπάρχουν και περιπτώσεις που η μέθοδος αποτυγχάνει, ή δεν δουλεύει άμεσα με επιτυχία. Σε τέτοιες περιπτώσεις συνήθως χρειάζεται να γίνονται αλλαγές στις τιμές παραμέτρων, αρχικές συνθήκες, κ.λ.π. μέχρις ότου διορθωθεί το πρόβλημα. Μερικές φορές ο χρόνος εκπαίδευσης είναι υπερβολικά μεγάλος. Χρειάζονται π.χ. πολλά εκατομμύρια κύκλοι διόρθωσης μέχρις ότου το σύστημα συγκλίνει, ή και να υπάρξει το ενδεχόμενο να μην συγκλίνει ποτέ. Σε τέτοιες περιπτώσεις πρέπει να αλλάζει το μέγεθος του βήματος.

Αυτό συμβαίνει διότι τα βάρη μπορεί να πάρουν μεγάλες τιμές. Εφόσον το σφάλμα που επιστρέφει από την έξοδο προς το κρυμμένο επίπεδο μέσα στο δίκτυο είναι ανάλογο της παραγώγου αυτής, μπορεί τότε η διαδικασία εκπαίδευσης να "κωλύσει". Τότε για να ξεπεραστεί αυτό γίνεται σμίκρυνση του μεγέθους του βήματος, αλλά αυτό έχει ως συνέπεια να μεγαλώσει ο χρόνος εκπαίδευσης. Ένα άλλο συχνό πρόβλημα είναι αυτό των τοπικών ελαχίστων. Η μέθοδος αυτή, χρησιμοποιεί την μαθηματική τεχνική της επικλινούς καθόδου.

Ακολουθείται η κλίση της επιφάνειας σφάλματος προς τα κάτω, μεταβάλλοντας συνεχώς τα βάρη μέχρι το σύστημα να φθάσει στο ελάχιστο. Το ελάχιστο αυτό όμως πρέπει να είναι το παγκόσμιο ελάχιστο. Αρκεί να αναλογιστεί κανείς ότι η επιφάνεια μπορεί να έχει πολλά βουνά, λόφους, κοιλάδες, φαράγγια, χαράδρες, κ.λ.π. Κατεπέκταση αυτό σημαίνει ότι υπάρχουν πολλά τοπικά ελάχιστα, που είναι ψηλότερα από το παγκόσμιο ελάχιστο και στα οποία μπορεί εύκολα να παγιδευτεί το δίκτυο στην προσπάθειά του να βρει το ελάχιστο. Καθ' όσον το σύστημα θέλει να πάει πάντα προς τα κάτω, αν πέσει σε ένα τοπικό ελάχιστο δεν έχει τρόπο να ξεφύγει, και να συνεχίσει το δρόμο του. Συνήθως χρησιμοποιούνται στατιστικές μέθοδοι εκπαίδευσης, για να αποφεύγεται το πρόβλημα αυτό.

Το μέγεθος του βήματος επίσης παίζει σημαντικό ρόλο στην ταχύτητα εκμάθησης. Εάν είναι πολύ μικρό, τότε η εκπαίδευση αργεί υπερβολικά, και πρέπει να αυξηθεί. Οι αλλαγές των βαρών θα πρέπει να γίνονται ταυτόχρονα σε όλα τα πρότυπα. Αν όμως το δίκτυο βρίσκεται σε ένα περιβάλλον το οποίο συνεχώς αλλάζει πρότυπα, τότε η εκπαίδευση του δικτύου δεν θα συγκλίνει ποτέ, λεχοντας ως αποτέλεσμα το δίκτυο να εκπαιδεύεται άσκοπα. Η μέθοδος της

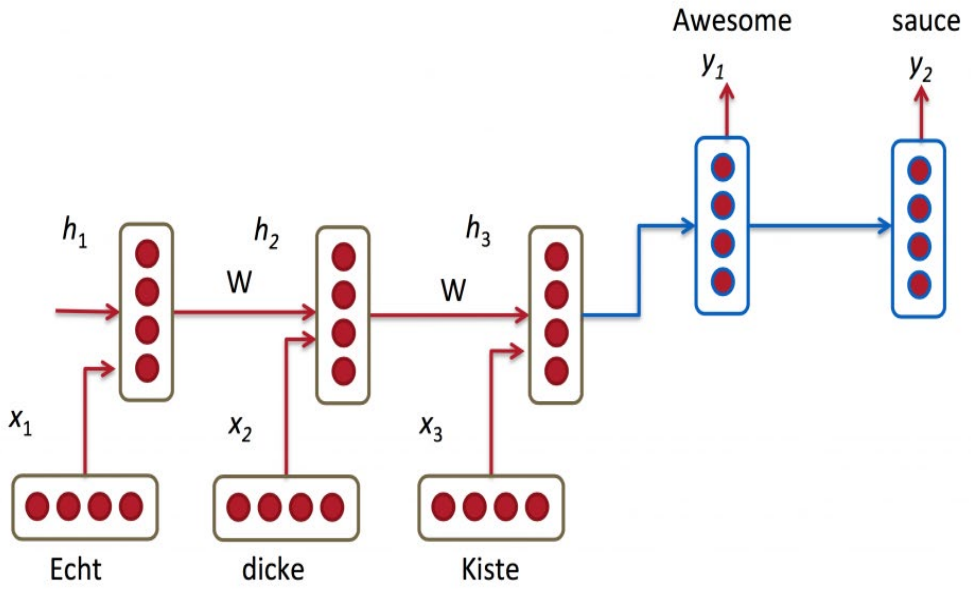
οπισθοδιάδοσης είναι η πιο κοινή και ευρέως χρησιμοποιούμενη μέθοδος σήμερα για εκπαίδευση νευρωσικών δικτύων.

2.8.1 Μοντελοποίηση γλωσσών και δημιουργία κειμένου

Λαμβάνοντας μια ακολουθία λέξεων, θέλουμε να προβλέψουμε την πιθανότητα κάθε λέξης με δεδομένες τις προηγούμενες λέξεις. Τα Μοντέλα Γλωσσών μας επιτρέπουν να μετρήσουμε πόσο πιθανό είναι μια πρόταση, η οποία είναι μια σημαντική συνεισφορά στην Μηχανική Μετάφραση (επειδή οι προτάσεις υψηλής πιθανότητας είναι συνήθως σωστές). Μια παρενέργεια της δυνατότητας να προβλέψουμε την επόμενη λέξη είναι ότι έχουμε ένα γενετικό μοντέλο, το οποίο μας επιτρέπει να δημιουργούμε νέο κείμενο με δειγματοληψία από τις πιθανότητες εξόδου. Και ανάλογα με τα δεδομένα εκπαίδευσης που διαθέτουμε, μπορούμε να δημιουργήσουμε όλα τα είδη. Στη Γλωσσική Μοντελοποίηση, η είσοδός μας είναι συνήθως μια ακολουθία λέξεων κωδικοποιημένων ως ένα διάνυσμα, και η έξοδος μας είναι η ακολουθία των προβλεπόμενων λέξεων. Κατά την εκπαίδευση του δικτύου ορίσαμε $o_t = x_{t+1}$ αφού θέλουμε η έξοδος στο βήμα t να είναι η πραγματική επόμενη λέξη.

2.8.2 Μηχανική μετάφραση

Η μηχανική μετάφραση είναι παρόμοια με τη μοντελοποίηση της γλώσσας, καθώς η εισαγωγή μας είναι μια ακολουθία λέξεων στη γλώσσα προέλευσής μας (π.χ. Γερμανικά). Θέλουμε να εξάγουμε μια ακολουθία λέξεων στη γλώσσα στόχου (π.χ. Αγγλικά). Μια βασική διαφορά είναι ότι η παραγωγή μας αρχίζει μόνο αφού έχουμε δει την πλήρη είσοδο, επειδή η πρώτη λέξη των μεταφρασμένων προτάσεών μας μπορεί να απαιτεί πληροφορίες που συλλέγονται από την πλήρη ακολουθία εισόδου.



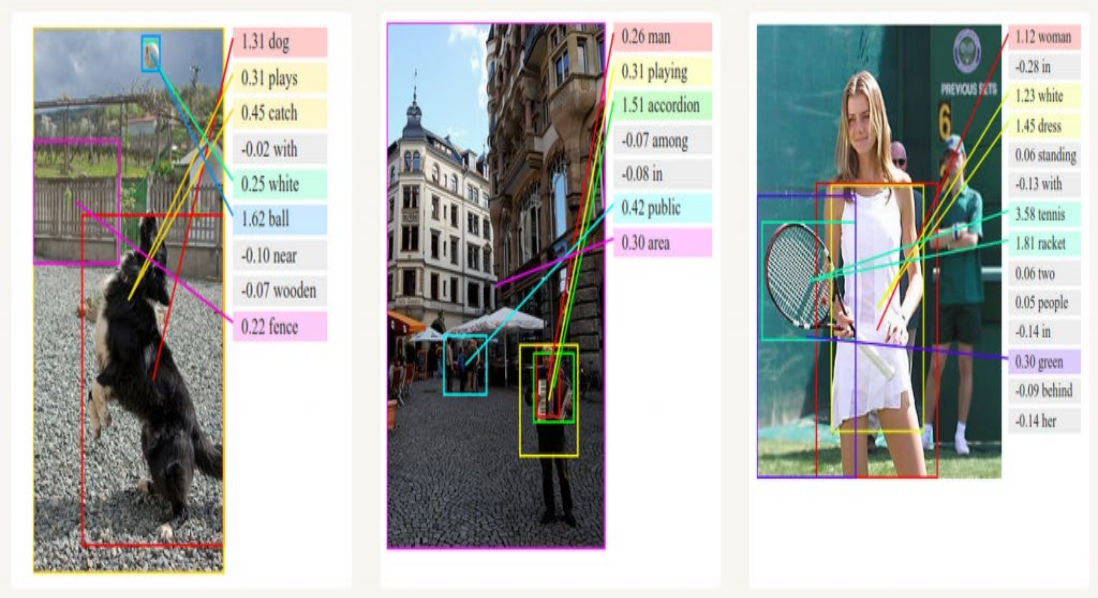
Εικόνα 2. 7 Σύστημα Μηχανικής Μετάφρασης

2.8.3 Αναγνώριση ομιλίας

Δεδομένης της ακολουθίας εισόδου ακουστικών σημάτων από ένα ηχητικό κύμα, μπορούμε να προβλέψουμε μια ακολουθία φωνητικών τμημάτων μαζί με τις πιθανότητες τους.

2.8.4 Δημιουργία περιγραφών εικόνας

Μαζί με τα συνελικτικά νευρικά δίκτυα, τα RNN χρησιμοποιήθηκαν ως μέρος ενός μοντέλου για τη δημιουργία περιγραφών για εικόνες χωρίς ετικέτα. Είναι εκπληκτικό το πόσο καλά φαίνεται αυτό να λειτουργεί. Το συνδυασμένο μοντέλο ευθυγραμμίζει ακόμη και τις δημιουργούμενες λέξεις με χαρακτηριστικά που βρίσκονται στις εικόνες.



Εικόνα 2. 8 Συνδυαστικό Μοντέλο CNN με RNN

2.9 Εκπαίδευση RNNs

Η εκπαίδευση ενός RNN είναι παρόμοια με την εκπαίδευση ενός παραδοσιακού νευρωνικού δικτύου. Χρησιμοποιούμε επίσης τον αλγόριθμο οποσθοδιάδοσης. Επειδή οι παράμετροι μοιράζονται σε όλα τα βήματα χρόνου στο δίκτυο, η κλίση σε κάθε έξοδο εξαρτάται όχι μόνο από τους υπολογισμούς του τρέχοντος βήματος χρόνου, αλλά και από τα προηγούμενα βήματα χρόνου. Για παράδειγμα, για να υπολογιστεί η κλίση στο t = 4 θα χρειαζόταν να κάνουμε back propagate 3 βήματα και να συνοψίσουμε τις κλίσεις. Αυτό ονομάζεται χρονική οποσθοδιάδοση (BPTT). Είναι γεγονός ότι τα πρωταρχικά RNN που εκπαιδεύονται με BPTT δυσκολεύονται να μάθουν μακροπρόθεσμες εξαρτήσεις εξαιτίας αυτού που ονομάζεται πρόβλημα εξαφανιζόμενων κλίσεων (vanishing gradients problem). Υπάρχουν κάποιοι μηχανισμοί για την αντιμετώπιση αυτών των προβλημάτων, όπως τα LSTM που σχεδιάστηκαν ειδικά για να λύσουν προβλήματα τέτοιου τύπου.

Τα δίκτυα LSTM είναι αρκετά δημοφιλή αυτές τις μέρες καθώς και έγινε σύντομη αναφορά παραπάνω. Εσωτερικά αυτά τα κελιά αποφασίζουν τι πρέπει να κρατήσουν (και τι να διαγράψουν) από τη μνήμη. Στη συνέχεια, συνδυάζουν την προηγούμενη κατάσταση, την τρέχουσα μνήμη και την είσοδο. Αποδεικνύεται ότι αυτοί οι τύποι μονάδων είναι πολύ αποτελεσματικοί στη λήψη μακροπρόθεσμων εξαρτήσεων.

Στην περίπτωση της εργασίας αυτής θα χρησιμοποιηθεί μια παρόμοια αρχιτεκτονική νευρωνικού δικτύου με χρήση των LSTM κελιών, ώστε να γίνει αποφυγή κάποιων προβλήματων όπως το πρόβλημα κατάβασης λόφου.

2.10 Είδη Αρχιτεκτονικών

2.10.1 Εισαγωγή στην Αρχιτεκτονική Word2Vec

Παρακάτω θα γίνει μια πιο αναλυτική προσέγγιση για την αρχιτεκτονική τύπου Word2vec (Date 2000) Η ενσωμάτωση λέξεων (word2vec), όπως η ενσωμάτωση εγγράφων, ανήκει στη φάση προ επεξεργασίας κειμένου. Συγκεκριμένα, στο τμήμα που μετατρέπει ένα κείμενο σε μια σειρά αριθμών. Στην επέκταση επεξεργασίας κειμένου KNIME (Berthold,Wiswedel 2009). ο κόμβος διάνυσμα εγγράφου μετατρέπει μια ακολουθία λέξεων σε μια ακολουθία 0/1 - ή αριθμών συχνότητας - με βάση την παρουσία / απουσία μιας συγκεκριμένης λέξης στο αρχικό κείμενο. Αυτό ονομάζεται επίσης "one hot encoding". Ωστόσο, το one hot encoding έχει δύο μεγάλα προβλήματα:

1) παράγει ένα πολύ μεγάλο πίνακα δεδομένων με τη δυνατότητα ενός μεγάλου αριθμού στηλών.
2) παράγει ένα πολύ αραιό πίνακα δεδομένων με πολύ μεγάλο αριθμό μηδενικών, που μπορεί να δημιουργεί πρόβλημα για την κατάρτιση ορισμένων αλγορίθμων μηχανικής μάθησης.

Η τεχνική Word2Vec συνεπώς σχεδιάστηκε με δύο στόχους: 1) Μείωση του μεγέθους του χώρου κωδικοποίησης λέξεων (χώρος ενσωμάτωσης). Συμπιέζουμε στην αναπαράσταση λέξεων την πιο ενημερωτική περιγραφή για κάθε λέξη. 2) Η ερμηνεία του χώρου ενσωμάτωσης καθίσταται δευτερεύουσα. Η τεχνική Word2Vec βασίζεται σε μια πλήρως συνδεδεμένη αρχιτεκτονική .

Ας ληφθεί σαν παράδειγμα μια απλή πρόταση όπως "η γρήγορη καστανή αλεπού πήδηξε πάνω από το τεμπέλικο σκυλί" και ας εξεταστεί η πρόταση λέξη προς λέξη. Για παράδειγμα, η λέξη "αλεπού" περιβάλλεται από μια σειρά άλλων λέξεων. Αυτό είναι το πλαίσιο της. Αν χρησιμοποιηθεί ένα πρόσθιο πλαίσιο μεγέθους 3, τότε η λέξη "αλεπού" εξαρτάται από το πλαίσιο "η γρήγορη καφέ". Η λέξη "πήδηξε" στο πλαίσιο "γρήγορη καφέ αλεπού" και ούτω καθεξής. Εάν χρησιμοποιηθεί ένα πίσω πλαίσιο μεγέθους 3, η λέξη "αλεπού" εξαρτάται από το πλαίσιο "πήδηξε πάνω". Η λέξη "πήδηξε" στο πλαίσιο "πάνω από το τεμπέλικο" και ούτω καθεξής. Ο συνηθέστερα χρησιμοποιούμενος τύπος περιβάλλοντος είναι ένα πλαίσιο προς τα εμπρός. Δεδομένου ενός πλαισίου και μιας λέξης σχετικής με αυτό το πλαίσιο, αντιμετωπίζονται δύο πιθανά προβλήματα:

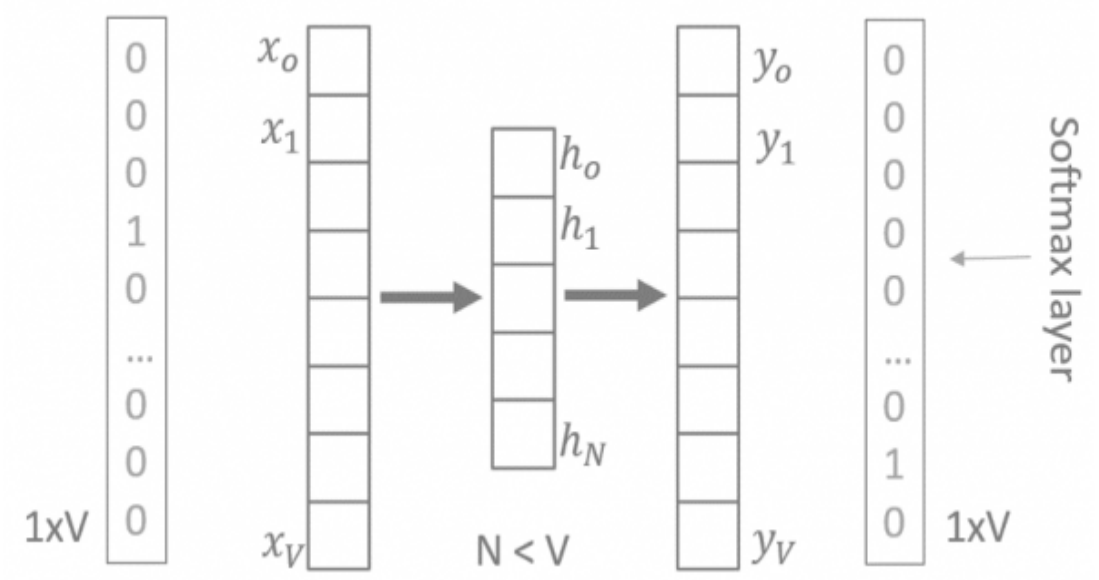
1) Από αυτό το πλαίσιο, να προβλέψουμε τη λέξη στόχου (συνεχής σακούλα λέξεων ή προσέγγιση(CBOW)

2) Από τη λέξη στόχο, να προβλέψουμε το πλαίσιο από το οποίο προήλθε (προσέγγιση Skip-gram)

