



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών &
Μηχανικών Γυπολογιστών
Τομέας Ηλεκτρονικής και Γυπολογιστών

Διπλωματική Εργασία

Ανάπτυξη Πολυτροπικού Συστήματος Ανάλυσης Συναισθήματος βάσει Κειμένου και Εικόνας

Εκπόνηση:

Γερογιάννης Κωνσταντίνος
ΑΕΜ: 9638

Επίβλεψη:

Καθ. Συμεωνίδης Ανδρέας
Υπ. Δρ. Νάστος
Δημήτριος-Νικήτας

Θεσσαλονίκη, Νοέμβριος 2023

*A computer would deserve to be called intelligent if it could deceive a human into believing
that it was human.*

— Alan Turing, *Hey Beau*

ΕΥΧΑΡΙΣΤΙΕΣ

Καταρχάς, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Νάστο Δημήτρη για την εξαιρετική συνεργασία μας στην εκπόνηση της διπλωματικής εργασίας, την αφοσίωσή του με την παροχή βοήθειας όπου την χρειάστηκα και την εμπιστοσύνη που μου έδειξε.

Θα ήθελα επίσης να ευχαριστήσω τον κύριο επιβλέποντα της διπλωματικής εργασίας και leader της ερευνητικής ομάδας ISSEL, τον καθηγητή Συμεωνίδη Ανδρέα, για την βοήθεια και την υποστήριξή του όπου χρειάστηκε. Επιπροσθέτως θέλω να αποδώσω τις ευχαριστίες μου στον υποψήφιο διδάκτορα του τμήματος Αλαγιαλόγλου Λεωνίδα για τις συμβουλές του σχετικά με τη διπλωματική εργασία.

Η οικογένεια μου σίγουρα αξίζει να βρίσκεται στη παρούσα σελίδα, που χάρη σε αυτήν βρίσκομαι εδώ που βρίσκομαι με τη στήριξή της και την αγάπη της στο παρελθόν, στο παρόν και στο μέλλον που έρχεται.

Και δε μπορώ να παραλείψω φυσικά τους φίλους μου για όλες αυτές τις εμπειρίες και στιγμές που ζήσαμε και συνέβαλλαν και αυτές με τον τρόπο τους στην μέχρι τώρα πορεία της επαγγελματικής και κοινωνικής μου ζωής.

Ευχαριστώ.

Περίληψη

Η Τεχνητή Νοημοσύνη (AI) έχει κεντρίσει το ενδιαφέρον της ερευνητικής κοινότητας και έχει καταφέρει να εισαχθεί σε τομείς και εφαρμογές που ίσως πριν κάποια χρόνια η πλειοφηφία των ανθρώπων να μη μπορούσε να φανταστεί, με απότερο σκοπό τη διευκόλυνση της ανθρώπινης καθημερινότητας, παρέχοντας εφαρμογές που μπορούν να φανούν χρήσιμες σε κάθε ανθρώπινη δραστηριότητα. Η βαθιά μάθηση, η επεξεργασία φυσικής γλώσσας και η υπολογιστική όραση αποτελούν κλάδους της τεχνητής νοημοσύνης που παρουσιάζουν τεράστιο ενδιαφέρον, καθώς αποσκοπούν στην αμεσότερη επικοινωνία μεταξύ του ανθρώπου και κάθε είδους υπολογιστή και στην κατανόηση περιεχομένου εικόνας από τον υπολογιστή.

Η εργασία αυτή επικεντρώνεται στην ανάλυση συναισθημάτων. Ως ανάλυση συναισθημάτων αναφερόμαστε στην εξαγωγή συναισθήματος από μία ψηφιακή είσοδο. Σε ένα πολυτροπικό σύστημα ανάλυσης συναισθημάτων η είσοδος αποτελείται από 2 ή περισσότερους τύπους δεδομένων π.χ. κείμενο, εικόνα, βίντεο, ήχος. Το πολυτροπικό σύστημα που υλοποιείται κατηγοριοποιεί την είσοδο στις κατηγορίες του αρνητικού, ουδέτερου ή θετικού συναισθήματος. Ελέγχονται μέθοδοι και μοντέλα επεξεργασίας κειμένου και εικόνας, ώστε να εξαχθούν συμπεράσματα σχετικά με το συναίσθημα του κάθε τύπου εισόδου. Επιπλέον, διερευνούνται μέθοδοι συνδυασμού των αποτελεσμάτων της επεξεργασίας των δύο διαφορετικών εισόδων με σκοπό τη βελτίωση των συμπερασμάτων. Ακόμη ελέγχονται οι διάφοροι μέθοδοι προεπεξεργασίας κειμένου και εικόνας.

Επιπροσθέτως ερευνάται η ικανότητα των μοντέλων να γενικευθούν σε νέα σύνολα δεδομένων εκτός από αυτά που εκπαιδεύτηκαν, καθώς και η δυνατότητα εντοπισμού του συναισθήματος με τη χρήση κατάλληλων πολυγλωσσικών μοντέλων, σε κείμενα γλωσσών στις οποίες δεν έχει προσαρμοσθεί το μοντέλο. Τέλος, δημιουργείται μια ιστοσελίδα στην οποία ο χρήστης μπορεί να εισάγει κείμενο και εικόνα που εκείνος επιθυμεί και το μοντέλο να εντοπίσει το συναίσθημα της εισόδου.

Κωνσταντίνος Γερογιάννης

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Ελλάδα

Οκτώβριος 2023

Title

Multimodal Sentiment Analysis System based on Text and Image

Abstract

Artificial Intelligence (AI) has piqued the interest of the research community and has managed to penetrate into sectors and applications that perhaps a few years ago the majority of people could not have imagined, with the ultimate goal of facilitating everyday human activities, providing applications that can be useful in every human activity. Deep learning, natural language processing, and computer vision are branches of artificial intelligence that are of great interest, as they aim at more direct communication between humans and all kinds of computers and in the understanding of image content by the computer.

This work focuses on sentiment analysis. Sentiment analysis refers to the extraction of emotion from a digital input. In a multimodal sentiment analysis system, the input consists of 2 or more types of data, e.g., text, image, video and sound. The implemented multimodal system categorizes the input into categories of negative, neutral, or positive sentiment. Methods and models for processing texts and images are tested to draw conclusions about the sentiment of each input type. In addition, methods for combining the results of processing the two different inputs are investigated with the aim of improving the conclusions. Various methods of preprocessing text and images are also examined.

Additionally, the ability of models to generalize to new data sets apart from those they were trained on is researched, as well as the ability to detect sentiment using appropriate multilingual models in texts of languages to which the model has not been fine-tuned. Finally, a website is created where the user can enter texts and images of their choice, and the system identifies the sentiment of the input.

Konstantinos Gerogiannis
Electrical & Computer Engineering Department,
Aristotle University of Thessaloniki, Greece
October 2023

Περιεχόμενα

Ευχαριστίες	iii
Περίληψη	v
Abstract	vii
Ακρωνύμια	xiv
1 Εισαγωγή	1
1.1 Περιγραφή του Προβλήματος	1
1.2 Σκοπός - Συνεισφορά της Διπλωματικής Εργασίας	2
1.3 Διάρθρωση της Αναφοράς	3
1.4 Τεχνητή Νοημοσύνη	5
2 Ανάλυση συναισθήματος βάσει κειμένου	9
2.1 Κατανόηση βασικών όρων	10
2.2 Προεπεξεργασία κειμένου	13
2.3 Datasets ανάλυσης συναισθήματος βάσει κειμένου	14
2.4 Γλωσσικά Μοντέλα/ Μοντέλα Κειμένου	15
2.5 Εργαλεία κειμένου	18
3 Ανάλυση συναισθήματος βάσει εικόνας	21
3.1 Προεπεξεργασία εικόνας	21
3.2 Datasets εικόνας	22
3.3 Μοντέλα υπολογιστικής όρασης/ Μοντέλα Εικόνας	23
4 Πολυτροπικό σύστημα ανάλυσης συναισθημάτων	35
4.1 Συνένωση χαρακτηριστικών	35
4.2 Πολυτροπικά σύνολα δεδομένων	38
5 Γλοποιήσεις	41
5.1 Λογισμικό	41
5.2 Επιλεγμένο Dataset: MVSA Single	46
5.3 Υπερπαράμετροι	47
5.4 Πολυτροπικό σύστημα ανάλυσης συναισθήματος κειμένου-εικόνας	50
6 Αποτελέσματα Πειραμάτων	59
6.1 Πίνακες αποτελεσμάτων πειραμάτων	60
6.2 Πειραματισμός με την προεπεξεργασία των εικόνων	65

ΠΕΡΙΕΧΟΜΕΝΑ

7 Επεκτάσεις	69
7.1 Δοκιμή του συστήματος σε διαφορετικές γλώσσες	69
7.2 Πειραματισμός με το ελληνικό bert	71
7.3 Δημιουργία συνόλου δεδομένων	72
7.4 Δημιουργία ιστοσελίδας	74
8 Συμπεράσματα	79
8.1 Γενικά συμπεράσματα	79
8.2 Προβλήματα	80
9 Μελλοντικές επεκτάσεις	83
Βιβλιογραφία	85