Εάν χρησιμοποιούμε ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο, με ένα κρυμμένο στρώμα, καταλήγουμε σε μια αρχιτεκτονική όπως αυτή του Σχήματος 1 και για τις δύο προσεγγίσεις. Τα πρότυπα εισόδου και εξόδου είναι κωδικοποιημένοι φορείς, με διάσταση 1xV όπου V είναι το μέγεθος λεξιλογίου. Στην περίπτωση της στρατηγικής CBOW, η λέξη-κλειδί μιας one hot encoded λέξης τροφοδοτεί την είσοδο και η λέξη-στόχο με μια one hot encoded προβλέπεται στο επίπεδο εξόδου. Στην περίπτωση της στρατηγικής Skip-gram, η κωδικοποιημένη λέξη τροφοδοτεί την είσοδο, ενώ το στρώμα εξόδου προσπαθεί να αναπαράγει το κωδικοποιημένο κείμενο μιας λέξης. Ο αριθμός των κρυμμένων νευρώνων είναι N, με N <V. Προκειμένου να εξασφαλιστεί η εκπροσώπηση της λέξης εξόδου με βάση την πιθανότητα, χρησιμοποιείται στη συνάρτηση ενεργοποίησης softmax στη στρώση εξόδου και υιοθετείται η ακόλουθη συνάρτηση σφάλματος E κατά τη διάρκεια της εκπαίδευσης:

$$E = \log\left(p(w_{0}|w_{i})\right)$$

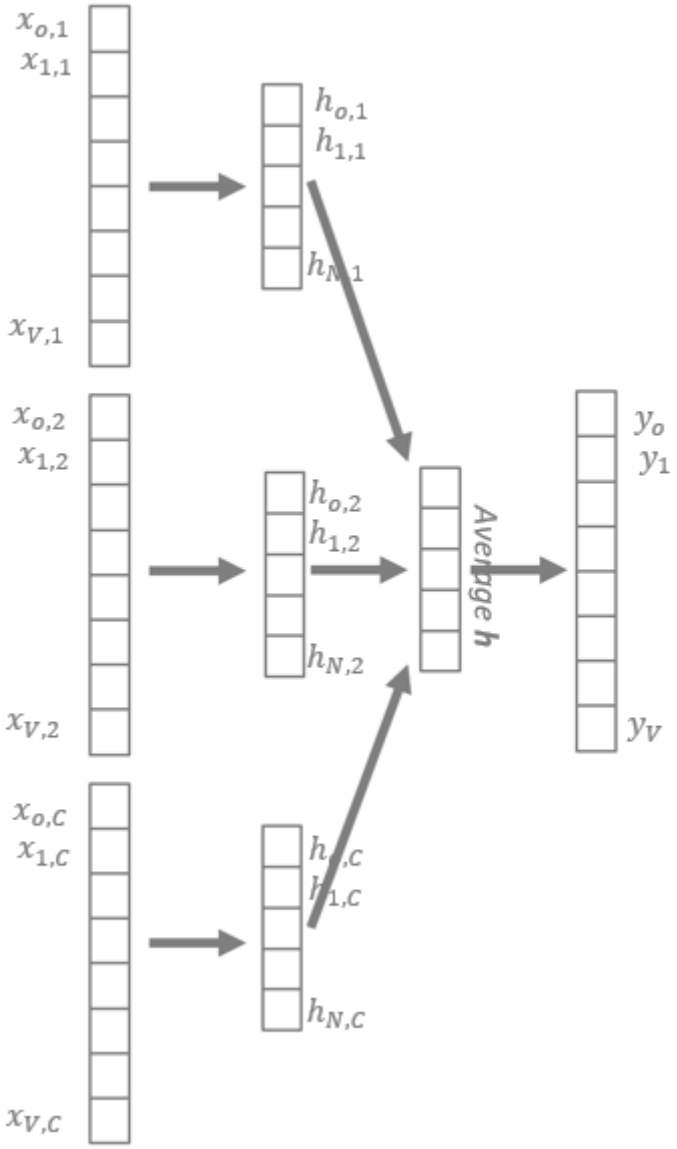
όπου w_{0} είναι η λέξη εξόδου και w_{i} είναι η λέξη εισόδου. Παράλληλα, για να μειωθεί η υπολογιστική προσπάθεια, χρησιμοποιείται μια γραμμική λειτουργία ενεργοποίησης για τους κρυμμένους νευρώνες και τα ίδια βάρη που χρησιμοποιούνται για την ενσωμάτωση όλων των εισόδων (CBOW) ή όλων των εξόδων (Skip-gram).



Εικόνα 2. 9 Word Embedding

Στην Εικόνα 2.9 αξίζει να παρατηρηθεί ότι τα στρώματα εισόδου και εξόδου έχουν αμφότερες τις διαστάσεις 1xV, όπου V είναι το μέγεθος του λεξιλογίου, αφού και οι δύο αντιπροσωπεύουν την κωδικοποίηση μιας λέξης. Παρατηρείται επίσης ότι το κρυφό στρώμα έχει λιγότερες μονάδες (N) από το στρώμα εισόδου (V). Έτσι, εάν αντιπροσώπευαν τη λέξη εισόδου με τις κρυφές εξόδους του νευρώνα και όχι με την αρχική κωδικοποίηση, έχει ήδη μειωθεί το μέγεθος του φορέα λέξεων, ελπίζοντας πάντα ότι διατηρήθηκαν αρκετές από τις αρχικές πληροφορίες. Οι κρυφές έξοδοι του νευρώνα παρέχουν την ενσωμάτωση λέξεων. Αυτή η αναπαράσταση λέξεων, που είναι πολύ πιο συμπαγής από one hot encoding, παράγει μια πολύ λιγότερο αραιή αναπαράσταση του χώρου εγγράφων. Εάν γίνει μετάβαση σε ένα μεγαλύτερο πλαίσιο, για παράδειγμα με μέγεθος C = 3, πρέπει να αλλάξει ελαφρώς τη δομή του δικτύου. Για την προσέγγιση CBOW, χρειάζονται στρώματα εισόδου C μεγέθους V για να συλλέγουμε C λέξεις-κλειδιά με ένα κωδικό. Το αντίστοιχο κρυμμένο στρώμα παρέχει τότε ενσωματωμένες λέξεις C, καθένα από το μέγεθος N.

Συνοψίζοντας τις ενσωματώσεις C, προστίθεται ένα ενδιάμεσο στρώμα για τον υπολογισμό της μέσης τιμής των ενσωματώσεων C . Το στρώμα εξόδου προσπαθεί να παράγει μία κωδικοποιημένη αναπαράσταση της λέξης προορισμού, με τις ίδιες λειτουργίες ενεργοποίησης και την ίδια λειτουργία σφάλματος όπως και για την αρχιτεκτονική δικτύου V-N-V στο σχήμα 1.



Εικόνα 2. 10 Word Embedding, Skip-gram

Μια παρόμοια αρχιτεκτονική (Εικόνα 10) δημιουργείται για την προσέγγιση Skip-gram και το μέγεθος περιβάλλοντος C> 1. Ακόμα και εδώ η λέξη dimensionality representation αντιπαραβάλλεται από V σε N και ο χώρος του εγγράφου παίρνει μια πιο συμπαγή αναπαράσταση. Αυτό θα βοηθήσει με τους επερχόμενους αλγόριθμους μηχανικής μάθησης, ειδικά σε μεγάλα σύνολα δεδομένων και μεγάλα λεξιλόγια.

Το σύνολο λεξιλογίου μπορεί να γίνει πολύ μεγάλο, με πολλές λέξεις να χρησιμοποιούνται μόνο μερικές φορές. Συνήθως, μια λειτουργία εξετάζει όλες τις λέξεις στο

λεξιλόγιο και αποφασίζει ποιοι θα κρατήσουν. Η λειτουργία επιβίωσης λαμβάνει την ακόλουθη μορφή:

$$P(w) = \frac{\left(\sqrt{\frac{z(w)}{s}} + 1\right)s}{z(w)}$$

όπου w είναι η λέξη, z (w) η συχνότητά της στο σύνολο εκπαίδευσης, και s είναι μια παράμετρος που ονομάζεται ρυθμός δειγματοληψίας. Όσο μικρότερος είναι ο ρυθμός δειγματοληψίας, τόσο λιγότερο πιθανή είναι η διατήρηση μιας λέξης.

Μερικές φορές, εκτός από τα κλασικά θετικά παραδείγματα στο σετ εκπαίδευσης, παρέχονται αρνητικά παραδείγματα. Τα αρνητικά παραδείγματα είναι λανθασμένες εισροές για τις οποίες δεν μπορεί να προσδιοριστεί έξοδος. Πρακτικά, τα αρνητικά παραδείγματα θα πρέπει να παράγουν έναν κωδικό όλων των 0. Γενικά, η προσέγγιση CBOW είναι ήδη μια βελτίωση σε σχέση με τον υπολογισμό μιας μήτρας συν-εμφάνισης, καθώς απαιτεί λιγότερη μνήμη. Ωστόσο, οι χρόνοι εκπαίδευσης για την προσέγγιση CBOW μπορεί να είναι πολύ μεγάλοι. Η προσέγγιση Skip-gram εκμεταλλεύεται την αρνητική δειγματοληψία και επιτρέπει τη σημασιολογική διαφοροποίηση μιας λέξης.

Όλη η διαίσθηση πίσω από την προσέγγιση Word2Vec συνίσταται στην εκπροσώπηση μιας λέξης με βάση το περιβάλλον της. Αυτό σημαίνει ότι οι λέξεις που εμφανίζονται σε παρόμοια περιβάλλοντα θα ενσωματωθούν παρομοίως.

2.11 Προγραμματιστικές Βιβλιοθήκες για Βαθιά Μάθηση

Πολλές από τις εξελίξεις στις πρακτικές εφαρμογές της Deep Learning οδήγησαν στη διαδεδομένη διαθεσιμότητα ισχυρών πακέτων λογισμικού ανοιχτού κώδικα. Αυτά επιτρέπουν στους προγραμματιστές να εξοικειώνονται εύκολα και αποτελεσματικά, γεγονός που επεκτείνει τον αριθμό των ανθρώπων που πιέζουν ενεργά την ανάπτυξη προς τα εμπρός. Δεδομένου ότι το Data Science γενικά έχει κινηθεί περισσότερο προς την Python, τα περισσότερα από αυτά τα πακέτα είναι ανεπτυγμένα για αυτή τη γλώσσα.

2.11.1 TensorFlow

TensorFlow είναι μια βιβλιοθήκη λογισμικού ανοικτού κώδικα για αριθμητικούς υπολογισμούς χρησιμοποιώντας γραφήματα ροής δεδομένων. Οι κόμβοι στο γράφημα αντιπροσωπεύουν μαθηματικές λειτουργίες, ενώ οι άκρες των γραφημάτων αντιπροσωπεύουν τις πολυδιάστατες συστοιχίες δεδομένων (tensors) που επικοινωνούν μεταξύ τους. Η ευέλικτη αρχιτεκτονική επιτρέπει να αναπτύσσονται υπολογισμοί σε μία ή περισσότερες CPU ή GPU

σε επιτραπέζιο υπολογιστή, διακομιστή ή κινητή συσκευή με ένα ενιαίο API. Για την λύση του προβλήματος που διαπραγματεύεται η συγκεκριμένη εργασία θα χρησιμοποιηθεί η συγκεκριμένη προγραμματιστική βιβλιοθήκη.

2.11.2 Caffe

Το Caffe2 έχει ως στόχο να προσφέρει έναν εύκολο και απλό τρόπο για να πειραματιστεί κανείς με βαθιά μάθηση και να επωφεληθεί από την κοινοτική συνεισφορά νέων μοντέλων και αλγορίθμων.

2.11.3 Torch

Το Torch είναι μια επιστημονική βιβλιοθήκη με ευρεία υποστήριξη για τους αλγόριθμους μηχανικής μάθησης που θέτουν πρώτα GPU σε λειτουργία. Είναι εύκολο στη χρήση και αποτελεσματικό, χάρη σε μια απλή και γρήγορη γλώσσα γραφής, LuaJIT, και μια υποκείμενη εφαρμογή C / CUDA.

2.11.4 Theano

Το Theano είναι μια βιβλιοθήκη Python που επιτρέπει να οριστούν, να βελτιστοποιηθούν και να αξιολογηθούν μαθηματικές εκφράσεις, ειδικά αυτές με πολυδιάστατες συστοιχίες. Χρησιμοποιώντας το Theano είναι δυνατόν να επιτευχθούν ταχύτητες ανταγωνιζόμενες με τις χειροποίητες υλοποιήσεις C για προβλήματα μεγάλης ποσότητας δεδομένων. Μπορεί επίσης να ξεπεράσει τη C σε μια CPU από πολλές τάξεις μεγέθους, αξιοποιώντας τις πρόσφατες GPU.

2.11.5 ConvNetJS

Το ConvNetJS είναι μια βιβλιοθήκη Javascript για την εκπαίδευση των μοντέλων Deep Learning (Νευρωνικά δίκτυα) εξ ολοκλήρου στο πρόγραμμα περιήγησης.

2.12 Δεδομένα Μεγάλης Κλίμακας

Μεγάλα δεδομένα είναι ένας όρος που ο καθένας πλέον μοιάζει να μιλάει για αυτά. Τι πραγματικά είναι τα μεγάλα δεδομένα; Από πού προέρχονται αυτά τα δεδομένα, πώς γίνεται η επεξεργασία τους και πώς χρησιμοποιούνται τα αποτελέσματα; Σε αυτό το σύντομο κεφάλαιο, θα γίνει μια εισαγωγή στα μεγάλα δεδομένα και τι σημαίνουν για τον μεταβαλλόμενο κόσμο στον οποίο ζούμε.

2.12.1 Τι είναι τα Δεδομένα Μεγάλης Κλίμακας

Δεν υπάρχει κανένας γρήγορος κανόνας σχετικά με το ακριβές μέγεθος μιας βάσης δεδομένων, ώστε τα δεδομένα εντός αυτής να θεωρούνται "μεγάλα". Αντίθετα, αυτό που συνήθως ορίζεται ως μεγάλα δεδομένα είναι η ανάγκη για νέες τεχνικές και εργαλεία για να μπορέσει να τα επεξεργαστούν. Για να χρησιμοποιηθούν και επεξεργαστούν μεγάλα δεδομένα, χρειάζονται προγράμματα που καλύπτουν πολλαπλές φυσικές και / ή εικονικές μηχανές που συνεργάζονται ώστε να επεξεργαστούν όλα τα δεδομένα σε εύλογο χρονικό διάστημα. Το να αποκτηθούν προγράμματα σε πολλαπλές μηχανές και να συνεργαστούν με αποτελεσματικό τρόπο έτσι ώστε κάθε πρόγραμμα να γνωρίζει ποια στοιχεία των δεδομένων θα επεξεργαστούν και στη συνέχεια να είναι σε θέση να βάλει τα αποτελέσματα από όλες τις μηχανές μαζί για να κατανοήσει ένα μεγάλο σύνολο δεδομένων, χρειάζεται ιδιαίτερες τεχνικές προγραμματισμού. Δεδομένου ότι συνήθως είναι πολύ ταχύτερο για τα προγράμματα να έχουν πρόσβαση στα δεδομένα που είναι αποθηκευμένα τοπικά αντί για ένα δίκτυο, η διανομή δεδομένων σε ένα σύμπλεγμα και ο τρόπος με τον οποίο τα μηχανήματα αυτά είναι συνδεδεμένα σε δίκτυο είναι επίσης σημαντικοί παράγοντες όταν σκεφτόμαστε μεγάλα προβλήματα δεδομένων.

Οι χρήσεις των μεγάλων δεδομένων είναι σχεδόν τόσο ποικίλες όσο είναι "μεγάλες". Σημαντικά παραδείγματα στα οποία ίσως υπάρχει μια ήδη εξοικείωση περιλαμβάνουν: Κοινωνικά δίκτυα μέσων ενημέρωσης που αναλύουν τα δεδομένα των μελών τους για να μάθουν περισσότερα σχετικά με αυτά και να τα συνδέσουν με περιεχόμενο και διαφήμιση σχετικό με τα ενδιαφέροντά τους ή μηχανές αναζήτησης που εξετάζουν τη σχέση μεταξύ ερωτημάτων και αποτελεσμάτων, απαντήσεις στις ερωτήσεις των χρηστών. Δύο από τις μεγαλύτερες πηγές δεδομένων σε μεγάλες ποσότητες είναι τα δεδομένα συναλλαγών, συμπεριλαμβανομένων όλων των στοιχείων από τις τιμές των μετοχών στα τραπεζικά δεδομένα, στις ιστορίες αγοράς μεμονωμένων εμπόρων και δεδομένων αισθητήρων, πολλά από τα οποία προέρχονται από αυτό που συνήθως αναφέρεται ως Διαδίκτυο των πραγμάτων (IoT). Αυτά τα δεδομένα αισθητήρων μπορεί να είναι οτιδήποτε, από μετρήσεις που λαμβάνονται από τα ρομπότ σε μια γραμμή κατασκευής του αυτοκινητοβιομήχανου εξοπλισμού, σε δεδομένα θέσης σε δίκτυο κινητής τηλεφωνίας, σε δεδομένα στιγμιαίας χρήσης ηλεκτρικού ρεύματος σε σπίτια και επιχειρήσεις, σε πληροφορίες επιβατών που επιβιβάζονται σε ένα σύστημα διαμετακόμισης. Με την ανάλυση αυτών των δεδομένων, οι οργανισμοί μπορούν να μάθουν τάσεις σχετικά με τα δεδομένα που μετρούν, καθώς και τους ανθρώπους που παράγουν αυτά τα δεδομένα. Η ελπίδα αυτής της μεγάλης ανάλυσης

δεδομένων είναι να παρέχει πιο προσαρμοσμένες υπηρεσίες και αυξημένη αποτελεσματικότητα σε οποιαδήποτε βιομηχανία συλλέγει τα δεδομένα.

Μία από τις πιο γνωστές μεθόδους για τη μετατροπή των πρώτων δεδομένων σε χρήσιμες πληροφορίες είναι αυτό που είναι γνωστό ως MapReduce (Lämmel 2008) . Το MapReduce είναι μια μέθοδος για τη λήψη ενός μεγάλου συνόλου δεδομένων και την εκτέλεση υπολογισμών σε αυτό σε πολλούς υπολογιστές, παράλληλα. Χρησιμεύει ως πρότυπο για τον τρόπο προγραμματισμού και χρησιμοποιείται συχνά για να αναφερθεί στην πραγματική εφαρμογή αυτού του μοντέλου.

Στην ουσία, το MapReduce αποτελείται από δύο μέρη. Η λειτουργία χαρτών κάνει την ταξινόμηση και φιλτράρισμα, λαμβάνοντας δεδομένα και τοποθετώντας τα μέσα σε κατηγορίες ώστε να μπορούν να αναλυθούν. Η λειτουργία της μείωσης (reduce) παρέχει μια περίληψη αυτών των δεδομένων συνδυάζοντας τα όλα μαζί. Ενώ σε μεγάλο βαθμό πιστώνεται στην έρευνα που πραγματοποιήθηκε στη Google, το MapReduce είναι τώρα ένας γενικός όρος και αναφέρεται σε ένα γενικό μοντέλο που χρησιμοποιείται από πολλές τεχνολογίες.