Κατάλογος Σχημάτων

1.1	Κλάδοι της επιστήμης της Τεχνητής Νοημοσύνης	5
1.2	Χωρισμός σε σύνολα εκπαίδευσης/επικύρωσης/ελέγχου	8
2.1	Η εξέλιξη στην ανάλυση συναισθήματος βάσει κειμένου, από τα λεξικά στη μεταφερόμενη μάθηση.	10
2.2	Η αρχιτεκτονική του transformer μοντέλου. Αριστερά φαίνεται το κομμάτι του κωδικοποιητή και δεξιά ο αποκωδικοποιητής.	11
2.3	Στοίβα από κωδικοποιητές-αποκωδικοποιητές	11
2.4	Αναπαράσταση της εισόδου στο BERT μοντέλο	16
3.1	Συνέλιξη με πίνακα για μέγεθος εικόνας MxNx3 και πυρήνα 3x3x3 .	24
3.2	Συγκέντρωση με πίνακα εισόδου 4x4 και μάσκα 2x2	24
3.3	Ταξινόμηση με CNN 2 στρωμάτων συνέλιξης	25
3.4	Αριστερά: VGG-19 (παλιότερο μοντέλο), κέντρο: Αρχιτεκτονική ResNet χωρίς προσπέραση συνδέσεων, δεξιά: Αρχιτεκτονική ResNet	26
3.5	Αριστερά: Μπλοκ του ResNet, δεξιά: μπλοκ του ResNeXt	27
3.6	Πυκνό Μπλοκ 5 στρωμάτων	27
3.7	Αρχιτεκτονική του DenseNet με 3 πυκνά μπλοκ	27
3.8	Αρχιτεκτονική του EfficientNet-B0	28
3.9	MBConv μπλοκ	28
3.10	Χωρισμός της εικόνας σε patches	29
3.11	Κωδικοποιητής του transformer εικόνας	30
3.12	Διαδικασία προ-εκμάθησης του BEiT	31
4.1	Early Fusion	36
4.2	Late Fusion	36
4.3	Συνδυασμός των modalities	37
4.4	Tensor Fusion	38
4.5	Attention Fusion	38
5.1	Σύγκριση πειραμάτων με το εργαλείο Comet ML	43
5.2	Πορεία εκμάθησης για διαφορετικούς ρυθμούς εκμάθησης	48
5.3	Εφαρμογή dropout 0.5 στο κρυφό στρώμα	49
5.4	Πολυτροπικό σύστημα ανάλυσης συναισθήματος κειμένου και εικόνας .	50
5.5	Πολυτροπικό σύστημα ανάλυσης συναισθήματος κειμένου και εικόνας με τη προσθήκη του εργαλείου Vader	51
5.6	Βασικό μοντέλο κειμένου/εικόνας για την διαδικασία της εκπαίδευσης	53

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

6.1	Fine-tuning των σημαντικότερων μοντέλων κειμένου για 6 εποχές	64
6.2	Fine-tuning των σημαντικότερων μοντέλων εικόνας για 6 εποχές	65
6.3	Εκπαίδευση των αρχιτεκτονικών συνένωσης για 25 εποχές	66
7.1	Κείμενο: "...οι μόνοι που έχουμε το προνόμιο να λέμε τον ουρανό "ουρανό" και την θάλασσα "θάλασσα", όπως την έλεγαν ο Ομηρος και ο Πλάτωνας πριν από δίαμισυ χιλιάδες χρόνια"..., Οδυσσέας Ελύτης, Τελευταία δύση του Αυγούστου	73
7.2	Βασική σελίδα της ιστοσελίδας. Ο χρήστης εισάγει το κείμενο και την εικόνα που επιθυμεί.	76
7.3	Τα αποτελέσματα της ανάλυσης συναισθήματος στα δεδομένα εισόδου	76
7.4	Γραφήματα με τις πιθανότητες της κάθε πρόβλεψης	77
7.5	Αποτελέσματα ανάλυσης συναισθήματος σε είσοδο κειμένου, απουσία εικόνας	77

Κατάλογος πινάκων

5.1	Βιβλιοθήκες της Python που χρησιμοποιήθηκαν	45
5.2	Προεπεξεργασία MVSA-Single	46
5.3	Προεπεξεργασία MVSA-Multiple	47
5.4	Υπερπαράμετροι μοντέλου κειμένου	53
5.5	Υπερπαράμετροι μοντέλου εικόνας	55
5.6	Υπερπαράμετροι μοντέλου συνένωσης	56
6.1	Σύγκριση του προτεινόμενου συστήματος με προηγούμενες δημοσιεύσεις πάνω στο ίδιο σύνολο δεδομένων	60
6.2	Υπερπαράμετροι και μετρικές μοντέλων κειμένου (BERT-ALBERT)	61
6.3	Υπερπαράμετροι και μετρικές μοντέλων κειμένου (ROBERTA-DEBERTA)	62
6.4	Υπερπαράμετροι και μετρικές των RESNET μοντέλων εικόνας	63
6.5	Υπερπαράμετροι και μετρικές των DENSENET μοντέλων εικόνας	63
6.6	Υπερπαράμετροι και μετρικές των EFFICIENTNET μοντέλων εικόνας	63
6.7	Υπερπαράμετροι και μετρικές των transformer μοντέλων εικόνας	64
6.8	Υπερπαράμετροι και μετρικές των μοντέλων συνένωσης	65
6.9	Πειράματα διαφορετικών μεθόδων προεπεξεργασίας των εικόνων	66
7.1	Εκπαίδευση στα αγγλικά κείμενα, πρόβλεψη σε όλες τις γλώσσες	70
7.2	Εκπαίδευση στα ελληνικά κείμενα, πρόβλεψη σε όλες τις γλώσσες	70
7.3	Εκπαίδευση και στα αγγλικά και στα ελληνικά κείμενα, πρόβλεψη σε όλες τις γλώσσες	71
7.4	Σύγκριση των μοντέλων κειμένου bert και greek bert	71
7.5	Συγκρούσεις στα labels μεταξύ των 2 annotators	72
7.6	Παράδειγμα σύγκρουσης απόφεων μεταξύ των δύο annotators	73
7.7	Χρήση τρίτου annotator	74
7.8	Εξαίρεση των δεδομένων που προκαλούν συγκρούσεις από το σετ δεδομένων	74
7.9	Συμπλήρωση των δεδομένων που προκαλούν συγκρούσεις όπως στο MVSA-Single	74
7.10	Βιβλιοθήκες της Python που χρησιμοποιήθηκαν για την υλοποίηση της ιστοσελίδας	75

Ακρωνύμια Εγγράφου

Παρακάτω παρατίθεται η λίστα με τα ακρωνύμια που χρησιμοποιούνται στην παρούσα διπλωματική εργασία:

AI	→ Artificial Intelligence
NLP	→ Natural Language Processing
SA	→ Sentiment Analysis
LLM	→ Large Language Model
CNN	→ Convolutional Neural Network
GPU	→ Graphics Processing Unit
CPU	→ Central Processing Unit
RAM	→ Random Access Memory

1

Εισαγωγή

Το κεφάλαιο αυτό είναι εισαγωγικό και επιχειρεί να εντάξει τον αναγνώστη στο θέμα που καλείται να αντιμετωπίσει η διπλωματική εργασία. Αρχικά γίνεται μια προσπάθεια αποσαφήνισης του τίτλου της, περιγράφοντας αναλυτικά τη σημασία κάθε όρου του. Έπειτα επισημαίνονται περιπτώσεις που μπορεί να συνεισφέρει η εργασία και παρουσιάζεται η διάρθρωση των κεφαλαίων που ακολουθούν. Στο τέλος του κεφαλαίου εισάγεται ο όρος της τεχνητής νοημοσύνης, συνοδευόμενος με βασικές πληροφορίες που η γνώση τους θα χρειαστεί στα κεφάλαια που ακολουθούν.

1.1 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Η παρούσα διπλωματική πραγματεύεται τη δημιουργία ενός πολυτροπικού συστήματος εντοπισμού συναισθήματος, οπότε και είναι σημαντικό να προσδιοριστεί η έννοια του συναισθήματος, το οποίο σαν όρος είναι γενικός και αόριστος. Αναζητώντας τον όρο *Sentiment* σε διάφορα λεξικά, βρίσκει κανείς τους ακόλουθους ορισμούς:

- Η σκέψη, η άποψη, η ιδέα, η οποία βασίζεται στην αίσθηση/στον τρόπο σκέψης απέναντι σε μία κατάσταση. (Cambridge Dictionary)
- Η γενική αίσθηση, στάση ή άποψη απέναντι σε κάτι. (Cambridge Dictionary)
- Η στάση, σκέψη ή κρίση που καθορίζεται από την αίσθηση. (Merriam-Webster dictionary)
- Μια συγκεκριμένη όψη ή αντίληψη (Merriam-Webster dictionary)

Παρόλο που οι παραπάνω ορισμοί μπορεί να μην είναι ξεκάθαροι και να μη συμβαδίζουν απόλυτα, τοποθετούν τον όρο σε ένα γενικό πλαίσιο. Το κύριο όμως

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

πρόβλημα είναι η σύγχυση μεταξύ των όρων emotion και sentiment, οι οποίοι αν μεταφραστούν αποδίδονται στην ελληνική λέξη συναίσθημα.

Ο όρος emotion χρησιμοποιείται για να αναφερθεί στην εσωτερική ψυχολογία του ατόμου και περιλαμβάνει τις αντιδράσεις του σε διάφορες καταστάσεις ή γεγονότα. Περιγράφει την υποκειμενική κρίση του ατόμου. Σαν όρος αναφέρεται σε συναίσθημα πλούσια σε ποικιλία και ένταση, όπως είναι η χαρά, η λύπη, ο θαυμασμός, ο φόβος, ο θυμός.