2.12.2 Εργαλεία για την Διαχείριση "Μεγάλων Δεδομένων"

Ίσως το πιο ισχυρό εργαλείο για την ανάλυση μεγάλων δεδομένων είναι γνωστό ως Apache Hadoop. Το Apache Hadoop (Evans 2013) είναι ένα πλαίσιο για την αποθήκευση και επεξεργασία δεδομένων σε μεγάλη κλίμακα και είναι απόλυτα ανοιχτό. Το Hadoop μπορεί να τρέξει σε υλικό βασικών προϊόντων, καθιστώντας εύκολη τη χρήση με ένα υπάρχον κέντρο δεδομένων, ή ακόμα και να διεξάγει ανάλυση στο σύννεφο (cloud). Το Hadoop χωρίζεται σε τέσσερα βασικά μέρη: Το σύστημα κατανομής αρχείων Hadoop, το οποίο είναι ένα κατανεμημένο σύστημα αρχείων σχεδιασμένο για πολύ υψηλό συνολικό εύρος ζώνης. YARN, μια πλατφόρμα για τη διαχείριση των πόρων του Hadoop και προγράμματα προγραμματισμού που θα λειτουργούν στην υποδομή Hadoop.

Το MapReduce, όπως περιεγράφηκε παραπάνω, αποτελεί πρότυπο για τη διεξαγωγή μεγάλης επεξεργασίας δεδομένων. Για περισσότερες πληροφορίες σχετικά με το Hadoop, μπορεί να ανατρέξει κανείς στην εισαγωγή του Apache Hadoop για μεγάλα δεδομένα. Φυσικά, αυτά δεν είναι τα μόνα μεγάλα εργαλεία δεδομένων. Υπάρχουν πραγματικά πολλές λύσεις ανοιχτού κώδικα για εργασία με μεγάλα δεδομένα, πολλά από τα οποία είναι εξειδικευμένα για την παροχή βέλτιστων δυνατοτήτων και επιδόσεων για μια συγκεκριμένη διαμόρφωση ή για συγκεκριμένες διαμορφώσεις υλικού.

- Το Apache Software Foundation (ASF) υποστηρίζει πολλά από αυτά τα μεγάλα έργα δεδομένων. Εδώ είναι μερικά που μπορείτε να βρείτε χρήσιμα.
- Το Apache Beam είναι ένα ενοποιημένο μοντέλο για τον προσδιορισμό αγωγών παράλληλης επεξεργασίας παρτίδων και συνεχούς ροής δεδομένων. " Επιτρέπει στους προγραμματιστές να γράψουν κώδικα που λειτουργεί σε πολλαπλές μηχανές επεξεργασίας.
- Η Apache Hive είναι μια αποθήκη δεδομένων που χτίστηκε στο Hadoop. Ένα έργο Apache κορυφαίου επιπέδου,
- Το Apache Impala είναι μια μηχανή ερώτησης SQL που τρέχει στον Hadoop.
- Το Apache Kafka επιτρέπει στους χρήστες να δημοσιεύουν και να εγγράφονται σε τροφοδοσίες δεδομένων σε πραγματικό χρόνο. Στόχος της είναι να φέρει την αξιοπιστία άλλων συστημάτων ανταλλαγής μηνυμάτων σε δεδομένα συνεχούς ροής.
- Το Apache Lucene είναι βιβλιοθήκη λογισμικού ευρετηρίασης και αναζήτησης πλήρους κειμένου που μπορεί να χρησιμοποιηθεί για μηχανές συστάσεων. Είναι επίσης η βάση για πολλά άλλα ερευνητικά έργα, όπως η Solr και η Elasticsearch.
- Το Apache Pig είναι μια πλατφόρμα για την ανάλυση μεγάλων συνόλων δεδομένων που τρέχουν στον Hadoop. Η Yahoo, η οποία την ανέπτυξε για να κάνει τις εργασίες MapReduce σε μεγάλα σύνολα δεδομένων, την συνέβαλε στην ASF το 2007.
- Η Apache Solr είναι μια πλατφόρμα αναζήτησης επιχείρησης που βασίζεται στη Lucene.

Το Apache Zeppelin είναι ένα project επωάσεως που επιτρέπει την ανάλυση αλληλεπιδραστικών δεδομένων με SQL και άλλες γλώσσες προγραμματισμού.

Δεδομένου ότι τα μεγάλα δεδομένα συνεχίζουν να αυξάνονται σε μέγεθος και σπουδαιότητα, ο κατάλογος των εργαλείων ανοιχτού κώδικα θα συνεχίσει να αυξάνεται επίσης.

3 Τα μέσα κοινωνικής δικτύωσης

Ο περισσότερος κόσμος μπερδεύει τις έννοιες. Το Facebook, το Twitter, το Youtube, κτλ. είναι μέσα μαζικής δικτύωσης. Είναι απλά εργαλεία των Social Media. Έχεις email; Κι αυτό είναι μέρος των Social Media. Πολλοί το υποτιμούν, αλλά είναι ένα από τα δυνατά εργαλεία όπως και το κινητό, ειδικά αν είναι smartphone. Ο όρος μέσα κοινωνικής δικτύωσης (ή αλλιώς social media) αναφέρεται στα μέσα αλληλεπίδρασης και επικοινωνίας ομάδων ανθρώπων μέσω διαδικτυακών κοινοτήτων. Τα social media εμφανίζονται σε διάφορες μορφές όπως πχ. το Facebook, το Twitter, κ.α. Τα Μέσα Κοινωνικής Δικτύωσης αποτελούν την κοινωνική διάδραση μεταξύ ανθρώπων που δημιουργούν, μοιράζονται ή ανταλλάσσουν πληροφορίες και ιδέες μέσα σε εικονικές κοινότητες και δίκτυα. Τα κοινωνικά δίκτυα σήμερα θεωρείται ότι αποτελούν κυρίαρχο κομμάτι της καθημερινότητας των σύγχρονων ανθρώπων. Τα κοινωνικά δίκτυα διαθέτουν τα παρακάτω χαρακτηριστικά:

- Υποστηρίζουν ποικιλία των μορφών περιεχομένου, όπως κείμενο, βίντεο , φωτογραφίες , ήχο , κ.τ.λ. Πολλά από αυτά κάνουν χρήση περισσοτέρων του ενός από αυτές τις επιλογές ως προς το περιεχόμενο
- Επιτρέπουν αλληλεπιδράσεις που περνούν μία ή περισσότερες πλατφόρμες μέσω διαμοιρασμού, email και τροφοδοσίες
- Χαρακτηρίζονται από διαφορετικά επίπεδα εμπλοκής του χρήστη οι οποίοι μπορούν να δημιουργήσουν , να σχολιάσουν ή να παρακολουθούν σε δίκτυα Social Media
- Απλοποιούν, βελτιώνουν την ταχύτητα και το εύρος της διάδοσης των πληροφοριών
- Προσφέρουν ενός- προς-ένα, ενός-προς-πολλούς και πολλών προς-πολλούς επικοινωνία
- Επιτρέπουν την επικοινωνία αυτή να πραγματοποιείται είτε σε πραγματικό χρόνο ή ασύγχρονη με την πάροδο του χρόνου
- Είναι ανεξάρτητα της συσκευής: Ο χρήστης μπορεί να χρησιμοποιήσει για τη διείσδυση σε Social Media έναν υπολογιστή, ή κινητές συσκευές (tablets και smartphones ιδιαίτερα)
- Επεκτείνει εμπλοκή με τρεις τρόπους: με τη δημιουργία σε πραγματικό χρόνο online εκδηλώσεις, με την επέκταση σε απευθείας σύνδεση αλληλεπιδράσεις offline εκδηλώσεις, και τελευταία με την υποστήριξη ζωντανών εκδηλώσεων

Τα μέσα κοινωνικής δικτύωσης, επιπρόσθετα, είναι ικανά να συντελέσουν στην προβολή αφενός της κοινής γνώμης, καθώς καθιστούν τους χρήστες δέκτες αλλά και εκδότες περιεχομένου μέσω της ανατροφοδότησης, και αφετέρου παρέχουν κοινωνική και συναισθηματική υποστήριξη (Evans, 2008).

Αυτό που κάνει τα online κοινωνικά δίκτυα να ξεχωρίζουν από τις υπόλοιπες διαδικτυακές υπηρεσίες είναι:

(1) Τα εξελιγμένα εργαλεία που επιτρέπουν στους χρήστες να διαμοιράζονται ψηφιακά αρχεία (π.χ. κείμενο, εικόνες και άλλα) και

(2) τα εξελιγμένα εργαλεία για την επικοινωνία και την κοινωνικοποίηση των χρηστών (Cachia, Compañó & Costa, 2007). Οι ίδιοι συγγραφείς μάλιστα ομαδοποίησαν τα online κοινωνικά δίκτυα ανάλογα με την αλληλεπίδραση και την κοινωνικοποίηση που προσφέρει η κάθε ιστοσελίδα.

Από μία εναλλακτική κατηγοριοποίηση προκύπτουν οι παρακάτω επτά τύποι κοινωνικών δικτύων:

- Social news and recommendations (π.χ digg.com)
- Social book marking sites (π.χ delicious.com)
- Micro blogging services (twitter)
- Blogging systems (π.χ.blogger.com)
- Social networks (π.χ. facebook, linkedin)
- Social sharing (π.χ. youtube, flickr)
- Wikis(π.χ. mediawiki.org)

Τα μέσα κοινωνικής δικτύωσης, συμβάλλουν επαναστατικά στην ενημέρωση για τα τρέχοντα ζητήματα, την απρόσκοπτη έκφραση απόψεων και ιδεών, ενθαρρύνοντας τη συζήτηση, την έννοια του πλουραλισμού και τη διαλογικότητα. Ακόμη, λειτουργούν επικουρικά στη σύναψη κοινωνικών δεσμών, όπως φιλικοί. Ως αποτέλεσμα της Θεωρίας Χρήσεων και Ηθικών Ικανοποιήσεων (U and G Theory), που αναλύει τους λόγους επιλογής των μέσων από τους χρήστες, οι κινητήριες δυνάμεις είναι η ενημέρωση, η ψυχαγωγία, η κοινωνική αλληλεπίδραση και η δημιουργία προσωπικής ταυτότητας (Mc Quail, 1994)[4]. Το πόρισμα καταδεικνύει την άποψη ότι τα media συμβάλουν στην ενίσχυση της επικοινωνίας μεταξύ ατόμων σε οικουμενικό (universal) επίπεδο, ενώ μέχρι πρότινος αυτό αποτελούσε προνόμιο λίγων, που ανήκαν σε συγκεκριμένη κοινωνική τάξη (Boyd and Ellison, 2008 [5]).

Συμπερασματικά, τα μέσα κοινωνικής δικτύωσης αναπτύσσονται με ταχύτατους ρυθμούς, ενώ κατέχουν τη δύναμη της σχηματοποίησης και της διαστρέβλωσης της κοινής

γνώμης. Επομένως, μπορούν να αλλάξουν τον τρόπο σκέψης, προσέγγισης ενός θέματος και να επιφέρουν ευρύτερες κοινωνικές ανακατατάξεις, που ποτέ δεν είχαμε φανταστεί (Joison, 2008).

Τέλος μέχρι πρόσφατα, οι εταιρείες ήταν σε θέση να ελέγχουν τις διαθέσιμες πληροφορίες σχετικά με αυτές, μέσα από ανακοινώσεις Τύπου και του οργανωμένου τμήματος δημόσιων σχέσεων. Σήμερα, ωστόσο, οι επιχειρήσεις έχουν όλο και περισσότερο υποβιβαστεί στο περιθώριο ως απλοί παρατηρητές, αφού δεν διαθέτουν ούτε τις γνώσεις, ούτε την ευκαιρία – ή, μερικές φορές, ακόμα και το δικαίωμα – να τροποποιήσουν τα δημόσια σχόλια που υποβάλλονται από τους πελάτες τους (Kaplan & Haenlein, 2010).

Στα πλαίσια μιας ολοκληρωμένης στρατηγικής για την παρουσία στα Social Media, θέτονται κατάλληλοι στόχουι, σχεδιάζονται, αναπτύσσονται και υλοποιούνται οι απαραίτητες τακτικές ενέργειες προκειμένου να αποκομιστούν τα βέλτιστα οφέλη για το brand μιας επιχείρησης. Γίνεται επιλογή ποια είναι τα καταλληλότερα Social Media για το κάθε brand και καταστρώνεται στρατηγικά η παρουσία στο καθένα από αυτό. Μέσω της κοινωνικής δικτύωσης μια επιχείριση μπορεί να αλληλεπιδρά με το κοινό της, που σημαίνει πως θέτει τις βάσεις για το χτίσιμο μακροχρόνιων σχέσεων με όλους τους stakeholders και στην ουσία να πετύχει το engagement. Ως γνωστό, στα social media έχουν παρουσία όχι μόνο το κοινό-στόχος, αλλά και οι συνεργάτες, οι προμηθευτές, το προσωπικό της επιχείρησης και άλλοι που σχετίζονται με αυτήν. Επικοινωνώντας τακτικά με τα μέλη η επιχείρηση δεν έχει απλά μια παρουσία στα social media αλλά τα χρησιμοποιεί με ουσιαστικό τρόπο προσθέτοντας παράλληλα αξία. Σύμφωνα όμως με τον B. Borges στο βιβλίο του «Marketing 2.0: Bridging the Gap Between Seller and Buyer Through Social Media Marketing» (2009) υπάρχουν 4 στάδια αλληλεπίδρασης (interaction):

• Συμμετοχή (Engaging): Χρειάζεται συστηματική και τακτική χρήση της σελίδας, έτσι ώστε να αυξηθεί η αλληλεπίδραση με τους χρήστες της σελίδας. Το πότε και πόσο συχνά μέσα στη μέρα ή εβδομάδα είναι διαφορετικό για τον καθένα.

• Να ακούει (Listening): Θετικά και αρνητικά σχόλια που γράφονται από τους χρήστες στις σελίδες των επιχειρήσεων μπορούν να αντιμετωπιστούν κατάλληλα, αλλά σίγουρα όχι με την αδιαφορία. Εκτός αυτού, απαντώντας η επιχείρηση δείχνει πως ενδιαφέρεται να χτίσει σχέσεις με τους «οπαδούς της».

• Αλληλεπίδραση (Interacting): Η αλληλεπίδραση στα social media έρχεται με φυσικό τρόπο. Για παράδειγμα, ένα link που θα ανεβάσει κανείς, μια φωτογραφία, μια χρήσιμη ή ενδιαφέρουσα πληροφορία ή ακόμα και μια δημοσκόπηση είναι αρκετά για να αρχίσουν να κάνουν like/comment/share τα μέλη. Η φωτογραφία επίσης, ενός νέου προϊόντος ή η

περιγραφή μιας νέας υπηρεσίας θα δώσει τη δυνατότητα να πληροφορηθεί περισσότερο το κοινό αλλά και να τους προτρέψει να τα δοκιμάσουν. Με την κατάλληλη αντίδραση μπορεί κανείς να δημιουργήσει συζητήσεις μαζί τους, οι οποίες θα διαδοθούν στους φίλους των μελών και κατ' επέκταση στους φίλους των φίλων των μελών (viral effect).

• Μέτρηση (Measuring): Όλα τα social networks είναι ένα επιπλέον μέσο που μπορεί να προσδώσει αξία σε μια επιχείρηση. Η αποτελεσματικότητά τους λοιπόν, είναι επιτακτική ανάγκη να μετριέται. Η συγκεκριμένη μέτρηση, ωστόσο, έχει σημασία όταν θέσει κάποιος ορισμένους στόχους, έτσι ώστε να γίνει μια εποικοδομητική αξιολόγηση. Κάποιος στόχος μπορεί να αφορά τον αριθμό των fans, άρα αυτό που ενδιαφέρει να είναι ο αριθμός αντίστοιχα των New Likes ανά χρήστη ή ανά περίοδο και συνολικά, ο αριθμός των Likes. Άλλος στόχος μπορεί να είναι η αλληλεπίδραση, κι επομένως να ενδιαφέρουν η ποσότητα και η ποιότητα των σχολίων, καθώς και η συναισθηματική χροιά (θετικό/αρνητικό)

3.1 Facebook

Το Facebook ανήκει στην κατηγορία των Social Media, ή αλλιώς των Μέσων Κοινωνικής Δικτύωσης, και πιο συγκεκριμένα στην κατηγορία των Κοινωνικών Δικτύων. Αυτή τη στιγμή είναι το δημοφιλέστερο στο είδος του, έχοντας πάνω από 829 εκατ. καθημερινούς ενεργούς χρήστες και πάνω από 1,32 δισ. μηνιαίους ενεργούς χρήστες, σύμφωνα με τα τελευταία οικονομικά του αποτελέσματα.

Η χρήση του Facebook έχει να κάνει κυρίως με τα νέα φίλων και γνωστών, με την καθημερινή επικοινωνία μέσω μηνυμάτων και με την ενημέρωση από επιχειρήσεις και σελίδες άλλων ενδιαφερόντων. Το Facebook θεωρείται το πληρέστερο Κοινωνικό Δίκτυο σε λειτουργίες, αν και η πληθώρα ρυθμίσεων και επιλογών μπερδεύουν τους περισσότερους χρήστες, οι οποίοι αρκούνται στην απλή χρήση. Κάθε χρήστης άνω των 13 ετών μπορεί να δημιουργήσει το δικό του προφίλ, το οποίο μπορεί στη συνέχεια να το ενημερώσει με προσωπικές πληροφορίες, φωτογραφίες, βίντεο, ταινίες και μουσική που του αρέσουν, αθλητές που θαυμάζει κλπ. Με τη δημιουργία του προφίλ, ο χρήστης μπορεί να στείλει αιτήματα φιλίας σε άλλους χρήστες, οι οποίοι θα πρέπει να τον αποδεχθούν ή να τον απορρίψουν. Από τη στιγμή που ο χρήστης αποκτά φίλους, μπορεί και βλέπει τις δημοσιεύσεις τους και μερική από τη δραστηριότητα τους στην αρχική του σελίδα. Η αρχική σελίδα είναι ένας χώρος όπου ο Facebook χρήστης περνά τον περισσότερο χρόνο του. Εκεί, εκτός από τις δημοσιεύσεις και τη δραστηριότητα φίλων, μπορεί να δει δημοσιεύσεις από τις σελίδες που ακολουθεί κάνοντας like σε αυτές, και από τις ομάδες στις οποίες συμμετέχει.

Κάθε σελίδα μπορεί να αντιπροσωπεύει μια επιχείρηση, έναν άνθρωπο (αθλητή, ηθοποιό κλπ), ένα πάθος, οτιδήποτε μπορεί κανείς να φανταστεί, όπως και τα groups μπορούν να έχουν οποιαδήποτε θεματολογία. Εκτός από τις δημοσιεύσεις με κείμενο, φωτογραφίες, βίντεο που μπορεί να δημοσιεύσει ένας χρήστης, υπάρχει και η δυνατότητα για εισερχόμενα μηνύματα μέσω του chat-πεδίο συνομιλίας. Οι λειτουργίες τις οποίες χρησιμοποιούν περισσότερο οι χρήστες βρίσκονται στις δημοσιεύσεις και είναι το μου αρέσει, την κοινοποίηση και τα σχόλια. Ειδικά το «Μου αρέσει», βρίσκεται παντού στο ίντερνετ και οι χρήστες με συνδεδεμένο προφίλ μπορούν να το πατούν και να δηλώνουν πως τους αρέσει συγκεκριμένο περιεχόμενο. Το Facebook μπορεί κανείς να επισκεφθεί μέσω του www.facebook.com από υπολογιστή, μέσω του m.facebook.com από κινητές συσκευές, ή μέσω των Android, iOS, Windows Phone κλπ, εφαρμογών. Επίσης, υπάρχει και η διεύθυνση 0.facebook.com, η οποία λειτουργεί για δωρεάν πλοήγηση μέσω των κινητών συσκευών με χρήση 3G/4G, όχι μέσω WiFi, και η οποία απλώς δε δείχνει οπτικό υλικό όπως φωτογραφίες και βίντεο.