Από την άλλη μεριά, ο όρος sentiment χρησιμοποιείται για να περιγράψει απόψεις και εκφράσεις για ένα συγκεκριμένο θέμα, αντικείμενο ή κατάσταση. Με το sentiment περιορίζεται το εύρος των διαφορετικών κατηγορίων που κρύβονται στον όρο emotion, κυρίως στις κατηγορίες του θετικού, του αρνητικού και του ουδέτερου, απλοποιώντας έτσι τον χώρο κατηγοριοποίησης του συναίσθηματος. Όταν γίνεται αναφορά στον όρο sentiment, η κατηγορία που αποδίδεται είναι η αντικειμενικότερη δυνατή που επικρατεί σαν αντίληψη στο ευρύτερο κοινωνικό σύνολο, χωρίς να βασίζεται στην άποψη ενός συγκεκριμένου ατόμου, επηρεασμένη από τις προσωπικές του εμπειρίες και βιώματα.

Από εδώ και στο εξής, όποτε χρησιμοποιείται η λέξη συναίσθημα θα αντιστοιχίζεται στην αγγλική λέξη sentiment. Στη διπλωματική εργασία τίθεται ως στόχος η ανάλυση συναίσθηματος. Η Ανάλυση Συναίσθηματος (Sentiment Analysis), που στη βιβλιογραφία ονομάζεται και Εξόρυξη Άποψης (Opinion Mining), αναφέρεται στον εντοπισμό του συναίσθηματος από τα δεδομένα. Η ανάλυση συναίσθηματος αρχικά εφαρμόσθηκε σε δεδομένα κειμένου, μετεξελίχθηκε όμως και σε διαφορετικές μορφές, όπως είναι η εικόνα, το βίντεο και ο ήχος. Βρίσκει πολλές εφαρμογές, αφού μπορεί να χρησιμοποιηθεί στις επιχειρήσεις για την εξαγωγή της γνώμης των καταγαλωτών για ένα προσφερόμενο προϊόν ή υπηρεσία, στα μέσα κοινωνικής δικτύωσης για τον εντοπισμό της κοινής γνώμης για ένα γεγονός και στην εξυπηρέτηση πελατών για την αποτελεσματικότητα των χρησιμοποιούμενων μεθόδων. Για την κατασκευή ενός τέτοιου συστήματος που καταφέρνει να αναγνωρίζει συναίσθηματα, μπορούν να ακολουθηθούν προσεγγίσεις που βασίζονται σε κανόνες ή μπορούν να χρησιμοποιηθούν AI αλγόριθμοι και τεχνικές. Ωστόσο, αυτά τα συστήματα καλούνται να αντιμετωπίσουν μορφές της ανθρώπινης επικοινωνίας που δεν είναι ξεκάθαρες, όπως η ειρωνεία, ο σαρκασμός και η χρήση λέξεων που κατέχουν διαφορετική ερμηνεία ανάλογα με το ευρύτερο περιεχόμενο. Επιπλέον καλούνται να αναγνωρίσουν φράσεις της γλώσσας και εκφράσεις του προσώπου που να διαφέρουν από κοινωνία σε κοινωνία. Ένα παράδειγμα που δείχνει τις διαφορές στο πως αντιλαμβάνονται διαφορετικές κουλτούρες τις εκφράσεις του προσώπου για συγκεκριμένα συναίσθηματα δίνεται στο [1], όπου φαίνεται η διαφορετική αντίληψη του κάθε συναίσθηματος μεταξύ των ανθρώπων που ζουν στον δυτικό Καύκασο και στην ανατολική Ασία.

1.2 ΣΚΟΠΟΣ - ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η παρούσα διπλωματική εργασία στοχεύει στη δημιουργία ενός συστήματος ανάλυσης συναίσθημάτων. Για να μπορέσουν να καλυφθούν με τον καλύτερο δυνατό τρόπο οι απαιτήσεις αναγνώρισης συναίσθηματος, αποπειράται ο σχεδιασμός

1.3. ΔΙΑΡΘΡΩΣΗ ΤΗΣ ΑΝΑΦΟΡΑΣ

ενός συστήματος που δέχεται δεδομένα διαφορετικών τύπων, καθώς μπορεί να αντιμετωπίσει συνθετότερα φαινόμενα της ανθρώπινης επικοινωνίας, όπως αναφέρθηκαν στην προηγούμενη ενότητα.

Πολυτροπικό σύστημα (Multimodal System) ονομάζεται ένα σύστημα το οποίο διαχειρίζεται δεδομένα που προέρχονται από διαφορετικούς τύπους πληροφορίας, όπως είναι το κείμενο, η εικόνα, ο ήχος και το βίντεο. Ο αγγλικός όρος που αντιστοιχίζεται στον τύπο δεδομένων ονομάζεται modality και θα χρησιμοποιείται εφεξής στη διπλωματική.

Στο πλαίσιο λοιπόν της διπλωματικής δημιουργείται ένα πολυτροπικό σύστημα, το οποίο περιλαμβάνει ένα μοντέλο κειμένου και ένα μοντέλο εικόνας. Για να γίνει η επιλογή αυτών των μοντέλων εξερευνούνται οι ερευνητικές περιοχές της επεξεργασίας φυσικής γλώσσας και της υπολογιστικής όρασης. Με την έρευνα στη βιβλιογραφία και τη διεξαγωγή πειραμάτων, η διπλωματική εργασία προτείνει ένα συγκεκριμένο πολυτροπικό σύστημα για την ανάλυση συναισθήματος προερχόμενο από κείμενο και εικόνα, το οποίο μάλιστα καταφέρνει να είναι ανταγωνιστικό στις επιδόσεις του συγκριτικά με τις πιο πρόσφατες δημοσιεύσεις που αλληλεπιδρούν με το ίδιο σύνολο δεδομένων.

Εκτός αυτού, η διπλωματική εργασία επιχειρεί να ελέγξει την ικανότητα γενίκευσης του προτεινόμενου συστήματος σε τελείως άγνωστα δεδομένα, μέσω διαφόρων πειραμάτων. Ενδεικτικά πειράματα είναι η μετάφραση των κειμένων σε διαφορετικές γλώσσες και η δημιουργία ενός καινούριου συνόλου δεδομένων κειμένων και εικόνας, συλλεγμένα από το ελληνικό Twitter. Στο τέλος δημιουργείται και ιστοσελίδα για την πρόσβαση στο προτεινόμενο σύστημα από οποιονδήποτε χρήστη του διαδικτύου.

Το θέμα που πραγματεύεται η παρούσα διπλωματική εργασία βρίσκει εφαρμογές σε πολλούς κλάδους της σύγχρονης κοινωνίας. Η ανάλυση συναισθήματος μπορεί να χρησιμοποιηθεί από εταιρείες για την αξιολόγηση και το αντίκτυπο των υπηρεσιών που προσφέρουν και από ερευνητές για την αναζήτηση λύσεων μέσω της επιστήμης σε ζητήματα που αφορούν τον άνθρωπο. Εκτός όμως των επαγγελματιών που εργάζονται στην ανάλυση συναισθήματος για επαγγελματικούς σκοπούς, το θέμα της διπλωματικής ιντριγκάρει και τον μέσο άνθρωπο, καθώς σχετίζεται άμεσα με την ανθρώπινη φύση προσπαθώντας να ανιχνεύσει το ανθρώπινο συναίσθημα.

1.3 ΔΙΑΡΘΡΩΣΗ ΤΗΣ ΑΝΑΦΟΡΑΣ

Η διάρθρωση της παρούσας διπλωματικής εργασίας είναι η εξής:

- **κεφάλαιο 2:** Αναλύονται βασικοί όροι που σχετίζονται με τα μοντέλα κειμένου και εξηγούνται τα κυριότερα μοντέλα κειμένου για ανάλυση συναισθημάτων και τα διαθέσιμα εργαλεία που μπορούν να χρησιμοποιηθούν συμπληρωματικά. Επίσης, παρουσιάζονται αρκετές τεχνικές προεπεξεργασίας κειμένων που χρησιμοποιούνται στον κλάδο της επεξεργασίας φυσικής γλώσσας. Δίνονται επιγραμματικά διάφορα σύνολα δεδομένων που αποτελούνται μόνο από κείμενο.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

- **κεφάλαιο 3:** Αναλύονται τα κυριότερα μοντέλα εικόνας για ανάλυση συναισθημάτων. Παρουσιάζονται τεχνικές προεπεξεργασίας εικόνων που χρησιμοποιούνται στον κλάδο της υπολογιστικής όρασης και σύνολα δεδομένων που αποτελούνται μόνο από εικόνες.
- **κεφάλαιο 4:** Παρουσιάζονται οι κυριότερες τεχνικές συνένωσης διαφορετικών modalities. Επιπλέον επισημαίνονται τα κυριότερα πολυτροπικά σετ δεδομένων για ανάλυση συναισθημάτων.
- **κεφάλαιο 5:** Δίνεται η λίστα με τα προγράμματα και τις βιβλιοθήκες που χρησιμοποιήθηκαν κατά την υλοποίηση του πολυτροπικού συστήματος. Αναλύεται το σετ δεδομένων που επιλέχθηκε ως κύριο για τον πειραματισμό. Εξετάζονται οι παράμετροι των μοντέλων και εξηγούνται αναλυτικά. Παρουσιάζεται σχηματικά το προτεινόμενο σύστημα.
- **κεφάλαιο 6:** Παρουσιάζονται και συγχρίνονται τα πειράματα για κάθε μοντέλο. Συγχρίνονται οι διαφορετικές τεχνικές επεξεργασίας των εικόνων προτού δοθούν στο μοντέλο.
- **κεφάλαιο 7:** Εξετάζεται η δυνατότητα χρήσης του συστήματος σε διαφορετικές γλώσσες. Δημιουργείται ένα νέο σετ δεδομένων αποτελούμενο από κείμενα και εικόνες. Δημιουργείται βοηθητική ιστοσελίδα για τη χρήση του μοντέλου.
- **κεφάλαιο 8:** Παρατίθενται τα τελικά συμπεράσματα και τα προβλήματα που προέκυψαν
- **κεφάλαιο 9:** Προτείνονται θέματα για μελλοντική μελέτη και δυνατές επεκτάσεις

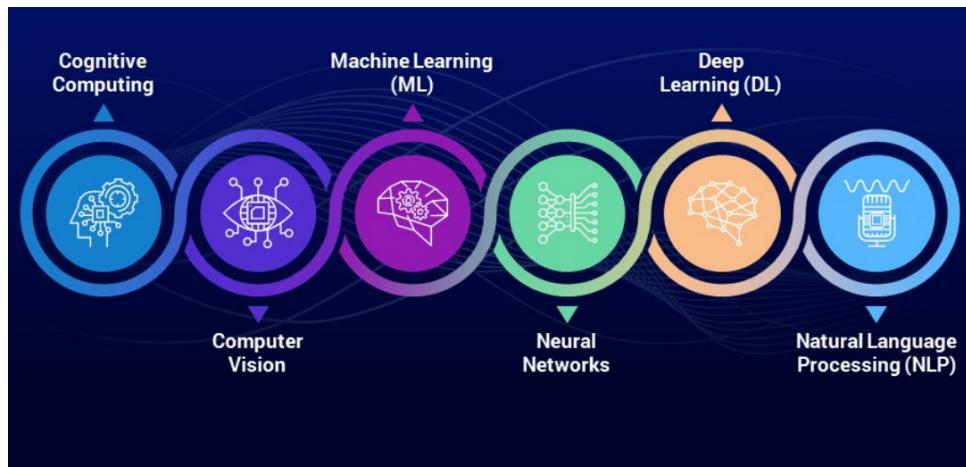
1.4 ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Η Τεχνητή Νοημοσύνη (AI: Artificial Intelligence) αναφέρεται στην ανάπτυξη συστημάτων υπολογιστών που μπορούν να εκτελούν εργασίες που συνήθως απαιτούν ανθρώπινη νοημοσύνη. Αυτές οι εργασίες περιλαμβάνουν την επίλυση προβλημάτων, την κατανόηση της φυσικής γλώσσας, την αναγνώριση προτύπων και τη λήψη αποφάσεων. Η τεχνητή νοημοσύνη συνεργάζεται άμεσα με πολλούς άλλους επιστημονικούς τομείς, όπως η επιστήμη των υπολογιστών, τα μαθηματικά, οι νευροεπιστήμες, η βιοϊατρική και άλλοι. Δίνοντας έναν πιο επίσημο ορισμό για την τεχνητή νοημοσύνη [2]:

Ο κλάδος της επιστήμης των υπολογιστών και της μηχανικής που επικεντρώνεται στη δημιουργία συστημάτων και μηχανών που είναι ικανές να εκτελούν εργασίες, οι οποίες τυπικά χρειάζονται ανθρώπινη νοημοσύνη. Οι εργασίες αυτές περιλαμβάνουν τη χρήση της λογικής, την ικανότητα επίλυσης προβλημάτων, την μάθηση από εμπειρία, την κατανόηση φυσικής γλώσσας και την αλληλεπίδραση με το περιβάλλον.

Κλάδοι της Τεχνητής Νοημοσύνης

Είναι αλήθεια πως η τεχνητή νοημοσύνη έχει μπει στη καθημερινότητα κάθε ανθρώπου, με εφαρμογές σε πολλές πτυχές της ζωής του. Μία κατηγοριοποίηση της τεχνητής νοημοσύνης με κριτήριο την εργασία που προσπαθεί να επιλύσει, δίνεται στους κλάδους που φαίνονται στο παρακάτω σχήμα (σχήμα 1.1). Το σχήμα αναδεικνύει την αλληλοσυσχέτιση μεταξύ των διαφορετικών κλάδων, όπως θα φανεί και στην περιγραφή που ακολουθεί.



Σχήμα 1.1: Κλάδοι της επιστήμης της Τεχνητής Νοημοσύνης

Η γνωστική υπολογιστική (Cognitive Computing) [3] στοχεύει στη δημιουργία συστημάτων που μπορούν να προσομοιώσουν την ανθρώπινη σκέψη. Οι δραστηριότητες που καλείται να επιλύσει συχνά περιλαμβάνουν την κατανόηση φυσικής γλώσσας, την αναγνώριση προτύπων και την επίλυση προβλημάτων, ενώ βρίσκει

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

εφαρμογές στον τομέα της υγείας, στην εξυπηρέτηση πελατών και στα τραπεζικά συστήματα.

Η υπολογιστική όραση (Computer Vision) [4] δημιουργεί μηχανές που είναι ικανές να ερμηνεύουν και να κατανοούν οπτικά δεδομένα. Αυτός ο τομέας έχει εφαρμογές σε αυτόνομα οχήματα, αναγνώριση προσώπου, ανάλυση ιατρικών εικόνων και γενικότερα ζητήματα ταξινόμησης εικόνων και βίντεο. Η πιο γνωστή αρχιτεκτονική μοντέλων αυτού του κλάδου είναι τα συνελικτικά νευρωνικά δίκτυα.

Ένα ακόμη υποσύνολο της τεχνητής νοημοσύνης είναι η μηχανική μάθηση (ML: Machine Learning). Είναι ένας γενικός κλάδος, που αναφέρεται στη δημιουργία αλγορίθμων ικανών να δέχονται δεδομένα στην είσοδο και με βάση αυτά να λαμβάνουν αποφάσεις. Ο αναγνώστης μπορεί να κατευθυνθεί στο [5], όπου γίνεται μια σύντομη αναφορά σε σύγχρονους αλγορίθμους που περιλαμβάνει η μηχανική μάθηση και στις δραστηριότητες που μπορούν να χρησιμοποιηθούν. Βρίσκεται εφαρμογές σε πάρα πολλούς τομείς, όπως είναι η υγειονομική περίθαλψη (διάγνωση), η οικονομία (αναγνώριση απάτης), το ηλεκτρονικό εμπόριο (σύστημα προτάσεων στον καταναλωτή). Στη συνέχεια του κεφαλαίου αναλύονται τα διαφορετικά είδη μάθησης σε κατηγορίες.

Τα νευρωνικά δίκτυα είναι υπολογιστικά μοντέλα που βασίζουν την αρχιτεκτονική τους στον ανθρώπινο εγκέφαλο. Αποτελούνται από βασικές μονάδες που ονομάζονται νευρώνες, οι οποίοι δέχονται μία είσοδο και παράγουν μία έξοδο εφαρμόζοντας μαθηματικές πράξεις στην είσοδο. Τα δίκτυα αυτά αποτελούνται από διάφορα στρώματα, με κυριότερα το στρώμα εισόδου, το στρώμα εξόδου και τα κρυφά στρώματα. Ένα νευρωνικό δίκτυο με πολλά κρυφά στρώματα μπορεί να ονομαστεί βαθύ δίκτυο και να ανήκει στον κλάδο της βαθιάς μάθησης. Κατά την εκπαίδευση οι νευρώνες προσαρμόζουν τα βάρη τους με την τεχνική της οπισθοδρόμησης. Μερικές από τις κύριες εφαρμογές τους είναι η αναγνώριση εικόνας/ήχου, η επεξεργασία φυσικής γλώσσας και η αυτόνομη οδήγηση [6].

Η βαθιά μάθηση αναφέρεται στο σύνολο μοντέλων τα οποία είναι μεγαλύτερα και πιο σύνθετα. Πιο συγκεκριμένα, χρησιμοποιεί νευρωνικά δίκτυα με πολλά κρυφά επίπεδα, επιτρέποντας την επιτυχία δυσκολότερων tasks με μεγαλύτερη ακρίβεια. Οι τεχνολογικές εξελίξεις και τα ισχυρότερα υπολογιστικά συστήματα των τελευταίων χρόνων επέτρεψαν την ανάπτυξη μοντέλων βαθιάς μάθησης.

Η επεξεργασία φυσικής γλώσσας μπορεί να οριστεί ως μια συλλογή υπολογιστικών τεχνικών για αυτόματη ανάλυση και αναπαράσταση ανθρώπινων γλωσσών[7]. Οι περισσότεροι αλγόριθμοι της NLP βασίζονται στην συντακτική αναπαράσταση της εισόδου, έχοντας έτσι τον περιορισμό να αποφασίζουν με βάση μόνο την είσοδο που βλέπουν και να αγνοούν πληροφορίες που πολλές φορές είναι προφανείς για τον ανθρώπινο αναγνώστη, αλλά δε περιέχονται στο κείμενο εισόδου.

Όπως χαρακτηριστικά αναδεικνύεται και στο σχήμα 1.1 και στις παραπόνω παραγράφους επεξήγησης του κάθε κλάδου, τα όρια τους είναι επικαλυπτόμενα και οποιαδήποτε εφαρμογή ή μοντέλο δε μπορεί αναγκαστικά να κατηγοριοποιηθεί σε μόνο έναν κλάδο. Στη διπλωματική εργασία μας το συνολικό σύστημα που προτείνεται περιέχει μοντέλα που ανήκουν και στις 6 παραπόνω κατηγορίες.