Εκτός, όμως, από τους προηγούμενους βασικούς τρόπους σύνδεσης, το Facebook παρέχει μία πληθώρα εφαρμογών για διάφορες λειτουργίες, όπως το υποχρεωτικό FB Messenger για τις συνομιλίες των χρηστών σε Android, iOS και Windows Phone, το FB Pages Manager για τη διαχείριση σελίδων σε Android και iOS, το FB Home ως launcher του Κοινωνικού Δικτύου για Android, και το FB Paper ως εναλλακτικό τρόπο πλοήγησης για iOS, προς το παρόν μονάχα για χρήστες στην Αμερική.

3.2 Twitter

Το Twitter είναι ένα Κοινωνικό Δίκτυο το οποίο ανήκει στην ευρύτερη κατηγορία των Social Media και θεωρείται το δεύτερο δημοφιλέστερο αυτή τη στιγμή πίσω από το Facebook. Σύμφωνα με τα τελευταία οικονομικά αποτελέσματα, το Twitter έχει 271 εκατ. ενεργούς μηνιαία χρήστες.

Αυτό που χαρακτηρίζει το συγκεκριμένο Κοινωνικό Δίκτυο από τα υπόλοιπα, είναι οι δημοσιεύσεις με όριο τους 140 χαρακτήρες, καθώς επίσης και η προώθηση του δημόσιου διαλόγου. Η χρήση του έχει να κάνει κυρίως με την ενημέρωση και στην Ελλάδα χαρακτηρίζεται ως το μέσο της ατάκας. Στο Twitter κάθε χρήστης δημιουργεί δωρεάν το δικό του προφίλ, το οποίο αποτελείται από εικόνες και διάφορες πληροφορίες, όπως μίνι βιογραφικό, περιοχή διαμονής, username. Το username, ή αλλιώς ψευδώνυμο, έχει πολύ μεγάλη σημασία καθώς χρησιμοποιείται περισσότερο και από το όνομα του χρήστη. Οι

περισσότεροι είναι γνωστοί από το ψευδώνυμο τους, καθώς ένας χρήστης για να αναφέρει έναν άλλον χρησιμοποιεί το username. Κάθε προφίλ μπορεί να δημοσιεύσει μηνύματα με μέγιστο όριο τους 140 χαρακτήρες, τα λεγόμενα και ως tweets. Από προεπιλογή, κάθε λογαριασμός δημοσιεύει δημόσια, με οποιονδήποτε χρήστη του ίντερνετ να μπορεί να προβάλλει τα μηνύματα του. Βέβαια, υπάρχει και επιλογή για ιδιωτικά tweets. Στο Twitter δε συναντάει κανείς φιλίες, αλλά ακόλουθους. Δεν είναι απαραίτητο για δύο χρήστες να ακολουθεί ο ένας τον άλλον. Κάθε χρήστης ακολουθεί όποιους θέλει και ακολουθείται από οποιονδήποτε. Αυτό σημαίνει πως ο καθένας διαμορφώνει την αρχική του σελίδα με περιεχόμενο το οποίο επιθυμεί, αν και το Twitter έκανε πρόσφατα κάποιες αλλαγές.

Σε σύγκριση με το Facebook το οποίο φιλτράρει και δείχνει στο χρήστη συγκεκριμένες δημοσιεύσεις στην αρχική σελίδα και όχι όλες, όποιες αυτό θεωρεί σημαντικότερες, στο Twitter εμφανίζονται όλα τα tweets των χρηστών που ακολουθεί κανείς. Μία από τις σημαντικότερες λειτουργίες των Κοινωνικών Δικτύων, τα hashtags, καθιερώθηκαν από το Twitter και χρησιμοποιούνται σε μεγάλο βαθμό για συγκέντρωση όλων των tweets γύρω από μια συζήτηση.

Οι κυριότερες λειτουργίες του Twitter είναι το retweet το οποίο χρησιμοποιείται για κοινοποίηση ενός tweet, το favorite το οποίο χρησιμοποιείται περισσότερο ως το like του Facebook, και το reply το οποίο χρησιμοποιείται για απάντηση σε ένα tweet. Το Twitter είναι ένα μέσο το οποίο ενισχύει το δημόσιο διάλογο, το οποίο το επιτυγχάνει με το όριο των 140 χαρακτήρων που δίνει σαν περιθώριο με το επιτρεπόμενο μέγεθος κειμένου.

3.3 Instagram

Το Instagram είναι μία δημοφιλής mobile social εφαρμογή και συγχρόνως μία υπηρεσία κοινωνικής δικτύωσης, η οποία επιτρέπει τη λήψη και το διαμοιρασμό φωτογραφιών και βίντεο. Έγινε γνωστό χάρη στα φίλτρα φωτογραφιών του, ενώ σήμερα διαθέτει φίλτρα και για βίντεο, καθώς επίσης και πληθώρα άλλων εργαλείων φιλικών προς τους χρήστες.

Το Instagram ανήκει πλέον στο Facebook, με την εξαγορά να έχει πραγματοποιηθεί τον Απρίλιο του 2012 έναντι 1 δισ. δολαρίων. Συνολικά υπάρχουν πάνω από 200 εκατ. εγγεγραμμένοι χρήστες, οι οποίοι έχουν ανεβάσει πάνω από 20 δισ. φωτογραφίες, ανεβάζουν καθημερινά πάνω από 60 εκατ. φωτογραφίες και πραγματοποιούν καθημερινά πάνω από 1,6 δισ. likes σε φωτογραφίες και βίντεο. Κάθε χρήστης έχει το δικό του προφίλ, το οποίο μπορεί να εμπλουτίσει με φωτογραφία προφίλ, βιογραφικό, σύνδεσμο (προς κάποιο site) κλπ. Από

προεπιλογή, κάθε προφίλ είναι δημόσιο προς προβολή, όπως και οι φωτογραφίες και τα βίντεο του, κάτι το οποίο μπορεί να αλλάξει μέσω των ρυθμίσεων.

Στην συγκεκριμένη εφαρμογή δεν υφίστανται φιλίες σε αντίθεση με το Facebook. Αντιθέτως, λειτουργεί η φιλοσοφία των followers, δηλαδή κάθε χρήστης ακολουθεί όποιους χρήστες επιθυμεί και ακολουθείται από όσους ενδιαφέρονται για το περιεχόμενο του. Ακολουθώντας χρήστες, οι φωτογραφίες και τα βίντεο τους προβάλλονται μέσω της αρχικής οθόνης του χρήστη. Κάθε χρήστης μπορεί να χρησιμοποιήσει τη λειτουργία της κάμερας της εφαρμογής, έτσι ώστε να τραβήξει μια φωτογραφία ή ένα βίντεο (μέγιστης διάρκειας 15 δευτερολέπτων) και στη συνέχεια να προβεί στην επεξεργασία μέσω των διαφόρων φίλτρων και εργαλείων. Υπάρχει επίσης και η δυνατότητα επιλογής παλαιών φωτογραφιών και βίντεο από τη μνήμη της συσκευής. Το ανέβασμα φωτογραφιών και βίντεο είναι εφικτό μονάχα μέσω των κινητών συσκευών και όχι μέσω υπολογιστή. Από υπολογιστή μπορεί κανείς να δει το προφίλ του, την αρχική σελίδα του με φωτογραφίες και βίντεο όσων ακολουθεί, τις οποίες μπορεί να κάνει like και να αφήσει σχόλιο, ή να κάνει follow νέους χρήστες. Κάθε φωτογραφία ή βίντεο δέχεται likes και σχόλια, αφού δεν υπάρχει η δυνατότητα για κοινοποίηση φωτογραφιών, ούτε για κατέβασμα και αποθήκευση. Κάθε χρήστης έχει στο προφίλ του δικό του περιεχόμενο.

Το Instagram υποστηρίζει τη λειτουργία των hashtags, ομαδοποιώντας φωτογραφίες και βίντεο γύρω από συγκεκριμένα θέματα, καθώς επίσης και τη λειτουργία tagging, προσθέτοντας με ετικέτα χρήστες σε φωτογραφίες, και τη λειτουργία προσωπικών μηνυμάτων. Τέλος, η εφαρμογή διαθέτει πολύ καλές λειτουργίες για εξερεύνηση νέων φωτογραφιών και βίντεο. Εκτός των hashtags, υπάρχει ξεχωριστή επιλογή «Εξερεύνηση» με φωτογραφίες και βίντεο τα οποία είτε είναι δημοφιλή σε όλο τον κόσμο, είτε έχουν σχολιαστεί ή δεχτεί like από όσους ακολουθεί ο χρήστης. Τέλος, υπάρχουν και ενημερώσεις οι οποίες μαρτυρούν τη δραστηριότητα όσων ακολουθούμε.

3.4 LinkedIn

Το LinkedIn ανήκει στα Social Media, λεγόμενα και ως Κοινωνικά Μέσα, και αποτελεί το σημείο συνάντησης όλων των επαγγελματιών. Πρόκειται συγκεκριμένα για ένα Κοινωνικό Δίκτυο, στόχος του οποίου είναι να συνδέσει όλους τους επαγγελματίες του κόσμου, κάνοντας τους πιο παραγωγικούς και καλύτερους στην εργασία τους. Το επαγγελματικό Κοινωνικό Δίκτυο μετρά αυτή τη στιγμή πάνω από 313 εκατομμύρια χρήστες, σύμφωνα με τα τελευταία οικονομικά αποτελέσματα τα οποία δημοσίευσε.

Η χρήση του LinkedIn έχει να κάνει κυρίως με την ενημέρωση γύρω από τις επαφές του χρήστη, τον κλάδο εργασίας του, όλα τα προηγούμενα σε πιο σοβαρό ύφος σε σχέση με τα υπόλοιπα Κοινωνικά Δίκτυα. Επίσης, το επαγγελματικό Κοινωνικό Δίκτυο χρησιμοποιείται από τις εταιρίες και ως ένα online βιογραφικό, με το οποίο οι χρήστες μπορούν να βρουν εργασία απαντώντας σε αγγελίες εργασίας μέσα σε αυτό.

Στο LinkedIn κάθε χρήστης δημιουργεί το δικό του προφίλ, στο οποίο μπορεί να προσθέσει πληροφορίες όπως την τωρινή και προηγούμενη απασχόληση του, τις εθελοντικές του δράσεις, τις γλώσσες τις οποίες γνωρίζει, τις δεξιότητες του, την εκπαίδευση του κλπ. Περιέχει πληροφορίες, δηλαδή, οι οποίες μπορούν να βοηθήσουν το χρήστη να βρει εργασία ή να επισυνάψει συνεργασίες και σχέσεις με άλλους επαγγελματίες. Κάθε προφίλ στέλνει αίτημα σύνδεσης σε άλλα προφίλ, κάτι το οποίο σημαίνει πως πρέπει τα άλλα προφίλ να αποδεχτούν ή να απορρίψουν το αίτημα. Τον τελευταίο καιρό παρατηρήθηκε επίσης μία επιλογή follow, η οποία συνδέει ένα χρήστη με τις δημόσιες πληροφορίες και τις δημοσιεύσεις ενός άλλου χρήστη. Εκτός από τις ενέργειες για like, σχόλιο και κοινοποίηση των δημοσιεύσεων άλλων χρηστών, οι χρήστες μπορούν να αφήνουν συστάσεις στα προφίλ του δικτύου τους, καθώς επίσης και να επιβεβαιώνουν δεξιότητες τις οποίες έχουν προσθέσει, ή ακόμη δεν έχουν προσθέσει, οι επαφές τους, μέσω της λειτουργίας Endorsements.

Το επαγγελματικό Κοινωνικό Δίκτυο περιέχει επίσης groups και εταιρικές σελίδες, στα οποία ο χρήστης μπορεί να ενημερώνεται γύρω από ένα συγκεκριμένο θέμα ή μία συγκεκριμένη εταιρία, αντίστοιχα. Συνεχίζοντας, το LinkedIn προσφέρει στους χρήστες δυνατότητα για εύρεση εργασίας από αγγελίες εργασίας τις οποίες δημοσιεύουν οι εταιρικές σελίδες. Τέλος, το LinkedIn διαθέτει πολλές ακόμη λειτουργίες, όπως για παράδειγμα για μαθητές οι οποίοι αναζητούν το επαγγελματικό μέλλον τους, για επαγγελματίες οι οποίοι θέλουν να μάθουν περισσότερα για τη σχέση σπουδών και εργασίας, για μαθητές οι οποίοι αναζητούν Πανεπιστήμιο και τομέα σπουδών, κλπ.

Σε αντίθεση με τα υπόλοιπα Κοινωνικά Δίκτυα, το LinkedIn παρέχει premium πακέτα στους χρήστες τους, με τα οποία εκείνοι αποκτούν πρόσβαση σε περισσότερα ή βελτιωμένα εργαλεία σε σύγκριση με τους απλούς χρήστες.

4 Ανάλυση του προβλήματος

4.1 Τεχνικές Προδιαγραφές του Προβλήματος

Τα Recurrent Neural Networks έχουν διερευνηθεί ευρέως και χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας και γενικά σε διαδοχικά προβλήματα. Στο κομμάτι που θα ασχοληθεί η συγκεκριμένη εργασία, για να προβλεφθεί ο τύπος της προσωπικότητας των υποψήφιων ατόμων από εταιρεία HR μέσω των δημοσιεύσεων τους στα κοινωνικά δίκτυα όπως το Facebook και το Twitter.

Το Facebook είναι μια πλατφόρμα κοινωνικής δικτύωσης που ξεκίνησε τον Φεβρουάριο του 2004. Οι χρήστες της μπορούν να επικοινωνούν μεταξύ τους μέσω μηνυμάτων με τις επαφές τους και να τους ειδοποιούν όταν ανανεώνουν τις προσωπικές πληροφορίες τους. Αρχικά δικαίωμα συμμετοχής είχαν μόνο οι φοιτητές του Χάρβαρντ αλλά από το 2006 η υπηρεσία έγινε προσβάσιμη σε κάθε άνθρωπο του πλανήτη που η ηλικία του ξεπερνούσε τα 13 χρόνια. Ομοίως και το Twitter είναι ένας ιστοχώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες να στέλνουν και να διαβάζουν μηνύματα μεγέθους έως 280 χαρακτήρων. Τα μηνύματα μπορούν να διαβαστούν και από μη συνδεδεμένους χρήστες αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν κείμενα. Και οι δύο προαναφερόμενοι ιστοχώροι ανήκουν στους δέκα πιο δημοφιλείς ιστότοπους του διαδικτύου. Υπενθυμίζοντας όπως ειπώθηκε και στην εισαγωγή στην συγκεκριμένη διατριβή θα φανεί πώς μπορεί να βοηθηθεί μια εταιρεία HR στην λήψη αποφάσεων αξιοποιώντας προς όφελός της το μέχρι τώρα πρόβλημά της και πιο συγκεκριμένα τα δεδομένα μεγάλης κλίμακας αξιοποιώντας τα με την βοήθεια της μηχανικής μάθησης αποκτώντας έτσι σημαντικό πλεονέκτημα έναντι των ανταγωνιστών της.

Για αυτό και θα παρουσιαστεί ο τρόπος επίλυσης αυτού του προβλήματος με την χρήση ενός ανατροφοδοτούμενου αλγορίθμου ο οποίος σαν είσοδο θα χρησιμοποιεί τα σχόλια των υποψηφίων από τους λογαριασμούς τους στα μέσα μαζικής δικτύωσης για μία θέση τα οποία στην πραγματικότητα θα είναι και τα δεδομένα μεγάλης κλίμακας. Παρακάτω παρατίθενται οι τύποι προσωπικότητας της επιστήμης του ανθρώπινου δυναμικού τα οποία και για τον αλγόριθμό μας θα λειτουργήσουν σαν τις ομάδες στις οποίες θα κατηγοριοποιούνται οι υποψήφιοι.

4.2 Τύποι Προσωπικότητας

Σε αυτήν την εργασία σκοπός είναι να προβλεφθεί η προσωπικότητα κάποιου συγκεκριμένου ατόμου μέσα από τον μεγάλο όγκο βιογραφικών που διαθέτουν πλέον οι εταιρείες εύρεσης προσωπικού, χρησιμοποιώντας τις αναρτήσεις του στα social media για τις ανάγκες της εταιρείας ώστε να γίνεται γρήγορα φιλτράρισμα και ταξινόμηση υποψηφίων και να προχωρούν σε συνέντευξη. Θα χρησιμοποιηθεί το σύνολο προσωπικοτήτων Myers-Briggs που δόθηκε από το Kaggle στο οποίο και γίνεται ιδιαίτερη ανάλυση από τον Drenth (2017). Παρακάτω, θα γίνει μια πιο ενδελεχής αναφορά στις διαφορετικές προσωπικότητες που γενικότερα υπάρχουν, ώστε να μπορέσουν να κατηγοριοποιηθούν οι υποψήφιοι σε μία από αυτές. Η εισαγωγή αυτή γίνεται επειδή είναι ιδιαίτερα σημαντικό να υπάρχει η κατανόηση των δεδομένων προτού εφαρμοστεί η μηχανική μάθηση σε αυτά.

Σε μια περιληπτική περιγραφή, η Myers-Briggs χωρίζει την προσωπικότητα σε τέσσερους διαφορετικούς τομείς με βάση το συνδιασμό των οποίων καθορίζεται η συνολική προσωπικότητα ενός ατόμου. Αυτοί οι τέσσερις τομείς είναι οι εξής:

1. Ο τρόπος με τον οποίο κάποιος επικεντρώνεται –Εξωστρέφεια(E) ή Εσωστρέφεια(I)
2. Ο τρόπος με τον οποίο λαμβάνει κάποιος πληροφορίες–Αίσθηση(S) ή Διαίσθηση(N)
3. Ο τρόπος με τον οποίο κάποιος λαμβάνει αποφάσεις –Σκέψη(T) ή Συναίσθημα(F)
4. Ο τρόπος με τον οποίο κάποιος αντιμετωπίζει τον εξωτερικό κόσμο–Κρίση(J) ή Αντίληψη(P)

Συνεπώς από τη Myers-Briggs προκύπτουν 16 διαφορετικοί τύποι προσωπικότητας οι οποίοι εκφράζονται μέσω ενός τετραψήφιου κώδικα και όλοι μαζί αποτελούν το δείκτη προσωπικότητας Myers-Briggs.

4.2.1 Φαινομενικός Τύπος Προσωπικότητας

Τα άτομα με φαινομενική προσωπικότητα τείνουν να είναι ζωντανοί, αισιόδοξοι, δυναμικοί και ξέγνοιαστοι. Αγαπούν την περιπέτεια και έχουν ανοχή υψηλού κινδύνου. Τυπικά, οι φαινομενικοί άνθρωποι είναι πολύ κακοί στην ανάρρηση της πλήξης και θα αναζητήσουν ποικιλία και διασκέδαση. Περιττό να ειπωθεί ότι αυτό το χαρακτηριστικό μπορεί μερικές φορές να επηρεάσει αρνητικά τις ρομαντικές σχέσεις τους. Επειδή αυτή η ιδιοσυγκρασία είναι επιρρεπής σε συμπεριφορές που αναζητούν ευχαρίστηση, πολλοί άνθρωποι με αυθόρμητη προσωπικότητα είναι πιθανό να αγωνιστούν με εθισμούς. Οι συνεχείς εθισμοί τους μπορεί να οδηγήσουν σε υπερκατανάλωση τροφής και προβλήματα βάρους.

Αυτοί οι άνθρωποι είναι πολύ δημιουργικοί και μπορεί να γίνουν σπουδαίοι καλλιτέχνες. Επιπλέον, είναι φανταστικοί διασκεδαστές και θα κάνουν φυσικά καλά αν επιλέξουν σταδιοδρομία στη βιομηχανία ψυχαγωγίας. Οι φυσικές τους ικανότητες θα τους εξυπηρετήσουν καλά αν επιλέξουν θέσεις εργασίας που σχετίζονται με μάρκετινγκ, αθλήματα, μαγειρική και γενικά οτιδήποτε έχει να κάνει με δημιουργικότητα.

4.2.2 Φλεγματικός τύπος προσωπικότητας

Κάποιος με φλεγματικό τύπο προσωπικότητας είναι συνήθως άνθρωπος ανθρώπων. Επιδιώκουν τη διαπροσωπική αρμονία και τις στενές σχέσεις. Οι φλεγματικοί άνθρωποι είναι πιστοί σύζυγοι και αγαπητοί γονείς. Διατηρούν τις σχέσεις τους με παλιούς φίλους, μακρινά μέλη της οικογένειας και γείτονες. Οι άνθρωποι με φλεγματικό ταπεραμέντο τείνουν να αποφεύγουν τις συγκρούσεις και πάντα προσπαθούν να μεσολαβούν μεταξύ των άλλων για να αποκαταστήσουν την ειρήνη και την αρμονία. Είναι φιλάνθρωποι και βοηθούν τους άλλους. Οι ιδανικές επαγγελματικές σταδιοδρομίες για φλεγματικούς τύπους προσωπικότητας πρέπει να σχετίζονται με: κοινωνικές υπηρεσίες, διδασκαλία, κοινωνική ανάπτυξη κλπ.

4.2.3 Χολερικός Τύπος Προσωπικότητας

Κάποιος με καθαρή «χολική ιδιοσυγκρασία» είναι συνήθως πρόσωπο με προσανατολισμό στόχου. Τα άτομα με χολική προσωπικότητα είναι πιο πολύ της ιδεολογίας του καταλαβαίνω, αναλυτικά και λογικά. Εξαιρετικά πρακτικοί και απλοί, χολικοί άνθρωποι δεν είναι απαραίτητα πολύ καλοί σύντροφοι ή ιδιαίτερα κοινωνικοί. Δεν τους αρέσουν οι μικρές συζητήσεις και απολαμβάνουν βαθιές και ουσιαστικές συνομιλίες. Θα προτιμούσαν να είναι μόνοι, παρά με αβαθή, επιφανειακά άτομα. Στην ιδανική περίπτωση, θέλουν να περάσουν χρόνο με ανθρώπους που έχουν παρόμοια επαγγελματικά συμφέροντα. Οι ιδανικές θέσεις εργασίας για τον τύπο της χολικής προσωπικότητας σχετίζονται με τις ακόλουθες βιομηχανίες: διαχείριση, τεχνολογία, στατιστική, μηχανική, προγραμματισμός.

4.2.4 Μελαγχολικός Τύπος Προσωπικότητας

Τα άτομα με μελαγχολική προσωπικότητα αγαπούν τις παραδόσεις. Οι γυναίκες μαγειρεύουν για τους άνδρες. οι άνδρες ανοίγουν τις πόρτες για γυναίκες. Αγαπούν τις οικογένειες και τους φίλους τους και, αντίθετα από την ιδιοσυγκρασία, δεν αναζητούν καινοτομία και περιπέτεια. Στην πραγματικότητα, το αποφεύγουν πάση θυσία. Κάποιος με μελαγχολική ιδιοσυγκρασία είναι πολύ απίθανο να παντρευτεί έναν ξένο ή να εγκαταλείψει την πατρίδα του για άλλη χώρα. Είναι πολύ κοινωνικοί και επιδιώκουν να συμβάλλουν στην κοινότητα. Όντας εξαιρετικά τακτοποιημένοι και ακριβείς, οι μελαγχολικοί άνθρωποι είναι

φανταστικοί διευθυντές ανθρώπων. Η τέλεια σταδιοδρομία για τον τύπο της μελαγχολικής προσωπικότητας πρέπει να είναι: διαχείριση, λογιστική, κοινωνική εργασία. Μέσω των παραπάνω τεσσάρων προσωπικοτήτων οι οποίες αναφέρονται αποκλειστικά στο "ταμπεραμέντο" του κάθε ανθρώπου έγινε μια επέκταση των παραπάνω σε δεκαέξι πιο συγκεκριμένες προσωπικότητες οι οποίες είναι και αυτές που θα χρησιμοποιήσουμε, παρακάτω λίγα λόγια για την κάθε μια.

4.2.5 Ο Επιθεωρητής – ISTJ Προσωπικότητα

Με την πρώτη ματιά, οι ISTJs είναι εκφοβιστικοί. Φαίνονται σοβαροί, επίσημοι και σωστοί. Αγαπούν επίσης τις παραδόσεις και τις παλιές σχολικές αξίες που υποστηρίζουν την υπομονή, τη σκληρή δουλειά, την τιμή και την κοινωνική και πολιτιστική ευθύνη. Είναι αποκλειστικά, ήρεμοι και ήσυχοι. Αυτά τα χαρακτηριστικά προέρχονται από το συνδυασμό I (Εξωστρέφεια), S (Αίσθηση), T (Σκέψη) και J (Κρίση), ενός τύπου προσωπικότητας που συχνά παρεξηγείται.

4.2.6 Ο Σύμβουλος – INFJ

Τα INFJs (Εξωστρέφεια, Διαίσθηση, Συναίσθημα, Κρίση) είναι οραματιστές και ιδεαλιστές που εξαντλούν τη δημιουργική φαντασία και τις λαμπρές ιδέες. Έχουν διαφορετικό και συνήθως πιο βαθύ τρόπο να κοιτάζουν τον κόσμο. Έχουν μια ουσία και βάθος στον τρόπο που σκέφτονται, δεν παίρνουν ποτέ τίποτα στο επίπεδο της επιφάνειας ή δεχόμενα τα πράγματα όπως είναι. Άλλοι μπορεί μερικές φορές να τους αντιλαμβάνονται ως περίεργους ή διασκεδαστικούς λόγω της διαφορετικής τους προοπτικής στη ζωή.

4.2.7 Ο Εγκέφαλος – INTJ

Οι INTJs (Εσωστρέφεια, Διαίσθηση, Σκέψη, Κρίση) ως εσωστρεφείς, είναι ήσυχοι, προστατευμένοι και άνετοι που είναι μόνοι. Συνήθως είναι αυτάρκεις και προτιμούν να δουλεύουν μόνοι τους παρά σε μια ομάδα. Η κοινωνικοποίηση αποστραγγίζει την ενέργεια ενός εσωστρεφούς, προκαλώντας την ανάγκη επαναφόρτισης. Οι INTJ ενδιαφέρονται για ιδέες και θεωρίες. Όταν παρατηρούν τον κόσμο, πάντα αμφισβητούν γιατί συμβαίνουν τα πράγματα. Διακρίνονται στην ανάπτυξη σχεδίων και στρατηγικών και δεν τους αρέσει η αβεβαιότητα.

4.2.8 Ο Δωρητής - ENFJ Προσωπικότητα

Τα ENFJs (Εξωστρέφεια, Διαίσθηση, Συναίσθημα, Κρίση) είναι άτομα με επίκεντρο τον άνθρωπο. Είναι εξωστρεφείς, ιδεαλιστές, χαρισματικοί, ειλικρινείς, υψηλόβαθμοι και ηθικοί και συνήθως ξέρουν πώς να συνδεθούν με τους άλλους ανεξάρτητα από το υπόβαθρο ή την προσωπικότητά τους. Βασιζόμενοι κυρίως στη διαίσθηση και τα συναισθήματα, τείνουν να ζουν στη φαντασία τους παρά στον πραγματικό κόσμο. Αντί να εστιάζουμε στο να ζούμε στο "τώρα" και αυτό που συμβαίνει σήμερα, οι ENFJ τείνουν να επικεντρώνονται στο αφηρημένο και τι θα μπορούσε ενδεχομένως να συμβεί στο μέλλον.

4.2.9 Ο Τεχνίτης - προσωπικότητα ISTP

Οι ISTPs (Εσωστρέφεια, Αίσθηση, Σκέψη, Αντίληψη) είναι μυστηριώδεις άνθρωποι που είναι συνήθως πολύ λογικοί, αλλά και αρκετά αυθόρμητοι και ενθουσιώδεις. Τα χαρακτηριστικά της προσωπικότητάς τους είναι λιγότερο εύκολα αναγνωρίσιμα από αυτά των άλλων τύπων και ακόμη και οι άνθρωποι που τα γνωρίζουν καλά δεν μπορούν πάντα να προβλέψουν τις αντιδράσεις τους. Σε βάθος, οι ISTPs είναι αυθόρμητα, απρόβλεπτα άτομα, αλλά κρύβουν αυτά τα χαρακτηριστικά από τον έξω κόσμο, συχνά με μεγάλη επιτυχία.

4.2.10 Ο Παροχέας - Προσωπικότητα του ESFJ

Τα ESFJ (Εξωστρέφεια, Αίσθηση, Συναίσθημα, Κρίση) είναι οι στερεότυποι εξωστρεφείς. Είναι κοινωνικές πεταλούδες και η ανάγκη τους να αλληλοεπιδρούν με τους άλλους και να κάνουν τους ανθρώπους ευτυχείς συνήθως καταλήγει να τους κάνουν δημοφιλείς. Ο ESFJ τείνει συνήθως να είναι μαζορέτα ή αθλητικός ήρωας στο γυμνάσιο και στο κολέγιο. Αργότερα στη ζωή, συνεχίζουν να απολαμβάνουν το επίκεντρο και επικεντρώνονται κυρίως στην οργάνωση κοινωνικών εκδηλώσεων για τις οικογένειες, τους φίλους και τις κοινότητές τους. Το ESFJ είναι ένας κοινός τύπος προσωπικότητας και ένας που αγαπά πολλοί άνθρωποι.

4.2.11 Ο Ιδεαλιστής - προσωπικότητα INFP

Οι INFP (Εσωστρέφεια, Διαίσθηση, Συναίσθημα, Αντίληψη), όπως και οι περισσότεροι εσωστρεφείς, είναι ήσυχοι και προστατευμένοι. Προτιμούν να μην μιλάνε για τον εαυτό τους, ειδικά στην πρώτη συνάντηση με ένα νέο άτομο. Τους αρέσει να περνούν μόνοι τους χρόνο σε ήσυχα μέρη όπου μπορούν να κατανοήσουν τι συμβαίνει γύρω τους. Τους αρέσει να αναλύουν σημάδια και σύμβολα και να τα θεωρούν ως μεταφορές με βαθύτερες έννοιες που

σχετίζονται με τη ζωή. Χάνονται στη φαντασία τους και στις ονειροπολήσεις τους, πνίγονται πάντα στο βάθος των σκέψεων, των φαντασιών και των ιδεών τους.

4.2.12 Ο Ερμηνευτής - προσωπικότητα ESFP

Τα ESFP (Εξωστρέφεια, Αίσθηση, Συναίσθημα, Αντίληψη), έχουν μια Εξαιρετική, Παρατηρητική, Αισθησιακή και Αντιληπτική προσωπικότητα και συνήθως θεωρούνται Διασκεδαστές. Γεννημένοι να είναι μπροστά σε άλλους και να καταγράψουν τη σκηνή, τα ESFP αγαπούν το φως του προβολέα. Τα ESFPs είναι προσεκτικοί εξερευνητές που αγαπούν να μαθαίνουν και να μοιράζονται αυτό που μαθαίνουν με άλλους. Είναι "άνθρωποι ανθρώπων" με ισχυρές διαπροσωπικές δεξιότητες. Είναι ζωντανές και διασκεδαστικές και απολαμβάνουν το κέντρο της προσοχής. Είναι ζεστοί, γενναιόδωροι και φιλικοί, συμπαθητικοί και προβληματισμένοι για την ευημερία των άλλων ανθρώπων.

4.2.13 Ο Πρωταθλητής - Προσωπικότητα της ENFP

Τα ENFP (Εξωστρέφεια, Διαίσθηση, Συναίσθημα, Αντίληψη) έχουν μια εξατομικευμένη, διαισθητική, αίσθηση και αντιληπτική προσωπικότητα. Αυτός ο τύπος προσωπικότητας είναι ιδιαίτερα ατομικιστικός και οι πρωταθλητές προσπαθούν να δημιουργήσουν τις δικές τους μεθόδους, εμφάνιση, ενέργειες, συνήθειες και ιδέες - δεν τους αρέσουν οι άνθρωποι που είναι αμέτοχοι και μισούν όταν αναγκάζονται να ζήσουν μέσα σε ένα κουτί. Τους αρέσει να είναι γύρω από άλλους ανθρώπους και να έχουν μια ισχυρή διαισθητική φύση όταν πρόκειται για τους εαυτούς τους και τους άλλους. Λειτουργούν από τα συναισθήματά τους τις περισσότερες φορές, και είναι πολύ αντιληπτικοί και προσεκτικοί.

4.2.14 Ο Δραστήριος - ESTP Προσωπικότητα

Οι ESTP (Εξωστρέφεια, Αίσθηση, Σκέψη, Αντίληψη) έχουν μια εξατομικευμένη, ευαίσθητη, σκεπτόμενη και αντιληπτή προσωπικότητα. Οι ESTPs διέπονται από την ανάγκη για κοινωνική αλληλεπίδραση, συναισθήματα, λογικές διαδικασίες και συλλογιστική, μαζί με την ανάγκη για ελευθερία. Η θεωρία και οι περιλήψεις δεν κρατούν να ESTP ενδιαφερόμενα για μεγάλο χρονικό διάστημα. Τα ESTPs καθορίζουν τα λάθη τους καθώς προχωράνε, αντί να κάθονται σε αδράνεια ή να προετοιμάζουν σχέδια έκτακτης ανάγκης.

4.2.15 Ο Επόπτης - ESTJ Προσωπικότητα

Οι ESTJs (Εξωστρέφεια, Αίσθηση, Σκέψη, Κρίση) είναι οργανωμένοι, ειλικρινείς, αφοσιωμένοι, αξιοπρεπείς, παραδοσιακοί και είναι μεγάλοι πιστούχοι να κάνουν αυτό που πιστεύουν ότι είναι σωστό και κοινωνικά αποδεκτό. Αν και τα μονοπάτια προς το "καλό" και το "σωστό" είναι δύσκολα, χαίρονται να πάρουν τη θέση τους ως ηγέτες του πακέτου. Είναι η επιτομή των καλών πολιτών. Οι άνθρωποι αναζητούν ESTJs για καθοδήγηση και συμβουλές, και οι ESTJs είναι πάντα ευτυχείς που προσεγγίζονται για βοήθεια.

4.2.16 Ο Διοικητής - ENTJ Προσωπικότητα

Ο βασικός τρόπος ζωής του ENTJ (Εξωστρέφεια, Διαίσθηση, Σκέψη, Κρίση) επικεντρώνεται σε εξωτερικές πτυχές και όλα τα πράγματα αντιμετωπίζονται λογικά Ο δευτερεύων τρόπος λειτουργίας τους είναι εσωτερικός, όπου η διαίσθηση και η συλλογιστική ισχύουν. Τα ENTJs είναι φυσικοί γεννημένοι ηγέτες ανάμεσα στους 16 τύπους προσωπικότητας και σαν να είναι υπεύθυνοι. Ζουν σε έναν κόσμο δυνατοτήτων και συχνά βλέπουν τις προκλήσεις και τα εμπόδια, όπως τις μεγάλες ευκαιρίες να προωθούνται. Φαίνεται να έχουν ένα φυσικό δώρο για ηγεσία, τη λήψη αποφάσεων και την εξέταση επιλογών και ιδεών γρήγορα αλλά προσεκτικά. Αυτοί είναι οι άνθρωποι που "παίρνουν" τα χρήματα που δεν τους αρέσει να κάθονται ακίνητοι.

4.2.17 Η προσωπικότητα του Λογικού – INTP

Οι INTPs (Εσωστρέφεια, Διαίσθηση, Σκέψη, Αντίληψη) είναι γνωστοί για τις λαμπρές θεωρίες τους και την αμείλικτη λογική, η οποία έχει νόημα, αφού είναι αναμφισβήτητα το πιο λογικό από όλους τους τύπους προσωπικότητας. Λατρεύουν τα σχέδια, έχουν έντονο μάτι για να καταλαβαίνουν τις διαφορές και μια καλή ικανότητα να διαβάζουν ανθρώπους, καθιστώντας την κακή ιδέα να βρεθεί σε ένα INTP. Οι άνθρωποι αυτού του τύπου προσωπικότητας δεν ενδιαφέρονται για πρακτικές καθημερινές δραστηριότητες και συντήρηση, αλλά όταν βρουν ένα περιβάλλον όπου η δημιουργική ιδιοφυΐα και το δυναμικό τους μπορούν να εκφραστούν, δεν υπάρχει όριο στον χρόνο και την ενέργεια που θα δαπανήσουν οι INTP αναπτύσσοντας μια διορατική και αμερόληπτη λύση.

4.2.18 Ο Φιλάνθρωπος - ISFJ Προσωπικότητα

Οι ISFJs (Εσωστρέφεια, Αίσθηση, Συναίσθημα, Κρίση) είναι φιλάνθρωποι και είναι πάντα έτοιμοι να δώσουν πίσω και να επιστρέψουν γενναιοδωρία με ακόμα μεγαλύτερη γενναιοδωρία. Οι άνθρωποι και τα πράγματα που πιστεύουν θα διατηρηθούν και θα υποστηριχθούν με ενθουσιασμό και ανιδιοτέλεια. Τα ISFJ είναι ζεστά και ευγενικά. Εκτιμούν την αρμονία και τη συνεργασία και είναι πιθανό να είναι πολύ ευαίσθητα στα συναισθήματα των άλλων ανθρώπων. Οι άνθρωποι εκτιμούν τα ISFJ για την εκτίμησή τους και την ευαισθητοποίησή τους, καθώς και την ικανότητά τους να φέρνουν το καλύτερο σε άλλους.