Είδη εκμάθησης μοντέλων/αλγορίθμων

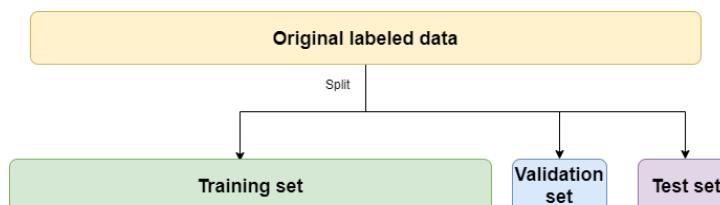
Στον τομέα της τεχνητής νοημοσύνης υπάρχουν διάφορες τεχνικές μάθησης, ώστε να μπορούν τα μοντέλα να μαθαίνουν από τα δεδομένα και να λαμβάνουν αποφάσεις.

- **Επιβλεπόμενη μάθηση (Supervised learning):** Ο αλγόριθμος εκπαιδεύεται σε ένα επισημασμένο σύνολο δεδομένων (labeled dataset), όπου στην κάθε είσοδο έχει ήδη αποδοθεί ένα συγκεκριμένο label. Ο στόχος είναι κατά την εκμάθηση το μοντέλο να επιλέγει τη σωστή κατηγορία σύμφωνα με το label και να μπορεί έτσι να κατηγοριοποιεί στη συνέχεια άγνωστα δεδομένα.
- **Μη επιβλεπόμενη μάθηση (Unsupervised learning):** Το μοντέλο μαθαίνει να προσαρμόζεται σε δεδομένα που δεν έχουν κάποια ετικέτα (label). Στοχεύει στην εύρεση μοτίβων ή δομών στα δεδομένα, ενώ τα πιο κοινά tasks των αλγορίθμων αυτών είναι η ομαδοποίηση και η μείωση διαστάσεων.
- **Ημι-επιβλεπόμενη μάθηση (Semi-supervised learning):** Το μοντέλο κατά την εκμάθηση του αντιμετωπίζει και επισημασμένα και μη επισημασμένα δεδομένα. Προσπαθεί να συνδυάσει τα ωφέλη των δύο προηγουμένων τεχνικών (προσαρμογή των προβλέψεων και ανίχνευση μοτίβων). Μπορεί να χρησιμοποιηθεί όταν δεν υπάρχει η δυνατότητα επισήμανσης ολόκληρου του dataset λόγω μεγέθους και χρόνου.
- **Ενισχυτική μάθηση (Reinforcement learning):** Ο πράκτορας (agent) αλληλεπιδρά με το περιβάλλον του και ανάλογα με τις αποφάσεις που επιλέγει, λαμβάνει και την αντίστοιχη ανταμοιβή. Ο στόχος του είναι η μεγιστοποίηση της συνολικής ανταμοιβής. Βρίσκει εφαρμογές κυρίως στη ρομποτική και σε παιχνίδια.
- **Αυτο-επιβλεπόμενη μάθηση (Self-supervised learning):** Στην αυτο-επιβλεπόμενη μάθηση, που αποτελεί παραλλαγή της μη επιβλεπόμενης μάθησης, το μοντέλο δημιουργεί τα δικά του labels για τα δεδομένα και προσπαθεί να προσαρμοστεί σε αυτά. Βρίσκει εφαρμογές στην επεξεργασία φυσικής γλώσσας (συμπλήρωση λέξεων σε πρόταση) και στην υπολογιστική όραση (συμπλήρωση μέρους μιας εικόνας που λείπει/έχει φθαρθεί).
- **Μεταφορά μάθησης (Transfer learning):** Η μεταφορά μάθησης συμβαίνει όταν ένα μοντέλο έχει ήδη προεκπαιδευτεί (pretraining) σε ένα συνήθως πολύ μεγαλύτερο σύνολο δεδομένων και καλείται να προσαρμοστεί (fine-tuning) σε ένα μικρότερο σύνολο δεδομένων, σε παρόμοιο όμως task. Επιλέγεται η συγκεκριμένη τεχνική για την εξοικονόμηση χρόνου εκμάθησης και όταν τα δεδομένα του dataset δεν επαρκούν για την κανονική εκπαίδευσή του.

Χωρισμός δεδομένων σε υποσύνολα

Στην περίπτωση που η εκμάθηση του μοντέλου γίνεται σε επισημασμένα δεδομένα, για την εκπαίδευση και τον έλεγχο των επιδόσεων του είναι απαραίτητος ο χωρισμός του συνόλου δεδομένων σε τρία υποσύνολα, ανεξάρτητα μεταξύ τους. Τα υποσύνολα ονομάζονται σύνολο εκπαίδευσης, επικύρωσης και ελέγχου. Ο χωρισμός στα 3 υποσύνολα γίνεται με τυχαίο τρόπο, προσδιορίζοντας μόνο το μέγεθος του κάθε υποσυνόλου ως ένα ποσοστό του αρχικού ([σχήμα 1.2](#)).

1. Σύνολο εκπαίδευσης (Train set): Χρησιμοποιείται για την εκμάθηση του μοντέλου. Από τα δεδομένα του συνόλου εκπαίδευσης το μοντέλο μαθαίνει να προσαρμόζεται σε αυτά για το task που έχει να αντιμετωπίσει. Για να το πετύχει αυτό, αλλάζει τις τιμές των παραμέτρων του, προσπαθώντας να μειώσει την τιμή της συνάρτησης απώλειας (loss function). Αποτελεί το μεγαλύτερο υποσύνολο, έχοντας συνήθως ένα ποσοστό της τάξεως του 60-80% του αρχικού dataset.
2. Σύνολο επικύρωσης (Validation set): Χρησιμοποιείται για την επιλογή των καταλληλότερων παραμέτρων στο μοντέλο. Βοηθάει στην αποφυγή του φαινομένου της υπερεκπαίδευσης, κατά το οποίο το μοντέλο εστιάζει μόνο στα δεδομένα όπου έχει εκπαιδευτεί και δε μπορεί να λειτουργήσει σωστά σε άγνωστα δεδομένα. Με το πέρας της εκπαίδευσης του μοντέλου στο σετ εκπαίδευσης, οι επιδόσεις αξιολογούνται στο σετ επικύρωσης και ανάλογα προσαρμόζονται οι υπερπαράμετροι του μοντέλου, όπως κρίνει ο ερευνητής. Έπειτα το μοντέλο ξαναεκπαιδεύεται από την αρχή με τις νέες παραμέτρους και η διαδικασία επαναλαμβάνεται, μέχρις ότου ο ερευνητής είναι ικανοποιημένος με τις επιδόσεις του μοντέλου στο σετ επικύρωσης. Το συνηθισμένο ποσοστό του συνόλου αυτού είναι 10-20%.
3. Σύνολο ελέγχου (test set): Το μοντέλο ελέγχεται σε αυτό το σύνολο και αποδεικνύει την ικανότητά του να γενικεύει σε νέα δεδομένα ή όχι. Είναι το τρίτο κατά σειρά σύνολο που χρησιμοποιείται, αφού πρώτα εκπαιδευτεί το μοντέλο στο σύνολο εκπαίδευσης και προσαρμοστούν οι παράμετροι του με τη βοήθεια του συνόλου επικύρωσης. Στις περισσότερες περιπτώσεις έχει παρόμοιο, αν όχι ίδιο, μέγεθος με το validation set (10-20%).



Σχήμα 1.2: Χωρισμός σε σύνολα εκπαίδευσης/επικύρωσης/ελέγχου

2

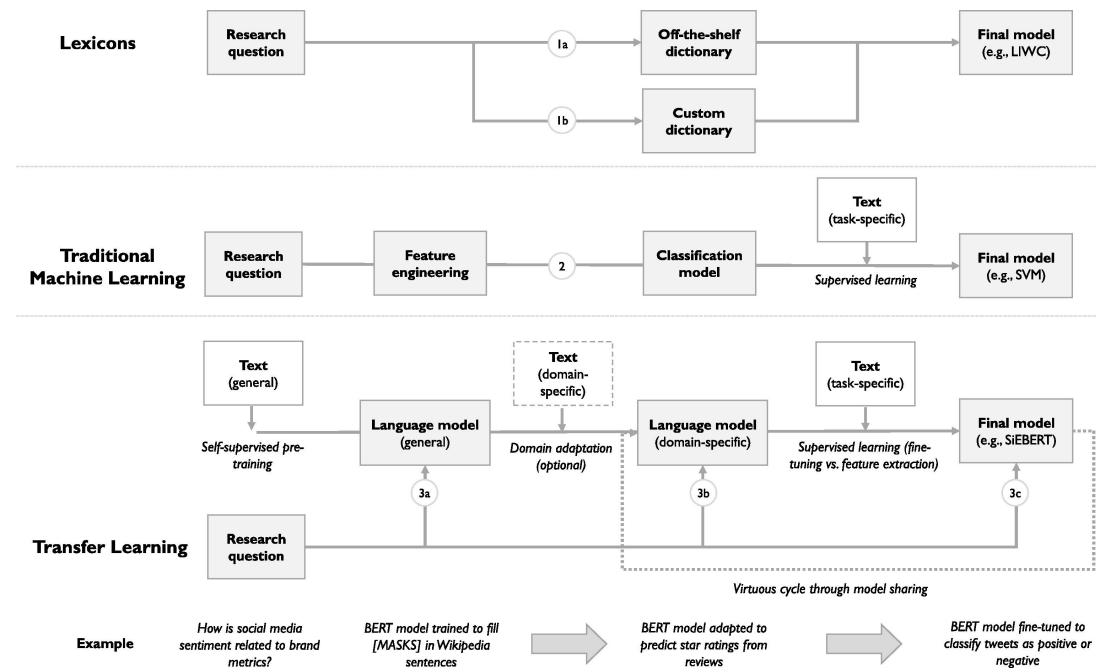
Ανάλυση συναισθήματος βάσει κειμένου

Τεράστιο ερευνητικό ενδιαφέρον παρουσιάζουν οι κλάδοι της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing, NLP) και της Υπολογιστικής Όρασης (Computer Vision, CV). Στη παρούσα διπλωματική εργασία αξιοποιείται ο κλάδος της NLP για την επεξεργασία του κειμένου και ο κλάδος της υπολογιστικής όρασης για την επεξεργασία της εικόνας, καθώς αναφέρεται σε πολυτροπικό σύστημα ταξινόμησης.