4.2.19 Ο Οραματιστής - Προσωπικότητα ENTP

Η προσωπικότητα ENTP (Εξωστρέφεια, Διαίσθηση, Σκέψη, Αντίληψη) είναι από τις πιο σπάνιες στον κόσμο, κάτι που είναι απολύτως κατανοητό. Αν και είναι εξωστρεφείς, δεν απολαμβάνουν μικρές ομιλίες και δεν μπορούν να ευδοκιμήσουν σε πολλές κοινωνικές καταστάσεις, ειδικά εκείνες που περιλαμβάνουν ανθρώπους που είναι πολύ διαφορετικοί από το ENTP. Τα ENTP είναι ευφυή και καλά ενημερωμένα πρέπει να είναι διαρκώς ψυχικά διεγερμένα. Έχουν τη δυνατότητα να συζητούν θεωρίες και γεγονότα σε εκτενείς λεπτομέρειες. Είναι λογικοί, ορθολογικοί και αντικειμενικοί στην προσέγγιση των πληροφοριών και των επιχειρημάτων.

4.2.20 Ο Συνθέτης - Προσωπικότητα ISFP

Τα ISFP (Εσωστρέφεια, Αίσθηση, Συναίσθημα, Αντίληψη) είναι εσωστρεφείς που δεν φαίνονται σαν εσωστρεφές. Είναι επειδή, ακόμη και αν έχουν δυσκολίες στη σύνδεση με άλλους ανθρώπους στην αρχή, γίνονται ζεστοί, προσιτοί και φιλικοί τελικά. Είναι διασκεδαστικοί και πολύ αυθόρμητοι, γεγονός που τους καθιστά ιδανικό φίλο για την ετικέτα μαζί σε οποιαδήποτε δραστηριότητα, ανεξάρτητα από το αν προγραμματιστεί ή να είναι απρογραμμάτιστο. Οι ISFPs θέλουν να ζήσουν τη ζωή τους στο έπακρο και να αγκαλιάσουν το παρόν, ώστε να βεβαιωθούν ότι είναι πάντα έξω για να εξερευνήσουν νέα πράγματα και να ανακαλύψουν νέες εμπειρίες. Πιστεύουν ότι βρίσκουν σοφία, έτσι ώστε να δουν περισσότερη αξία στη συνάντηση με τους νέους ανθρώπους από άλλοι εσωστρεφείς.

4.3 Η Λύση του Προβλήματος

Σε αυτή την εργασία σκοπός μας είναι να προβλέψουμε την προσωπικότητα κάποιου συγκεκριμένου ατόμου χρησιμοποιώντας τις αναρτήσεις του στα social media μέσω του αλγορίθμου μηχανικής μάθησης. Για να γίνει πιο κατανοητός ο τρόπος επίλυσης ακολουθεί περιγραφή του κώδικα που χρησιμοποιήθηκε.

Αρχικά φαίνεται ότι υπάρχουν απλά 2 στήλες και 8675 δείγματα. Η πρώτη αντιπροσωπεύει τον τύπο προσωπικότητας και η 2η τα ποστ που έγιναν χωρισμένα με έναν ειδικό χαρακτήρα. Επίσης, υπενθυμίζεται πως οι τύποι προσωπικότητας είναι 20.

Σε αυτό το σημείο γίνεται η κωδικοποίηση των ετικετών μας, το οποίο προφανώς είναι αναγκαίο ώστε να περαστούν στον αλγόριθμο μας, μιας και οι αλγόριθμοι μηχανικής μάθησης δέχονται μόνο αριθμητικά και όχι λεκτικά δεδομένα. Χρησιμοποιήθηκε η τεχνική του one hot encoding η οποία είναι πιο αποδοτική για τις περιπτώσεις που έχουμε πολλαπλές κλάσεις όπως είναι και η δική μας.

Παρουσιάζεται ότι υπάρχουν 329090 διαφορετικές λέξεις, οι οποίες ακολούθως θα πρέπει να μετατραπούν σε αριθμούς, ώστε να μπορέσουν να περαστούν μέσα στον αλγόριθμο. Για αυτό το σκοπό, θα χρησιμοποιηθεί η τεχνική του lookup table.

Σε αυτό το σημείο είναι ευδιάκριτο ότι υπάρχει πολύ μεγάλη διαφορά στα βήματα του αλγορίθμου, αφού η μικρότερη εισαγωγή έχει 13 λέξεις ενώ η μεγαλύτερη έχει 9538 λέξεις. Επομένως, είναι πολύ σημαντικό να μετατρέψουμε τα ποστ σε ίδιου μήκους. Αυτό είναι ένα πολύ σημαντικό βήμα καθώς ο αλγόριθμος δεν δέχεται άνισα entries. Τα ποστ που έχουν πάνω από 13 λέξεις θα συμπληρωθούν με μηδενικά ενώ τα μεγαλύτερα από 500 λέξεις απλά θα κοπούν. Επιπλέον, αυτό γίνεται και για λόγους μείωσης των διαστάσεων καθώς δεν υπάρχει λόγος να υπερτερούν μεταξύ τους τόσο πολύ οι διαστάσεις σε οποιουδήποτε τύπου δεδομένα.

Ακολουθεί η RNN αρχιτεκτονική την οποία θα διαδεχτεί ο τομέας Embedding. Σε αυτό το σημείο πρέπει να προσθέσουμε ένα embedding layer, το οποίο χρειάζεται λόγω ύπαρξης 176000 λέξεων στο λεξιλόγιό μας.

Στη συνέχεια θα δημιουργήσουμε τα LSTM κελιά με σκοπό να χρησιμοποιηθούν στο νευρωνικό δίκτυο. Στην παρακάτω κλάση θα οριστούν define τα κελιά πάνω στον γράφο.

Έπειτα θα πρέπει να περαστούν τα δεδομένα απο το νευρωνικό δίκτυο, κάτι το οποίο είναι εφικτό με την χρήση της εντολής tf.nn.dynamic_rnn.

Επιπροσθέτως, δημιουργείται ένα initial state για το πέρασμα στο RNN. Αυτό το cell state περνάει μέσα απο τα hidden layers σε συνεχόμενα time steps. Το tf.nn.dynamic_rnn

στην ουσία κάνει όλη την διεργασία . Επιστρέφει κάθε στιγμή το βήμα και το final state του hidden layer.

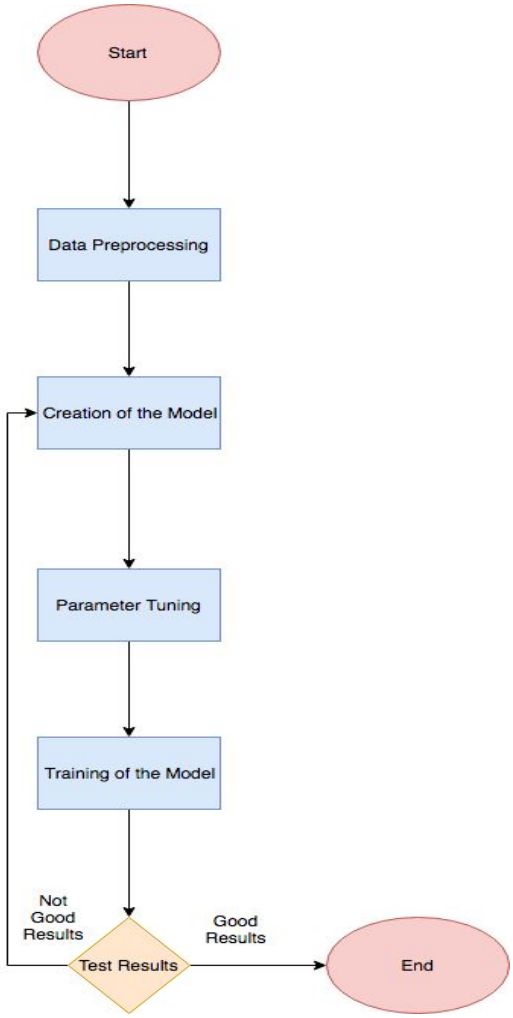
Όλα αυτά καταλήγουν στην τελική έξοδο η οποία θα χρησιμοποιηθεί για να επιτευχθεί η πρόβλεψη του τύπου προσωπικότητας (personality prediction). Επομένως θα πρέπει να πιαστεί το τελευταίο output με την χρήση του output[:,-1] και να υπολογιστεί το κόστος απο αυτό και τα labels.

Επιπλέον, προηγείται δημιουργία μερικών nodes στον γράφο για να υφίσταται και η ακρίβεια επικύρωσης (validation accuracy) την ώρα που εκπαιδευέται το νευρωνικό δίκτυο. Είναι πολύ σημαντικό η εκπαίδευση να γίνεται σε παρτίδες για υπολογιστικούς λόγους αλλά και για να μπορεί να υπάρχει πρόσβαση σε αποτελέσματα την ώρα εκπαίδευσης του νευρωνικού δικτύου.

Παρακάτω στο Παράρτημα που φαίνεται ο κώδικας μπορεί να εντοπίσει κανείς αναλυτικά όλα όσα περιγράφησαν. Χρησιμοποιείται η κλάση Model, και με την χρήση της μεθόδου fit με arguments τα δεδομένα εκπαίδευσης(train_x, train_y) και τα δεδομένα επικύρωσης(val_x, val_y) εκπαιδεύουν το μοντέλο. Τέλος, δημιουργήθηκε και η καμπύλη απώλειας του μοντέλου.

4.3.1 Διάγραμμα Ροής

Παρακάτω απεικονίζεται το οργανόγραμμα της εργασίας καθώς και το κώδικας που γράφτηκε για την λύση του προβλήματος.



Εικόνα 4. 1 Οργανόγραμμα εργασιών

Η Python είναι μια γλώσσα προγραμματισμού που επιλέχτηκε να χρησιμοποιηθεί γιατί χρησιμοποιεί απλό συντακτικό, έχει εξαιρετική αναγνωσιμότητα, φορητότητα και μοντέρνα χαρακτηριστικά που την κάνουν κατάλληλη ως πρώτη γλώσσα προγραμματισμού. Είναι μία γλώσσα υψηλού επιπέδου (άλλα παραδείγματα τέτοιων γλωσσών είναι η FORTRAN, C, Java κλπ). Ο κώδικας μία τέτοιας γλώσσας πρέπει να μετατραπεί σε γλώσσα μηχανής ώστε να εκτελεστεί από τον υπολογιστή. Η επεξεργασία αυτή γίνεται από διερμηνευτές και μεταγλωττιστές. Η γλώσσα αυτή γράφτηκε από τον Ολλανδό προγραμματιστή Guido van Rossum στα τέλη της δεκαετίας 1980-90. Η έκδοση 2.0 δημοσιεύτηκε στις 16 Οκτωβρίου 2000 και η έκδοση 3.0, η οποία δεν είναι, εν γένει, συμβατή με τις προηγούμενες εκδόσεις, στις 3 Δεκεμβρίου 2008. Αν ο υπολογιστής χρησιμοποιεί το λειτουργικό σύστημα Linux ή το MacOS τότε η Python είναι ήδη εγκατεστημένη.

Έγινε επιλογή της python σαν γλώσσα προγραμματισμού στη συγκεκριμένη εργασία για αρκετούς λόγους. Αρχικά, είναι γλώσσα ανοικτού κώδικα καθώς και είναι γενικής χρήσης αλλά και υψηλού επιπέδου σαν γλώσσα προγραμματισμού. Ακόμα χρησιμοποιεί απλό συντακτικό και διαθέτει εξαιρετική αναγνωσιμότητα με σημαντικές δυνατότητες προς διάφορες κατευθύνσεις. Τέλος κατά γενική ομολογία κρίνεται κατάλληλη τόσο για αρχάριους, όσο και για έμπειρους προγραμματιστές. Θεωρείται ιδιαίτερα αντικειμενοστραφής σαν γλώσσα. Υπάρχουν αρκετά πακέτα υποστήριξης.

4.3.2 Προεπεξεργασία των Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για το μοντέλο μας όπως προαναφέρθηκε είναι το σύνολο δεδομένων Myers Briggs από το Kaggle, η επεξεργασία των δεδομένων και η δημιουργία του μοντέλου έγιναν με την χρήση της γλώσσας προγραμματισμού Python και με την χρήση της προγραμματιστικής βιβλιοθήκης για Deep Learning, Tensorflow. Το σύνολο δεδομένων μας είχε 9859 δείγματα τα οποία χρησιμοποιήσαμε για την εκπαίδευση του μοντέλου μας. Το σύνολο δεδομένων συνίστατο από μια στήλη που περιγράφει τον τύπο της προσωπικότητας του ατόμου που έκανε τις δημοσιεύσεις (post) και μια στήλη στην οποία έχουμε όλες τις δημοσιεύσεις χωρισμένες με ειδικό χαρακτήρα όπως φαίνεται στην εικόνα 4.2 παρακάτω.

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Εικόνα 4. 2 Στήλη Προσωπικότητας και Στήλη Δημοσιεύσεων

Για το πρώτο μέρος έπρεπε να γίνει κάποια προεπεξεργασία δεδομένων καθώς οι δημοσιεύσεις είχαν πολλούς περιττούς χαρακτήρες οι οποίοι θα ήταν επιπλέον χαρακτηριστικά που θα έκανε το μοντέλο να καταναλώνει πολύ περισσότερη υπολογιστική

δύναμη κατά την διαδικασία της εκπαίδευσης. Αφαιρώντας έτσι όλα τα σημεία στίξης και τους συνδέσμους στις δημοσιεύεσεις μειώθηκαν οι διαστάσεις.

Μετά από αυτό το μέρος δημιουργήθηκε ένα λεξικό στο οποίο μεταμορφώθηκε κάθε λέξη σε έναν αριθμό χρησιμοποιώντας την τεχνική του πίνακα αναζήτησης (lookup table). Αυτό συνέβη επειδή θέλουμε να τροφοδοτήσουμε μόνο αριθμούς (οι ταξινομητές δέχονται μόνο αριθμητικά δεδομένα) στον ταξινομητή μας (classifier) και επομένως δεν μπορούμε να έχουμε τα δεδομένα μας ως λεκτικά. Για το τελευταίο μέρος δημιουργήθηκε ένα μήκος ακολουθίας 512 αφού το μέγεθος των post είναι άνισο και ο ταξινομητής δεν μπορεί να δεχτεί δεδομένα διαφορετικής διάστασης. Αυτό σημαίνει ότι κάθε post που είχε μήκος μικρότερο από 512 μετατράπηκε σε 512 συμπληρωμένα με μηδενικά στο τέλος της προεπεξεργασίας χωρίστηκαν οι δημοσιεύσεις σε ένα σετ εκπαίδευσης (training), σε ένα σετ δοκιμής(testing) αλλά και σε ένα σετ επικύρωσης (validation). Ο λόγος που έγινε αυτό αναγράφεται σε παραπάνω κεφάλαιο.

4.3.3 Δημιουργία του Μοντέλου

Για το τεχνικό κομμάτι του μοντέλου, δημιουργήθηκε ένα νευρωνικό δίκτυο Reccurent με ένα στρώμα ενσωμάτωσης όπως καθορίστηκε προηγουμένως. Η αρχιτεκτονική Word2Vec (στρώμα ενσωμάτωσης) σημαίνει ότι το λεξιλόγιο μας είναι διακριτό και θα μάθουμε έναν χάρτη που θα ενσωματώνει κάθε λέξη σε συνεχή διανυσματικό χώρο. Χρησιμοποιώντας αυτή τη διανυσματική απεικόνιση θα μας επιτραπεί να έχουμε μια συνεχή, κατανεμημένη εκπροσώπηση του λεξιλογίου μας. Αν για παράδειγμα, το σύνολο δεδομένων μας αποτελείται από n-λέξεις, μπορούμε τώρα να χρησιμοποιήσουμε τις συνεχείς λειτουργίες λέξεων για να δημιουργήσουμε μια κατανεμημένη αναπαράσταση των n-λέξεων. Στη διαδικασία εκμάθησης ενός γλωσσικού μοντέλου θα γίνει η εκμάθηση αυτής της λέξης ενσωματώνοντας την τον χάρτη.

Για την δημιουργία του μοντέλου, το πρώτο πράγμα που έπρεπε να γίνει ήταν να δημιουργηθούν οι placeholders (άδεια nodes που θα τοποθετηθούν τα δεδομένα κατά την διαδικασία της εκπαίδευσης) του γράφου. Μετέπειτα δημιουργήθηκε ο σκελετός του RNN χρησιμοποιώντας κελιά LSTM (Holland 1992) και Multi rnn για την δημιουργία του πολυστρωματικού μοντέλου.

Για την βελτιστοποίηση χρησιμοποιήθηκε ο Adam (Drenth 2017), καθώς είναι μία επέκταση του κλασσικού μοντέλου κατάβασης λόφου και επίσης δίνει αποτελέσματα αρκετά γρήγορα.

4.3.4 Αποτελέσματα

Φαίνεται σίγουρα (βλ. Εικόνα 4.3 Αποτελέσματα) ότι τα αποτελέσματα φαίνονται αρκετά περίεργα. Η πρώτη σκέψη είναι ότι μπορεί να έχουμε μάθει απέξω τα δεδομένα εκπαίδευσης μας (overfit) η ότι απλά το μοντέλο εξάγει εντελώς λανθασμένα αποτελέσματα. Για να σιγουρευτούμε μπορούμε εύκολα να ανατρέξουμε σε μία καμπύλη μάθησης (learning curve) ώστε να μπορούμε σίγουρα να αναγνωρίσουμε την πιθανότητα να ισχύει το δεύτερο ενδεχόμενο (λανθασμένα αποτελέσματα).

```
Epoch: 1/3 Iteration: 5 Train loss: 0.063
Epoch: 1/3 Iteration: 10 Train loss: 0.062
Epoch: 1/3 Iteration: 15 Train loss: 0.062
Epoch: 1/3 Iteration: 20 Train loss: 0.062
Epoch: 1/3 Iteration: 25 Train loss: 0.062
Val acc: 0.938
Epoch: 2/3 Iteration: 30 Train loss: 0.062
Epoch: 2/3 Iteration: 35 Train loss: 0.062
Epoch: 2/3 Iteration: 40 Train loss: 0.062
Epoch: 2/3 Iteration: 45 Train loss: 0.062
Epoch: 2/3 Iteration: 50 Train loss: 0.062
Val acc: 0.938
Epoch: 3/3 Iteration: 55 Train loss: 0.062
Epoch: 3/3 Iteration: 60 Train loss: 0.062
Epoch: 3/3 Iteration: 65 Train loss: 0.062
Epoch: 3/3 Iteration: 70 Train loss: 0.062
Epoch: 3/3 Iteration: 75 Train loss: 0.062
Val acc: 0.938
Epoch: 3/3 Iteration: 80 Train loss: 0.062
```

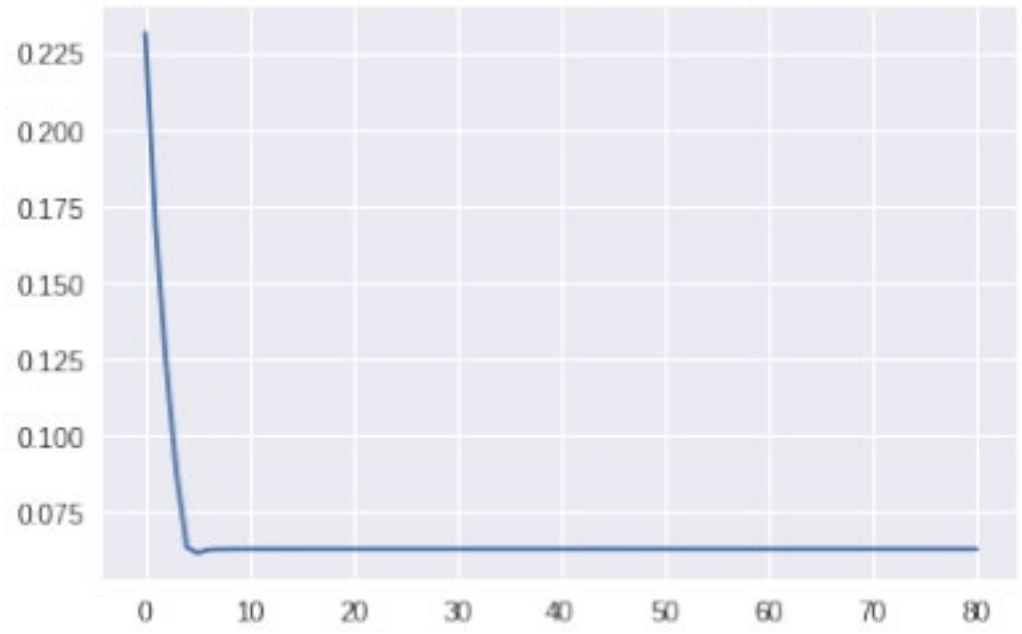
Εικόνα 4. 3 Αποτελέσματα

Ανάλυση Εικόνας 4.3:

Epoch: Είναι όταν ένα ολόκληρο σύνολο δεδομένων περνάει εμπρός και προς τα πίσω από το νευρωτικό δίκτυο μόνο μια φορά. Δεδομένου ότι ένα epoch είναι πολύ μεγάλο για να το τροφοδοτήσουμε στον υπολογιστή με την μια, το χωρίζουμε σε πολλές μικρότερες παρτίδες. Όπως φαίνεται και από τον πίνακα 4.3, τα δεδομένα μας προσπέλασαν αυτήν την διαδικασία τρεις φορές (3 running sessions).