Σε αυτό το κεφάλαιο γίνεται μια εκτενής μελέτη των σύγχρονων μεθόδων και εργαλειών που χρησιμοποιούνται για κατηγοριοποίηση κειμένου. Πιο συγκεκριμένα, θα παρουσιαστούν αρχικά κάποιες βασικές έννοιες συνυφασμένες με τα εργαλεία αυτά, όπως είναι τα μεγάλα γλωσσικά μοντέλα (Large Language Models, ή LLM), ο μηχανισμός προσοχής (Attention Mechanism), η οικογένεια μοντέλων που ονομάζονται μετασχηματιστές (Transformers). Εφεξής οι αντίστοιχοι όροι θα χρησιμοποιούνται στα αγγλικά, ώστε να συμβαδίζουν με τη διεθνή βιβλιογραφία.

Στη συνέχεια παρουσιάζονται αναλυτικά τα μοντέλα που μπορούν να χρησιμοποιηθούν για τους σκοπούς της ταξινόμησης. Εκτός όμως από τα διάφορα μοντέλα, αναλύονται οι διαθέσιμες επιλογές προεπεξεργασίας των δεδομένων, όπως αυτές έχουν χρησιμοποιηθεί σε δημοσιευμένες έρευνες του παρελθόντος. Επίσης γίνεται αναφορά στα πιο ευρέως χρησιμοποιούμενα σε δεδομένων για τους σκοπούς του προβλήματος που θέλουμε να επιλύσουμε. Στο [σχήμα 2.1](#) [8] φαίνεται η εξελικτική πορεία των μεθόδων που χρησιμοποιήθηκαν στην ανάλυση συναισθήματος βάσει κειμένου. Τα πρώτα μοντέλα στηρίχτηκαν σε λεξικά, στη συνέχεια χρησιμοποιήθηκαν κλασσικοί ταξινομητές της μηχανικής μάθησης και πλέον γίνεται χρήση μεγάλων γλωσσικών μοντέλων με τη βοήθεια της μεταφοράς μάθησης (transfer learning). Τέλος, το κεφάλαιο κλείνει αναλύοντας εργαλεία που έχουν χρησιμοποιηθεί σε αντίστοιχες έρευνες του παρελθόντος και έχουν αποδόσει.

ΚΕΦΑΛΑΙΟ 2. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΆΣΕΙ ΚΕΙΜΕΝΟΥ



Σχήμα 2.1: Η εξέλιξη στην ανάλυση συναισθήματος βάσει κειμένου, από τα λεξικά στη μεταφερόμενη μάθηση.

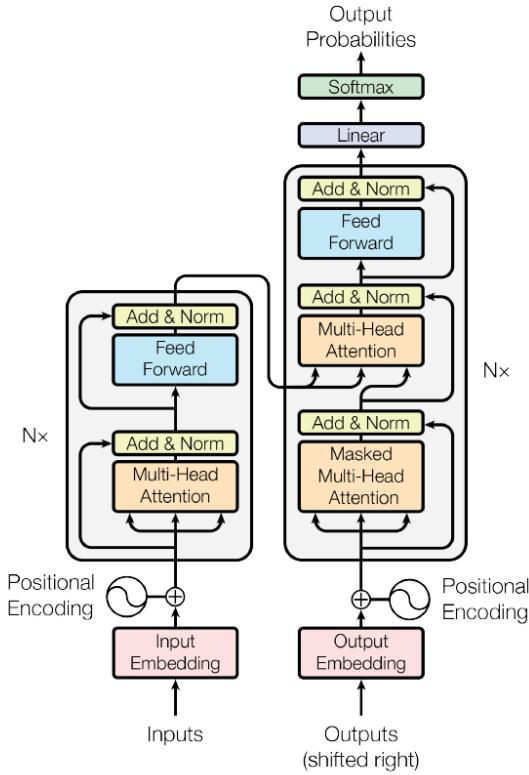
2.1 ΚΑΤΑΝΟΗΣΗ ΒΑΣΙΚΩΝ ΟΡΩΝ

Σε αυτή την ενότητα αναλύονται κάποιοι βασικοί όροι που χρειάζονται για την κατανόηση των μοντέλων που ακολουθούν και γενικά για την υλοποίηση που επιλέχθηκε.

Transformers

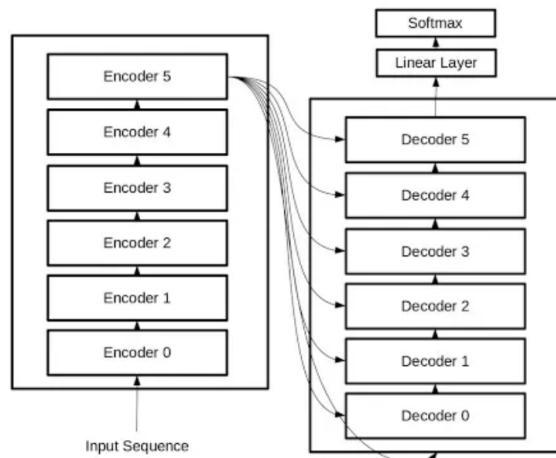
Ο όρος Transformers παρουσιάστηκε για πρώτη φορά στη δημοσίευση με τίτλο *Attention is all you need*[9], η οποία δημοσιεύθηκε το 2017 από τη Google. Το paper αυτό έφερε επανάσταση στον κλάδο του NLP, επιτρέποντας τη δημιουργία πολλών μοντέλων αυτού του είδους, τα οποία αυξήσαν σημαντικά τις επιδόσεις σε προβλήματα που κλήθηκαν να αντιμετωπίσουν. Πριν την εμφάνιση των transformers, στον κλάδο του NLP χρησιμοποιούνταν κατά κόρον τα αναδρομικά νευρωνικά δίκτυα (RNN) και τα δίκτυα μακράς βραχύχρονης μνήμης (LSTM), τα οποία όμως εμφάνιζαν προβλήματα όταν χρησιμοποιούνταν σε μεγάλα σετ δεδομένων, επειδή αυξανόταν σημαντικά ο χρόνος εκπαίδευσης του μοντέλου και υπήρχε το φαινόμενο να χάνονται οι κλίσεις (gradients) από παλιότερα δεδομένα, οδηγώντας σε σημαντική απώλεια πληροφορίας. Για την αντιμετώπιση αυτού του προβλήματος, τα transformers λειτουργούν παράλληλα, επιτρέποντας την ταυτόχρονη επεξεργασία πολλών δεδομένων, μειώνοντας έτσι τον χρόνο εκμάθησης του μοντέλου. Η αρχιτεκτονική του μοντέλου αυτού φαίνεται στην παρακάτω εικόνα (σχήμα 2.2). Είναι ένα μοντέλο που λαμβάνει ως είσοδο μια ακολουθία και εξάγει στην έξοδο επίσης μια ακολουθία (sequence to sequence model). Η βασική δομή αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή (encoder, decoder), ωστόσο μπορούν να χρησιμοποιηθούν σειριακά περισσότεροι για τη βελτίωση της ικανότητας του μοντέ-

2.1. ΚΑΤΑΝΟΗΣΗ ΒΑΣΙΚΩΝ ΌΡΩΝ



Σχήμα 2.2: Η αρχιτεκτονική του transformer μοντέλου. Αριστερά φαίνεται το κομμάτι του κωδικοποιητή και δεξιά ο αποκωδικοποιητής.

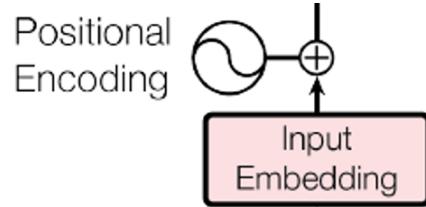
λου να μαθαίνει και να προβλέπει ([σχήμα 2.3](#)). Αξίζει να γίνει ανάλυση μερικών εκ των blocks της αρχιτεκτονικής του transformer, αφού θα χρησιμοποιηθούν εκτενώς στη συνέχεια της εργασίας.



Σχήμα 2.3: Στοίβα από κωδικοποιητές-αποκωδικοποιητές

Embedding layer

Στο κείμενο εισόδου αντιστοιχίζεται σε κάθε λέξη ένας αριθμός, ο οποίος περιέχει και πληροφορία για τη θέση της συγκεκριμένης λέξης στο κείμενο. Η ίδια διαδικασία συμβαίνει ταυτόχρονα και στο κείμενο εξόδου.