Iteration: Είναι ένας γενικός όρος για τη λήψη κάθε στοιχείου από κάτι, το ένα μετά το άλλο. Κάθε φορά που χρησιμοποιείται ένας βρόγχος, για να περάσει από μια ομάδα στοιχείων, πρόκειται για επανάληψη. Όπως φαίνεται και απο τον πίνακα 4.3, είχαμε 80 επαναλήψεις.

Train loss: Δείχνει πόσο καλά το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης, ενώ η απώλεια επικύρωσης δείχνει πόσο καλά προσαρμόζεται στα νέα δεδομένα. Όπως φαίνεται και από τον πίνακα 4.3 κατά την 80η επανάληψη είχαμε 6.2% αστοχία προσαρμογής στα δεδομένα εκπαίδευσης (train loss).



Εικόνα 4. 4 Καμπύλη Εκμάθησης

Καμπύλη εκμάθησης καλείται η γραφική αναπαράσταση του μέσου ρυθμού μάθησης που διέπει το μοντέλο μας. Φαίνεται ξεκάθαρα ότι η σύγκλιση του μοντέλου μας έγινε κανονικά οπότε αποκλείουμε την δεύτερη περίπτωση που συζητήθηκε παραπάνω. Τώρα στόχος μας είναι να ελέγξουμε την πρώτη περίπτωση, δηλαδή αν το μοντέλο μας έμαθε απέξω τα δεδομένα εκπαίδευσης. Αυτό μπορεί να γίνει πολύ εύκολα χρησιμοποιώντας το test set πάνω στο ήδη εκπαιδευμένο μοντέλο μας. Αν τα αποτελέσματα μας είναι κοντά στο training

accuracy (εκπαιδευτική ακρίβεια μοντέλου) τότε σίγουρα δεν έχει μάθει απέξω τα δεδομένα εκπαίδευσης.

Παρατηρούμε πως όσο οι επαναλήψεις του κώδικα αυξάνονται (Άξονας χ'χ: Iterations), τόσο η αστοχία επικύρωσης μειώνεται (Άξονας ψ'ψ: Train Loss). Από 22.5% αστοχία προσαρμογής στα δεδομένα εκπαίδευσης (train loss), κατά τις πρώτες δοκιμές, καταλήξαμε να έχουμε λιγότερο από 7.5%, όπως απεικονίζεται στην καμπύλη εκμάθησης 4.4. Μετά την παραπάνω ενέργεια το μοντέλο μας έβγαλε ακρίβεια δοκιμής 0.93 η 93% (100% -7% Train Loss όπως απεικονίζεται στον πίνακα 4.4) το οποίο με την σειρά του σημαίνει ότι δεν έχουμε κάνει overfit επομένως τα αποτελέσματα μας είναι αντιπροσωπευτικά έως και αξιόπιστα. Overfit συμβαίνει όταν ένα μοντέλο μαθαίνει τις λεπτομέρειες και την διακύμανση (noise) στα δεδομένα εκπαίδευσης στο βαθμό που επηρεάζει αρνητικά την απόδοση του μοντέλου σε νέα δεδομένα.

5 Μελλοντική Εργασία

Σε αυτό το σημείο θα γίνει μια συζήτηση για το πως μπορεί να επεκταθεί αυτό το μοντέλο και επίσης πού μπορεί να χρησιμοποιηθεί στην αγορά, Παραπάνω όπως βλέπουμε λύσαμε το πρόβλημα της κατηγοριοποίησης το οποίο συζητήθηκε και στην αρχή αυτής της

εργασίας με αρκετά ικανοποιητικά αποτελέσματα. Το 93% accuracy στο testing set μας σημαίνει (σχεδόν αντιπροσωπευτικά) πως οποιαδήποτε καινούρια λεκτικά δεδομένα περαστούν στον αλγόριθμο μας έχουν μια πιθανότητα 93% να κατηγοριοποιηθούν με ακρίβεια. Παρακάτω θα παραθέσω μερικούς τρόπους με τους οποίους το μοντέλο αυτό θα μπορούσε να επεκταθεί και επίσης που θα μπορούσε να χρησιμοποιηθεί σε δύο διαφορετικά υποκεφάλαια της συζήτησης.

5.1 Χρήση του Μοντέλου (B2B deployment)

Το μοντέλο το οποίο δημιουργήθηκε θα μπορούσε εύκολα να εφαρμοστεί σε ένα τμήμα Ανθρώπινου Δυναμικού οποιασδήποτε εταιρείας. Οι περισσότερες εταιρείες πρωτού καλέσουν τους υποψήφιους εργαζόμενους για μια συνέντευξη συνήθως έχουν ανάγκη από μια εικόνα του ψυχογραφήματος τους. Πολλές φορές υπάρχουν ψυχολόγοι στις εκάστοτε εταιρείες οι οποίοι κάνουν ακριβώς αυτήν την δουλειά (κάποιες φορές παρέχουν και υποστήριξη στους ήδη εργαζομένους), και επομένως εύκολα θα μπορούσαν να αποκτήσουν πρόσβαση στα social media των υποψηφίων (καθώς το 95% έχουν ανοιχτά τα προφίλ τους με την θέληση τους) και με την χρήση εξειδικευμένων εργαλείων εξόρυξης δεδομένων (data mining tools) να εξάγουν τα λεκτικά τους δεδομένα εισάγοντας τα στο μοντέλο μας ώστε να αποκτήσουν αυτή την εικόνα της προσωπικότητας τους για την οποία συζητάμε. Με λίγα λόγια η εφαρμογή του μοντέλου μας απευθείας δημιουργεί επιπλέον κέρδος στην εταιρεία καθώς δεν θα χρειάζεται να πληρώνει κάποιον εξωτερικό ψυχολόγο ώστε να κάνει αυτήν την δουλειά την οποία το μοντέλο μας ίσως την κάνει καλύτερα και απο αυτόν.

Για μια πιο επιχειρηματικά προσανατολισμένη προσέγγιση θα μπορούσαμε να προσφέρουμε μια τέτοια υπηρεσία εύκολα μέσω ενός σέρβερ σε αρκετές εταιρείες που έχουν ανάγκη κάτι τέτοιο και έτσι να δημιουργήσουμε μια πλατφόρμα η οποία προσφέρει μια πιο φθηνή λύση σε προβλήματα εταιρειών που αναθέτουν αυτή τη δουλειά σε εξωτερικούς επαγγελματίες (individuals) ή πολλές φορές σε εταιρείες που προσφέρουν τέτοιου είδους υπηρεσίες σε τμήματα εύρεσης προσωπικού.

5.2 Τρόποι επέκτασης του Μοντέλου

Παρακάτω θα παρατεθούν κάποιοι τρόποι με τους οποίους το μοντέλο μας θα μπορούσε να μας δώσει καλύτερα αποτελέσματα.

1. **Μεγαλύτερο εύρος δεδομένων**, καθώς τα 9000 δείγματα δεν θεωρούνται αρκετά δεδομένα στον τομέα του βαθειάς εκμάθησης, βέβαια αυτό πολλές φορές εξαρτάται

και απο το πρόβλημα το οποίο καλούμαστε να λύσουμε. Στο συγκεκριμένο πρόβλημα περισσότερα δεδομένα σίγουρα θα οδηγούσαν σε καλύτερα αποτελέσματα

2. **Μεγαλύτερο και πιο πολύπλοκο μοντέλο,** στην περίπτωση μας το μέγεθος του μοντέλου μας έχει μόνο ένα LSTM layer και δεν χρειάζεται και παραπάνω επειδή όπως είδαμε τα αποτελέσματα μας εμφανίστηκαν σχεδόν κατευθείαν κατά την διάρκεια της εκπαίδευσης του μοντέλου μας. Στην περίπτωση που είχαμε περισσότερα δεδομένα της τάξεως των 2.000.000 ή παραπάνω δειγμάτων, τότε θα έπρεπε να γίνει σίγουρα και μια αρκετά εμφανή αλλαγή στο μέγεθος του μοντέλου μας.
3. **Χρήση φιλτραρισμένων δεδομένων,** αυτή είναι μια σχεδόν ακραία περίπτωση επέκτασης καθώς δεν χρειάζεται όταν έχουμε να κάνουμε με λίγα δεδομένα. Στην ουσία αυτό που θα μπορούσε να γίνει, είναι να γίνεται αποκλειστικά η χρήση δεδομένων τα οποία προσδίδουν ακρίβεια πρόβλεψης στα αποτελέσματα μας, υπάρχουν εξειδικευμένοι αλγόριθμοι οι οποίοι το κάνουν αυτό (feature selection field).

Επίσης καλό είναι σε αυτό το σημείο να τονιστεί και με ποιους άλλους τρόπους θα μπορούσε το ίδιο μοντέλο (εκπαιδευμένο με άλλα δεδομένα) να βοηθήσει ένα τμήμα Ανθρώπινου Δυναμικού. Παρακάτω θα παρατεθούν μερικά παραδείγματα.

Α) Παρακολούθηση και αξιολόγηση αιτούντων. Η παρακολούθηση και η αξιολόγηση των αιτούντων ήταν πάντα το πρώτο στην λίστα των εφαρμογών Μηχανικής εκμάθησης όσον αφορά το Ανθρώπινο Δυναμικό, ειδικά για εταιρείες και ρόλους που λαμβάνουν μεγάλους όγκους αιτούντων. Το Glint δεν είναι εταιρεία Τεχνητής Νοημοσύνης(AI), αλλά χρησιμοποιεί εργαλεία σχετικής τεχνολγίας AI για να βοηθήσει τις εταιρείες να εξοικονομήσουν χρήματα και να προσφέρουν καλύτερη εργασιακή εμπειρία. Τα εργαλεία μάθησης μηχανών βοηθούν το ανθρώπινο δυναμικό και το διευθυντικό προσωπικό να προσλαμβάνουν νέα μέλη της ομάδας παρακολουθώντας το ταξίδι ενός υποψήφιου σε όλη τη διαδικασία συνέντευξης και συμβάλλοντας στην επιτάχυνση της διαδικασίας εξ ορθολογισμού των σχολίων προς τους αιτούντες.

Β) Ο ανταγωνισμός για τους καλύτερους ανθρώπους έχει οδηγήσει πολλά τμήματα ανθρώπινου δυναμικού να χρησιμοποιούν εκτιμήσεις βασισμένες σε αλγόριθμους. Δεν αρκεί να ενεργήσουμε άμεσα σε πληροφορίες σχετικά με τα δεδομένα, αλλά να χρησιμοποιήσουμε αυτές τις πληροφορίες παράλληλα με την ερώτηση σχετικά με την οδήγηση, όπως: 1) πώς γίνεται να συνδεθούν τα χαρακτηριστικά των αιτούντων με τα επιχειρηματικά αποτελέσματα;

Γ) Ποια αποτελέσματα πρέπει να επικεντρωθούν κατά την πρόσληψη; και

Δ) Οι προβλέψεις (μίσθωση) να γίνονται με αμερόληπτο τρόπο.

Αυτό βέβαια είναι ένα σχεδόν υπερβολικό παράδειγμα επέκτασης του προαναγραφόμενου μοντέλου αλλά στην ουσία δεν είναι τίποτε άλλο από πολλά μοντέλα ίδιου τύπου που συνεργάζονται μεταξύ τους (ensembles) για να εξάγουν το επιθυμητό αποτέλεσμα.

6 Συμπεράσματα

Η προσέλκυση ταλέντων πριν από την πρόσληψη έχει επίσης δει μια ανάκαμψη στις εφαρμογές που βασίζονται στη μηχανική μάθηση τα τελευταία χρόνια. Ο Black, ο οποίος είναι ο ανώτερος διευθυντής της Οργανωτικής Ανάπτυξης του Glint, με την επωνυμία LinkedIn, ως παράδειγμα μιας εταιρείας που χρησιμοποιεί μία από τις πιο κοινές εκδόσεις μηχανικής μάθησης που συνιστούν θέσεις εργασίας. Άλλοι ιστότοποι εύρεσης εργασίας, όπως το Glassdoor και το Search, χρησιμοποιούν παρόμοιους αλγόριθμους για να χτίζουν χάρτες

αλληλεπίδρασης με βάση τα δεδομένα των χρηστών από προηγούμενες αναζητήσεις, συνδέσεις, αναρτήσεις και κλικ.

Το Phenom People είναι ένα παράδειγμα μιας σειράς εργαλείων που βασίζονται στη μηχανική μάθηση, τα οποία βοηθούν να οδηγήσουν τα ταλέντο στο χώρο της καριέρας μιας εταιρείας μέσω πολλαπλών καναλιών κοινωνικής δικτύωσης και αναζήτησης εργασίας. Ο Black σημειώνει ότι αυτό είναι πραγματικά ένα βήμα μετά από μια αναζήτηση λέξεων-κλειδιών, αν και ένα μεγάλο βήμα υπολογιστικά, καθώς υπάρχουν πολλά ακόμα πράγματα να γίνουν. Στην ουσία πάλι μπορούμε να δούμε ότι παρόμοια μοντέλο ανάλυσης λεκτικών αλλά και αριθμητικών χαρακτηριστικών χρησιμοποιούνται για την εξαγωγή που συζητάμε. Το μόνο που χρειάζεται στην κάθε περίπτωση είναι τα κατάλληλα δεδομένα. Η κατανόηση των ανθρώπων και ο λόγος για τον οποίο αποφασίζουν να παραμείνουν ή να εγκαταλείψουν μια θέση εργασίας είναι αναμφισβήτητα ένα από τα πιο σημαντικά ερωτήματα που πρέπει να απαντήσει η επιστήμη του Ανθρώπινου Δυναμικού. Ο εντοπισμός του κινδύνου φθοράς απαιτεί προηγμένη αναγνώριση προτύπων κατά την τοποθέτηση μιας σειράς μεταβλητών.

Ας φανταστεί κανείς μια υποθετική κατάσταση εντοπισμού συγκεκριμένων παραγόντων κινδύνου με βάση βαθμολογίες σε μια έρευνα εργαζομένων. Εάν ένας άνθρωπος προσπαθήσει να ανιχνεύσει τον κίνδυνο φθοράς μεταξύ των γυναικών μηχανικών σε μία εταιρεία με λιγότερο από 2 χρόνια θητείας, οι αναλύσεις διακύμανσης για την κατάληξη σε αυτό το συμπέρασμα είναι αναρίθμητες, όπως η εύρεση μιας βελόνας στα άχυρα. Η μηχανική μάθηση μας επιτρέπει να συνδέσουμε αυτές τις κουκίδες σε δευτερόλεπτα, απελευθερώνοντας εκπροσώπους του ανθρώπινου δυναμικού για να αφιερώσουν χρόνο στη στήριξη ομάδων αντί να αναλύσουν δεδομένα. Θα μπορούσε να προσφερθεί λύση και σε αυτό το κομμάτι απλά αντιμετωπίζοντας την καθημερινή φθορά των εργαζομένων ως ένα χρονικό πρόβλημα αναγνώρισης προτύπων. Στην συγκεκριμένη περίπτωση θα πρέπει να γίνουν βέβαια αρκετές μικρές αλλαγές στο μοντέλο καθώς τώρα πια το πρόβλημα θα πρέπει να αναχθεί σε χρονικό όπως προαναφέρθηκε. Παρόλα αυτά και πάλι αυτό το πρόβλημα θα μπορούσε να αντιμετωπιστεί έχοντας τα κατάλληλα δεδομένα και την τεχνογνωσία της διπλωματικής αυτής.

Στην συγκεκριμένη διατριβή χρησιμοποιήθηκε η μηχανική μάθηση για την αντιμετώπιση προβλήματος στον τομέα του Ανθρώπινου Δυναμικού. Το Ανθρώπινο Δυναμικό αποτελεί αναμφισβήτητα έναν τομέα που δίδεται ιδιαίτερη βαρύτητα στην παγκόσμια βιομηχανία. Ένα σοβαρό πρόβλημα που αντιμετωπίζει ο συγκεκριμένος κλάδος είναι ότι δεν έχει εκμεταλλευτεί επαρκώς τα πλεονεκτήματα της μηχανικής μάθησης για την επίλυση των προβλημάτων που αντιμετωπίζει με στόχο την αύξηση της αποδοτικότητάς του για αυτόν τον λόγο η παρούσα εργασία ασχολήθηκε με την αναγνώριση διαφόρων τύπων

προσωπικότητας χρησιμοποιώντας λεκτικά δεδομένα χρηστών των μέσων κοινωνικής δικτύωσης, (Facebook, Twitter).

Το μοντέλο που χρησιμοποιήθηκε για την επίλυση του προβλήματος αυτού είναι ένα νευρωνικό δίκτυο βαθιάς μάθησης. Η βασική διαφορά του μοντέλου που χρησιμοποιήθηκε έναντι των βασικών μοντέλων βαθιάς μάθησης είναι η αρχιτεκτονική συγκεκριμένου τύπου μνήμης νευρωνικού δικτύου η οποία και αποδείχθηκε ως κατά πολύ ανώτερη από τις κλασικές αρχιτεκτονικές ειδικά όταν πρέπει να αντιμετωπισθεί πρόβλημα αλληλουχιών όπως είδαμε στην συγκεκριμένη διατριβή μιας και στην εν λόγω περίπτωση η ανάλυση λεκτικού περιεχομένου είναι πρόβλημα αλληλουχίας. Τα αποτελέσματα του αλγορίθμου δείχνουν ιδιαίτερα υψηλά ποσοστά επιτυχίας, κάτι το οποία είναι ιδιαίτερα ενθαρρυντικό για την λειτουργία του αλγορίθμου και σίγουρα δίνει περιθώρια για περαιτέρω εξέλιξη.

7 Βιβλιογραφία

1)Bottou L. (2010), "Large-Scale Machine Learning with Stochastic Gradient Descent". Leon Bottou, Princeton.

2)Date C.J. (2000), "An Introduction to Database Systems", International Edition, (7th Edition), Addison-Wesley Publishing Company.

3)Drenth A.J. (2017), "The 16 Personality Types: Profiles, Theory, & Type Development". Andrew Drenth., ISBN: 978-0979216831

4)Ellis K., Gregory A., Mears-Young B.R. and Ragsdell G. (1995), "Critical Issues in Systems Theory and Practice". Springer US

5)Goodfellow Ian, Bengio Yoshua and Courville Aaron , (2016). "*Deep Learning*". MIT Press.

6)Holland, J.H., (1992), "Genetic Algorithms". Scientific American.

7)Klaus Greff. (2017)., "LSTM: A Search Space Odyssey".

8)Laudon K.C and Laudon J.P., (2009), "Essentials of Management Information Systems" (8th Edition). Prentice-Hall

9)McQuail Denis., (2017), "The Theorist of Mass Communication:, 1935-2017 - Social Science Space"

10)Singh H.S., (1997), "Data Warehousing: Concepts, Technologies, Implementation and Management",Upper Saddle River, NJ : Prentice-Hall

11)Tomas Mikolov. (2013), "Distributed Representations of Words and Phrases and their Compositionality".

12)Turban E. and Aronson J.E. (2007), "Business Intellligence and Decision Support Systems", (8th Edition), Upper Saddle River, NJ: Prentice Hall

13)Zadeh L.A. (1994), "Fuzzy Logic, Neural Networks and Soft Computing". Communications of the ACM, Vol.37, No.3.

14)Λι, Πίτερ Μ. (2012). «Κεφάλαιο 1». Μπεϋζιανή Στατιστική. Wiley. ISBN 978-1-1183-3257-3.

15)Ροβιθάκης, Γ. Α. (2007). Τεχνικές Βελτιστοποίησης. Θεσσαλονίκη: Εκδόσεις Τζιόλα.

16)Russell, Stuart; Norvig, Peter (1995). Artificial Intelligence : A Modern Approach. Englewood Cliffs: Prentice Hall. p. 578. ISBN 0-13-103805-2.

17)Berthold, Michael R.; Cebron, Nicolas; Dill, Fabian; Gabriel, Thomas R.; Kötter, Tobias; Meinl, Thorsten; Ohl, Peter; Thiel, Kilian; Wiswedel, Bernd (16 November 2009). "KNIME - the Konstanz information miner" (PDF). ACM SIGKDD Explorations Newsletter. 11 (1): 26. doi:10.1145/1656274.1656280. S2CID 408188.

18)Lämmel, R. (2008). "Google's Map Reduce programming model —Revisited". Science of Computer Programming.

19)Evans, Chris (October 2013). "Big data storage: Hadoop storage basics". computerweekly.com. Computer Weekly.

20)Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research, 12, 2121–2159. Retrieved from http://jmlr.org/papers/v12/duchi11a.html

21)Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V, … Ng, A. Y. (2012). Large Scale Distributed Deep Networks. NIPS 2012: Neural Information Processing

Systems, 1–11. http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf

22)Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543. http://doi.org/10.3115/v1/D14-1162

23)Kingma P. Diederik, Jimmy Ba Adam: A Method for Stochastic Optimization. 24) Diederik P. Kingma, OpenAI, 2014.

25) Data for the problem provided by Kaggle.

26)Neural Networks: All you Need to Know - Towards Data Science,

27)url: https://towardsdatascience.com/nns-aynk-c34efe37f15a

28) Sequence Classification with LSTM Recurrent Neural Networks,

29)Url:https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras

Παράρτημα Α –Διάγραμμα Ροής & Κώδικας

Ο Κώδικας

Personality type Predictor
Author: Georgiadis Vasilios.

Σε αυτή την εργασία σκοπός μας είναι να προβλέψουμε την προσωπικότητα κάποιου συγκεκριμένου ατόμου χρησιμοποιώντας τις αναρτήσεις του στα social media.

```python
# import dependencies
import pandas as pd
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt
```

```python
# read the csv file
import io

data = pd.read_csv("mbti_1.csv")
data[0:5]
```

```python
len(data)
```

8675

Φαίνεται ότι έχουμε απλά 2 στήλες και 8675 samples. Η πρώτη αντιπροσωπεύει τον τύπο προσωπικότητας και η 2η τα ποστ που έγιναν χωρισμένα με έναν ειδικό χαρακτήρα ("|||"). Επίσης οι τύποι προσωπικότητας είναι 20.

Data Preprocessing

```python
# let's do the easy part first. We'll do some magic over our labels to binarize them in order to feed them in our model.

mbti_dict = {0:'ENFJ',1:'ENFP',2:'ENTJ',3:'ENTP',4:'ESFJ',5:'ESFP',6:'ESTJ',7:'ESTP',8:'INFJ',9:'INFP',
            10:'INTJ',11:'INTP',12:'ISFJ',13:'ISFP',14:'ISFP',15:'ISTP'}

labels_clfs = data['type'].values.tolist()

labels = []
for per_type in data['type']:  # iter through the dictionary.
   for i in range(len(mbti_dict)):
       if per_type == mbti_dict[i]:
           labels.append(i)

labels_mat = np.zeros((len(labels),16))
for i in range(len(labels)):
   labels_mat[i][labels[i]] = 1

labels = labels_mat
print(labels[0:5])
```

```
[[ 0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.]
 [ 0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]]
```

Σε αυτό το σημείο βλέπουμε ότι ήδη έχει γίνει το encoding των label μας, το οποίο προφανώς είναι αναγκαίο ώστε να περαστούν στον αλγόριθμο μας, επειδή όπως έχει ειπωθεί παραπάνω στην εργασία οι αλγόριθμοι μηχανικής μάθησης δεν

δέχονται δεδομένα που είναι λεκτικά αλλά μόνο αριθμητικά. Χρησιμοποιήθηκε η τεχνική του one hot encoding η οποία είναι πιο αποδοτική για τις περιπτώσεις που έχουμε πολλαπλές κλάσεις όπως η δική μας.

```python
import re

def remove_punc(post):
   post = post.lower()

   post = post.replace('|||',"")

   post = re.sub(r'''(?i)\b((?:https?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s()<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()<>]+\)))*\)|[^\s`!()\[\]{};:'".,<>?«»“”‘’]))''', '', post, flags=re.MULTILINE)
    removables = ['@','#','$','%','^','&','*','(',')','-','_','+','=','{','}','[',']','|','\\','"',"'",';',':','<','>','/',',','.','?','!','\n']
   for punc in removables:
       post = post.replace(punc,'')

   post = re.sub( '\s+', ' ', post ).strip()
   return post
```

```python
posts = data.posts.tolist()
posts = [remove_punc(post) for post in posts]
```

```python
# Count total words
from collections import Counter

word_count = Counter()
for post in posts:
   word_count.update(post.split(" "))
```

```python

# Size of the vocabulary available to the RNN
vocab_len = len(word_count)
print(vocab_len)

print(len(posts[0]))
```

```
329090
3075
```

Βλέπουμε ότι έχουμε 329090 διαφορετικές λέξεις, το επόμενο βήμα είναι να τις μετατρέψουμε σε αριθμούς ώστε να μπορούμε να τις περάσουμε μέσα στον αλγόριθμο μας. Θα χρησιμοποιηθεί η τεχνική του lookup table.

```python
# Create a look up table
vocab = sorted(word_count, key=word_count.get, reverse=True)
# Create your dictionary that maps vocab words to integers here
vocab_to_int = {word: ii for ii, word in enumerate(vocab, 1)}

posts_ints=[]
for post in posts:
    posts_ints.append([vocab_to_int[word] for word in post.split()])

print(len(posts_ints[0]))
```

```
547
```

```python
non_zero_idx = [ii for ii, post in enumerate(posts_ints) if len(post) != 0]
len(non_zero_idx)

```

```
8675
```

```python
# generalize..
posts_ints = np.array([posts_ints[ii] for ii in non_zero_idx])
labels = np.array([labels[ii] for ii in non_zero_idx])
```

```python
posts_lens = Counter([len(x) for x in posts])
print("Review Length (zero): {}".format(posts_lens[0]))
print("Review Length (min): {}".format(min(posts_lens)))
print("Review Length (max): {}".format(max(posts_lens)))
```

```
Review Length (zero): 0
Review Length (min): 13
Review Length (max): 9538
```

Εδώ βλέπουμε ότι έχουμε πολύ μεγάλη διαφορά στα βήματά μας αφού το μικρότερο entry μας έχει 13 λέξεις ενώ το μεγαλύτερο έχει 9538 λέξεις. Σε αυτό το σημείο είναι πολύ σημαντικό να μετατραπούν τα ποστ μας σε ίδιου μήκους ποστ, Αυτό είναι ένα πολύ σημαντικό βήμα καθώς ο αλγόριθμος δεν δέχεται άνισα entries. Τα ποστ που έχουν πάνω απο 13 λέξεις θα συμπληρωθούν με μηδενικά ενώ τα μεγαλύτερα απο 500 λέξεις απλά θα κοπούν. Επίσης αυτό γίνεται και για λόγους μείωσης των διαστάσεων μας καθώς δεν υπάρχει λόγος να υπερτερούν μεταξύ τους τόσο πολύ οι διαστάσεις σε οποιουδήποτε τύπου δεδομένα.

```python
seq_len = 500
features = np.zeros((len(posts_ints),seq_len),dtype=int)
for i, row in enumerate(posts_ints):
    features[i, -len(row):] = np.array(row)[:seq_len]
print(features[:10])
```

```
[[    5   140  1288 ...,   366     6   649]
 [   23   739     3 ...,  1028    26     4]
```

```
    [   77    49   383 ...,    25   642    55]
    ...,
    [    1   248     2 ...,    11 87704  1098]
    [   23    21   120 ...,    18   262    14]
    [   11    20  1184 ...,  1865     2  3453]]
```

```python
from sklearn.model_selection import train_test_split

train_x, test_x, train_y, test_y = train_test_split(features, labels,
test_size=0.2)
test_x, val_x, test_y, val_y = train_test_split(test_x, test_y,
test_size=0.5)
train_x1, test_x1, train_y1, test_y1 = train_test_split(features,
labels_clfs, test_size=0.2)

print("Training set length: ",len(train_x))
print("Test set length: ", len(test_x))
print("Validation set length: ", len(val_x))
```

```
   Training set length:  6940
   Test set length:  867
   Validation set length:  868
```

```python
from sklearn.manifold import TSNE

#sample code for tsne, not working..
tsne = TSNE(n_components=2, random_state=0)

X_2d = tsne.fit_transform(train_x)


plt.figure(figsize=(6, 5))
for i, label in zip(range(16), train_y):
   plt.scatter(X_2d[train_y == i, 0], X_2d[train_y == i, 1], label=label)
plt.legend()
plt.show()
```

```
   ---------------------------------------------------------------------------
--

   IndexError                                Traceback (most recent call
last)

   <ipython-input-23-33eb8a2ad94a> in <module>()
         8 plt.figure(figsize=(6, 5))
         9 for i, label in zip(range(16), train_y):
   ---> 10     plt.scatter(X_2d[train_y == i, 0], X_2d[train_y == i, 1],
label=label)
        11 plt.legend()
        12 plt.show()


   IndexError: too many indices for array



   <Figure size 432x360 with 0 Axes>
```

RNN Architecture.

Παρακάτω θα γίνει η περιγραφή της αρχιτεκτονικής του νευρωνικού μας δικτύου.

```python
# Hyperparameters

lstm_size = 256
lstm_layers = 1
batch_size = 256
learning_rate = 0.001
```

Embedding

Σε αυτό το σημείο πρέπει να προσθέσουμε ένα embedding layer.Χρειάζεται να γίνει αυτό επειδή υπάρχουν 176000 λέξεις στο λεξιλόγιό μας.

LSTM cell

Μετά θα δημιουργήσουμε τα LSTM cells για να τα χρησιμοποιήσουμε στο νευρωνικό μας δίκτυο. Στην παρακάτω κλάση θα οριστούν define τα κελιά πάνω στον γράφο μας.

RNN forward pass

Τώρα πρέπει στην ουσία να κάνουμε το pass των δεδομένων μας απο το νευρωνικό μας δίκτυο. Μπορούμε εύκολα να το κάνουμε αυτό με την χρήση της εντολής tf.nn.dynamic_rnn.

```
outputs, final_state = tf.nn.dynamic_rnn(cell, inputs,
initial_state=initial_state)
```

Παραπάνω όπως βλέπετε δημιούργησα ένα initial state για το πέρασμα στο RNN. Αυτό το cell state παιρνάει μέσα απο τα hidden layers σε συνεχόμενα time steps. Το tf.nn.dynamic_rnn στην ουσία κάνει όλη την δουλειά για εμάς . Επιστρέφει κάθε στιγμή το βήμα και το final state του hidden layer.

Output

Μας Ενδιαφέρει μόνο το final output το οποίο θα χρησιμοποιήσουμε για να κάνουμε και το personality prediction μας. Οπότε πρέπει να πιάσουμε το τελευταίο output με την χρήση του output[:,-1] και να υπολογίσουμε το κόστος απο αυτό και τα labels_.

Validation accuracy

Δημιουργήθηκαν μερικά nodes στον γράφο για να έχουμε και το validation accuracy μας την ώρα που εκπαιδεύουμε το νευρωνικό μας δίκτυο.

Batching

Είναι πολύ σημαντικό η εκπαίδευση να γίνεται σε batches για υπολογιστικούς λόγους αλλά και για να μπορούμε να έχουμε πρόσβαση σε αποτελέσματα την ώρα που εκπαιδεύουμε το νευρωνικό μας δίκτυο.

Παρακάτω μπορείτε να δείτε όλα αυτά που αναφέρθηκαν γραμμένα. Χρησιμοποιούμε την κλάση Model, και με την χρήση της μεθόδου fit με arguments τα δεδομένα εκπαίδευσης(train_x, train_y) και τα δεδομένα επικύρωσης(val_x, val_y) εκπαιδεύουμε το μοντέλο μας.

```python
n_words = len(vocab_to_int)
embed_size = 300
# Create the graph object
graph = tf.Graph()
# Add nodes to the graph
```

```
with graph.as_default():
    inputs_ = tf.placeholder(tf.int32, [None, None], name='inputs')
    labels_ = tf.placeholder(tf.int32, [None, None], name='labels')
    keep_prob = tf.placeholder(tf.float32, name='keep_prob')
    embedding = tf.Variable(tf.random_uniform((n_words, embed_size), -1, 1))
    embed = tf.nn.embedding_lookup(embedding, inputs_)
```

```python
class Model:

    def get_batches(x, y, batch_size=100):

        n_batches = len(x)//batch_size
        x, y = x[:n_batches*batch_size], y[:n_batches*batch_size]
        for ii in range(0, len(x), batch_size):
            yield x[ii:ii+batch_size], y[ii:ii+batch_size]

    def fit(self, train_x, train_y, val_x, val_y):

        with graph.as_default():
            lstm = tf.contrib.rnn.BasicLSTMCell(lstm_size)

            dropout = tf.contrib.rnn.DropoutWrapper(lstm,
output_keep_prob=keep_prob)

            multi_cells = tf.contrib.rnn.MultiRNNCell([dropout] *
lstm_layers)

            init = multi_cells.zero_state(batch_size, tf.float32)

            outputs, final_state = tf.nn.dynamic_rnn(multi_cells, embed,
initial_state=init)

            predictions = tf.contrib.layers.fully_connected(outputs[:, -1],
1, activation_fn=tf.sigmoid)
            cost = tf.losses.mean_squared_error(labels_, predictions)

            optimizer = tf.train.AdamOptimizer(learning_rate).minimize(cost)

            correct_pred = tf.equal(tf.cast(tf.round(predictions), tf.int32),
labels_)
            accuracy = tf.reduce_mean(tf.cast(correct_pred, tf.float32))
```

```
        epochs = 1

        with graph.as_default():
            saver = tf.train.Saver()

        loss_arr = []
        with tf.Session(graph=graph) as sess:
            sess.run(tf.global_variables_initializer())
            iteration = 1
            for e in range(epochs):
                state = sess.run(init)

                for ii, (x, y) in enumerate(get_batches(train_x, train_y,
batch_size), 1):
                    feed = {inputs_: x,
                            labels_: y,
                            keep_prob: 0.5,
                            init: state}
                    loss, state, _ = sess.run([cost, final_state, optimizer],
feed_dict=feed)
                    loss_arr.append(loss)


                    if iteration%5==0:
                        print("Epoch: {}/{}".format(e+1, epochs),
                              "Iteration: {}".format(iteration),
                              "Train loss: {:.3f}".format(loss))

                    if iteration%25==0:
                        val_acc = []
                        val_state =
sess.run(multi_cells.zero_state(batch_size, tf.float32))
                        for x, y in get_batches(val_x, val_y, batch_size):
                            feed = {inputs_: x,
                                    labels_: y,
                                    keep_prob: 1,
                                    init: val_state}
                            batch_acc, val_state = sess.run([accuracy,
final_state], feed_dict=feed)
                            val_acc.append(batch_acc)
                        print("Val acc: {:.3f}".format(np.mean(val_acc)))
                    iteration +=1
            saver.save(sess, "checkpoints/sentiment.ckpt")
```

```
            return loss_arr
```

```
model = Model()
loss = model.fit(train_x, train_y, val_x, val_y)
```

```
   Epoch: 1/1 Iteration: 5 Train loss: 0.071
   Epoch: 1/1 Iteration: 10 Train loss: 0.062
   Epoch: 1/1 Iteration: 15 Train loss: 0.062
   Epoch: 1/1 Iteration: 20 Train loss: 0.062
   Epoch: 1/1 Iteration: 25 Train loss: 0.062
   Val acc: 0.938
```

Loss Curve

Παρακάτω δημιουργήθηκε και το loss curve του μοντέλου μας.

```
%matplotlib inline
plt.plot(loss)
plt.show()
```

Benchmarks using older ML algorithms.

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier

clf = RandomForestClassifier().fit(train_x1, train_y1)
clf_1 = GaussianNB().fit(train_x1, train_y1)
clf_2 = DecisionTreeClassifier().fit(train_x1, train_y1)


score = clf.score(test_x1, test_y1)
score_1 = clf_1.score(test_x1, test_y1)
score_2 = clf_2.score(test_x1, test_y1)
print(score)
print(score_1)
```

```
print(score_2)



test_acc = []
with tf.Session(graph=graph) as sess:
    saver = tf.train.Saver()
    saver.restore(sess, tf.train.latest_checkpoint('checkpoints'))
    test_state = sess.run(multi_cells.zero_state(batch_size, tf.float32))
    for ii, (x, y) in enumerate(get_batches(test_x, test_y, batch_size), 1):
        feed = {inputs_: x,
                labels_: y,
                keep_prob: 1,
                init: test_state}
        batch_acc, test_state = sess.run([accuracy, final_state],
feed_dict=feed)
        test_acc.append(batch_acc)
    print("Test accuracy: {:.3f}".format(np.mean(test_acc)))
```