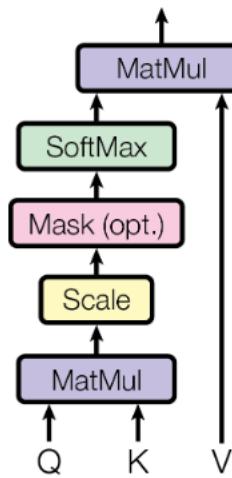
Μηχανισμός Attention

Ο μηχανισμός αυτός χρησιμοποιεί 3 πίνακες, οι οποίοι ονομάζονται Q (Queries), K (Keys) και V (Values). Οι πίνακες αυτοί προκύπτουν με πολλαπλασιασμό της εισόδου με 3 διαφορετικούς πίνακες βαρών W_q , W_k και W_v . Τα βάρη αυτά αλλάζουν κατά τη διάρκεια της εκπαίδευσης (αρχικά αρχικοποιούνται σε τυχαίες τιμές). Μαθηματικά, οι τιμές στην έξοδο αυτού του στρώματος προκύπτουν από τον τύπο:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Η παράμετρος στον παρονομαστή χρησιμοποιείται για scaling των εξόδων. Για να αναδειχθεί η σημασία του attention, αναφέρονται ενδεικτικά κάποιες δημοσιεύσεις που δε σχετίζονται άμεσα με το αντικείμενο μας, αλλά δείχνουν τη σημασία του αλγορίθμου. Έχει χρησιμοποιηθεί λοιπόν για μετάφραση κειμένου σε διαφορετικές γλώσσες [10], για αναγνώριση φωνής [11] αλλά και σε προβλήματα ενισχυτικής μάθησης [12].

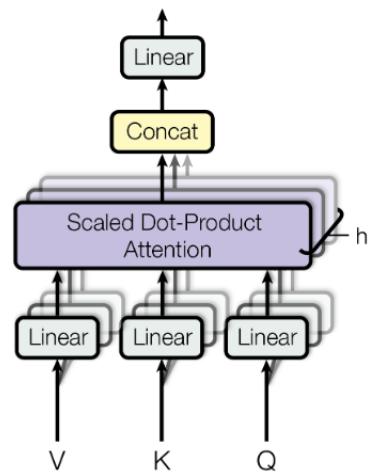
Scaled Dot-Product Attention



Μηχανισμός Multi Head Attention

Εδώ χρησιμοποιούνται ταυτόχρονα πολλοί μηχανισμοί attention, με διαφορετικά βάρη στους πίνακες που αναφέραμε, για την ενίσχυση της ικανότητας του μοντέλου. Μετά από κάθε τέτοιο block, ακολουθεί μια πρόσθεση της εξόδου με την αρχική είσοδο και έπειτα η κανονικοποίηση του αποτελέσματος.

Multi-Head Attention



Μεγάλα Γλωσσικά Μοντέλα (LLM)

Ένα μεγάλο γλωσσικό μοντέλο (LLM) είναι ένα μοντέλο βασισμένο σε νευρωνικά δίκτυα, που χαρακτηρίζεται από τον μεγάλο αριθμό των παραμέτρων του, συχνά στις δεκάδες εκατομμύρια ή ακόμα και δισεκατομμύρια, και εκπαιδεύεται σε τεράστια σύνολα δεδομένων κειμένου για την εκτέλεση καθηκόντων σχετικών με τη φυσική κατανόηση, δημιουργία και επεξεργασία της φυσικής γλώσσας. Αυτά τα μοντέλα έχουν προχωρήσει σημαντικά την τεχνολογία στον τομέα της επεξεργασίας της φυσικής γλώσσας και έχουν επιτύχει την προηγμένη απόδοση σε διάφορες εργασίες, συμπεριλαμβανομένης της ανάλυσης συναισθημάτων. Τα LLMs δε μπορούν να εκπαιδευτούν άμεσα από έναν απλό προσωπικό υπολογιστή, λόγω του μεγάλου όγκου πληροφορίας που χρειάζονται για την εκπαίδευσή τους και του πλήθους των παραμέτρων τους, γι' αυτό και αξιοποιούνται από τους χρήστες αλληλεπιδρώντας με κάποιο API, έχοντας ήδη προεκπαιδευτεί από κάποια εταιρεία/ερευνητικό κέντρο. Μερικά από τα γνωστότερα παραδείγματα αυτής της κατηγορίας μοντέλων είναι τα BERT [13] (θα αναφερθούμε εκτενώς στη συνέχεια), GPT-3 [14], GPT-4 [15], Llama [16], PaLM [17].

2.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ

Προτού ένα μοντέλο τροφοδοτηθεί με κείμενα για ταξινόμηση, είναι απαραίτητο τα δεδομένα να περάσουν από κάποια στάδια προεπεξεργασίας. Στην αρχική τους μόρφη, τα κείμενα δεν είναι καθαρά, εννοώντας πως περιέχουν σημεία στίξης, κεφαλαία γράμματα, αριθμητικές τιμές, ειδικούς χαρακτήρες. Ανάλογα με τη περίσταση, θα πρέπει να ληφθεί προσεκτικά η απόφαση διαχείρισης των παραπάνω περιπτώσεων. Για παράδειγμα, σε ένα σετ δεδομένων που προέρχεται από κάποια πλατφόρμα κοινωνικής δικτύωσης, οι ειδικοί χαρακτήρες, όπως είναι τα hashtags (#), είναι πιθανό να περιέχουν σημαντική πληροφορία για την αναγνώριση του συναισθήματος. Μπορεί όμως το ίδιο σύμβολο σε κάποιο άλλο σετ δεδομένων να προσδιορίζει για παράδειγμα τον αύξοντα αριθμό σε μια ιστοσελίδα με κριτικές, όπου σε αυτή τη περίπτωση δε προσφέρει κάποια πληροφορία για την εκδήλωση του συναισθήματος. Παρακάτω παραθέτονται οι πιο διαδεδομένες τεχνικές προεπεξεργασίας κειμένου.

Tokenization: Tokenization ονομάζεται η διαδικασία διαχωρισμού του αρχικού κειμένου σε λέξεις/προτάσεις. Όπως θα αναλυθεί και στη συνέχεια, η διαδικασία αυτή ειναι απαραίτητη στα LLMs και καθένα από αυτά διαθέτει ένα εργαλείο για το tokenization, που ονομάζεται tokenizer.

Lowercasing: Η διαδικασία μετατροπής των κεφαλαίων γραμμάτων σε μικρά ονομάζεται lowercasing. Βοηθάει στην κανονικοποίηση των κειμένων, ωστόσο υπάρχει η περίπτωση τα κεφαλαία γράμματα να εκφράζουν συναισθήματα οργής, μίσους. Άρα το αν θα πραγματοποιηθεί lowercasing είναι μια απόφαση που ωφείλει να λάβει ο ερευνητής ανάλογα με τα δεδομένα που προσπαθεί να ταξινομήσει.

Αφαίρεση stop words: Συχνά χρησιμοποιούμενες λέξεις που δεν εμπεριέχουν πληροφορία έχουν ενταχθεί σε μια οικογένεια λέξεων που ονομάζονται stop words. Διαφορετικές λίστες με stop words δημοσιεύονται για διαφορετικές γλώσσες [18] [19], αλλά και για συγκεκριμένους τομείς με ξεχωριστή ορολογία [20].

Stemming/Lemmatization:

- Κατά το stemming, αφαιρείται η κατάληξη από τη λέξη εισόδου, που συχνά όμως μπορεί να οδηγεί σε λάθος αποτέλεσμα. Χρησιμοποιείται για την αύξηση της ταχύτητας σε μεγάλα datasets.
- Κατά το lemmatization, λαμβάνεται υπόψη το περιεχόμενο των συμφραζόμενων και αφαιρείται η κατάληξη της λέξης, διατηρώντας όμως το νόημα της. Το μειονέκτημα στη διαδικασία είναι το υπολογιστικό κόστος.

Χειρισμός ειδικών χαρακτήρων και συμβόλων: Η επιλογή αφαίρεσης ή διατήρησης των ειδικών χαρακτήρων και συμβόλων έγκειται στη χρησιμότητά τους ανάλογα με το task, όπως τονίσθηκε και στην εισαγωγή της ενότητας.

Έλεγχος για λάθη: Το κείμενο μπορεί να περιέχει λέξεις με ορθογραφικά/τυπογραφικά λάθη. Έχουν προταθεί μέθοδοι ελέγχου και επιδιόρθωσης τέτοιων λαθών στη βιβλιογραφία [21].

2.3 DATASETS ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΆΣΕΙ ΚΕΙΜΕΝΟΥ

Η ανάλυση συναισθήματος για πολλά χρόνια βασιζόταν μόνο σε είσοδο κειμένων. Παρόλο που η διπλωματική εργασία εστιάζει σε πολυτροπική είσοδο, από ερευνητικό ενδιαφέρον αξίζει να αναφερθούν σύντομα μερικά διάσημα datasets κειμένων που θεμελίωσαν την ανάλυση συναισθημάτων.

- Amazon Customer Reviews [22]: Το dataset αποτελείται από περίπου 35 εκατομμύρια κριτικές από πελάτες της ιστοσελίδας [Amazon](#). Οι κριτικές ταξινομούνται σε θετικές και αρνητικές.
- IMDB Review Dataset [23]: Το dataset συλλέγει 50000 κριτικές χρηστών σε ταινίες από το [IMDb](#). Οι κριτικές χωρίζονται σε θετικές ($\geq 7/10$) και αρνητικές ($\leq 4/10$). Οι ουδέτερες κριτικές δε λαμβάνονται υπόψη.
- Yelp Reviews Dataset [24]: Κριτικές από διάφορες επιχειρήσεις (εστιατόρια, καταστήματα κτλπ.) συλλέχθηκαν από την ιστοσελίδα [Yelp](#) ώστε να σχηματιστεί αυτό το dataset. Οι κριτικές συνοδεύονται από βαθμολογία κλίμακας 1 εώς 5.
- Twitter Sentiment Analysis Dataset [25]: Το σετ δεδομένων δημιουργήθηκε χρησιμοποιώντας το Twitter API για συλλογή tweets από τη γνωστή πλατφόρμα κοινωνικής δικτύωσης. Το σετ εκπαίδευσης περιέχει 1.600.000 tweets και είναι το πρώτο μεγάλο dataset που λαμβάνει πληροφορίες από μέσα κοινωνικής δικτύωσης.
- Stanford Sentiment Treebank [26]: Αποτελείται από κριτικές ταινιών, έχοντας labels σε 5 κατηγορίες, από πολύ αρνητικό συναίσθημα μέχρι πολύ θετικό. Έχει το πλεονέκτημα πως εισάγει ιεραρχική δομή για την απόδοση του συναισθήματος σε μία φράση χωρίζοντας την σε λέξεις, βιοηθώντας έτσι τους

ερευνητές να εξερευνήσουν το συναίσθημα σε μικρότερα κομμάτια του αρχικού κειμένου. Το μειονέκτημά του είναι το μικρό του μέγεθος (10662 προτάσεις).

- Webis-CLS-10 Dataset [27]: Ειδικεύεται στην ανάλυση συναίσθηματος από διαφορετικές γλώσσες. Αποτελείται και αυτό από κριτικές ταινιών, γραμμένες στα αγγλικά, στα γερμανικά, στα γαλλικά και στα ιαπωνικά. Μπορεί να χρησιμοποιηθεί για εκπαίδευση μοντέλων που να μπορούν να διαχειριστούν πολλές γλώσσες. Περιέχει 2000 δείγματα για εκπαίδευση, 2000 δείγματα για έλεγχο και πολλά ακόμη (9000 με 50000 ανάλογα τη γλώσσα και τον τομέα) που δεν τους έχει αποδοθεί συναίσθημα. Η ταξινόμηση γίνεται στις κατηγορίες θετικό και αρνητικό.

2.4 Γλωσσικά Μοντέλα/ Μοντέλα Κειμενού

Στην ενότητα αυτή συνοψίζονται τα χρησιμοποιούμενα μοντέλα στη βιβλιογραφία για την κατηγοριοποίηση κειμένου (text classification) στην ανάλυση συναίσθηματος. Τα μοντέλα αυτά μπορούν επίσης να χρησιμοποιηθούν γενικότερα σε tasks που εντάσσονται στο κλάδο της επεξεργασίας φυσικής γλώσσας.

BERT

Το BERT (Bidirectional Encoder Representations from Transformers, ελληνικά: Αναπαραστάσεις Αμφίδρομου Κωδικοποιητή από Μετασχηματιστές) [13] είναι ένα μοντέλο γλωσσικής αναπαράστασης. Το BERT είναι ένα γενικό μοντέλο, δηλαδή δεν προορίζεται για ένα συγκεκριμένο task στο κλάδο του NLP, αλλά μπορεί να προσαρμοσθεί αποτελεσματικά στα περισσότερα tasks.

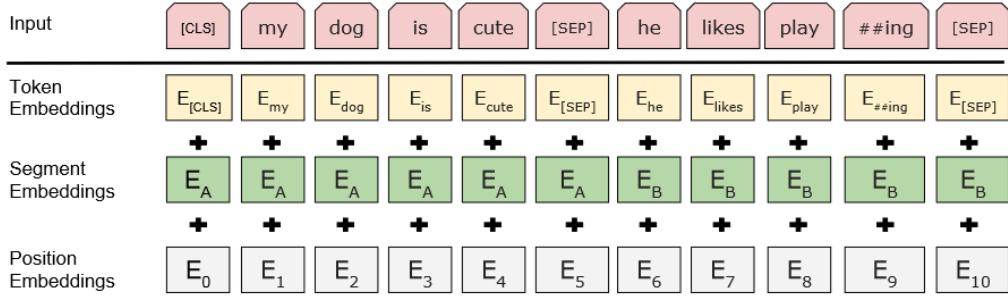
Το BERT είναι ένα *pretrained* μοντέλο. Αυτό σημαίνει πρακτικά πως το μοντέλο έχει εκπαιδευτεί με κείμενα που βρίσκονται στο διαδικτύο και τα οποία δεν είναι annotated, δεν έχουν κατηγοριοποιηθεί δηλαδή πρωτύτερα από κάποιον άνθρωπο. Τέτοια μοντέλα μπορούν έπειτα να προσαρμοστούν σε κάποιο σετ δεδομένων που προσπαθεί να επιλύσει κάποιο συγκεκριμένο task, μια διαδικασία που ονομάζεται fine-tuning. Η διαδικασία της πρώιμης εκμάθησης επιτρέπει στο μοντέλο να επεξεργαστεί πολύ μεγαλύτερο όγκο πληροφορίας, απ' όση θα μπορούσε να επεξεργαστεί λαμβάνοντας σαν είσοδο μόνο annotated δεδομένα.

Το BERT χρησιμοποιεί την τεχνολογία των transformers που αναλύθηκε προηγουμένως, με αποτέλεσμα να μπορεί να συνδύαζει το περιεχόμενο ενός κειμένου διαβάζοντας και προς τις δύο κατευθύνσεις. Έτσι, όταν προσπαθεί να συμπληρώσει για παράδειγμα την άγνωστη λέξη σε μία πρόταση, λαμβάνει υπόψη όχι μόνο την πρόταση μέχρι εκείνο το σημείο, αλλά και τη συνέχεια της, όπως ακριβώς θα έκανε και ένας άνθρωπος όταν θα καλούνταν να συμπληρώσει το κενό σε μια αντίστοιχη ασκηση. Η συμβολή των προηγούμενων και επόμενων λέξεων είναι ταυτόχρονη.

Προτού η είσοδος διθεί στο μοντέλο, πρέπει να της προστεθούν κάποια μεταδεδομένα ([σχήμα 2.4](#)). Πιο συγκεκριμένα:

- Ενσωμάτωση ειδικών tokens: Το [CLS] token εισάγεται στην αρχή του κειμένου εισόδου και το [SEP] token εισάγεται σε κάθε αλλαγή πρότασης.

- Τμηματοποίηση: Εισάγεται δείκτης ο οποίος καθορίζει για κάθε λέξη σε ποιά πρόταση ανήκει
- Θέση: Εισάγεται δείκτης που αναδεικνύει τη θέση της λέξης στο κείμενο εισόδου.



Σχήμα 2.4: Αναπαράσταση της εισόδου στο BERT μοντέλο

Για την εκπαίδευση του μοντέλου, χρησιμοποιούνται δύο τεχνικές, στις οποίες εκπαιδεύεται ταυτόχρονα. Στην πρώτη, η οποία ονομάζεται *Masked LM*, τοποθετούνται μάσκες στο 15% των λέξεων της εισόδου, τις οποίες το BERT καλείται να προβλέψει. Η δεύτερη τεχνική ονομάζεται πρόβλεψη επόμενης πρότασης (NSP), όπου δίνονται στο μοντέλο δύο προτάσεις με το token [SEP] αναμεσά τους και το BERT αποφασίζει αν η δεύτερη πρόταση αποτελεί λογική συνέχεια της πρώτης. Το BERT δημοσιεύθηκε με 2 αρχιτεκτονικές, όπως φαίνονται στον παρακάτω πίνακα:

Model	Layers	Hidden nodes	Attention heads	parameters
BERT base	12	768	12	110M
BERT large	24	1024	16	340M

Για κάθε μία αρχιτεκτονική, δημοσιεύθηκαν δύο μοντέλα, το ένα είναι ικανό να διαχειρίζεται είσοδο μόνο με μικρά γράμματα (uncased), ενώ το δεύτερο διαχειρίζεται μικρούς και κεφαλαίους χαρακτήρες (cased).

Με τη δημοσίευσή του, το μοντέλο αυτό κατάφερε να αποκτήσει κορυφαίες επιδόσεις σε πολλά tasks του τομέα, με κυριότερα την απάντηση ερωτήσεων, την ανάλυση συναισθημάτων, την ανίχνευση παραφράσεων. Για την χρήση του μοντέλου, είναι απαραίτητο πρώτα να επιλεγεί ένα σετ δεδομένων πάνω στο οποίο το μοντέλο θα γίνει fine-tuned. Έπειτα από λίγες μόνο επαναλήψεις, το BERT θα είναι έτοιμο να ανταπεξέλθει στο συγκεκριμένο task που επιδιώκει το dataset.

RoBERTa

Το μοντέλο Roberta (A Robustly optimized BERT pretraining Approach) [28] ήρθε ως απάντηση της Facebook AI στο BERT της Google AI. Το μοντέλο αυτό εκπαιδεύτηκε σε περίπου δεκαπλάσιο πλήθος δεδομένων συγκριτικά με το BERT και κατάφερε να ξεπεράσει το BERT στα benchmarks που συγκρίθηκαν.

2.4. ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ/ MONTELA KEIMENOY

Εκτός από τον μεγαλύτερο αριθμό δεδομένων, οι ερευνητές αποφάσισαν να αφαιρέσουν τη πρόβλεψη επόμενης πρότασης από τα tasks του μοντέλου κατά την εκπαίδευση του. Επιπλέον στο RoBERTa χρησιμοποιήθηκε μεγαλύτερο μέγεθος batch και μακρύτερες ακολουθίες δόθηκαν στην είσοδο. Προστέθηκε ακόμη ένας υπχανισμός δυναμικής τοποθέτησης της μάσκας.

Βέβαια και αυτό το μοντέλο είναι pretrained λόγω του μέγεθους του και για να χρησιμοποιηθεί πρέπει πρώτα να προσαρμοστεί στο συγκεκριμένο task/dataset.

ALBERT

Η συνεχής αύξηση στο μέγεθος των μοντέλων παρουσιάζει δυσκολίες, αφού αυξάνει το χρόνο εκμάθησης και υπάρχουν περιορισμοί όσον αφορά τους διαθέσιμους υπολογιστικούς πόρους. Γνωρίζοντας τα παραπάνω προβλήματα, το ALBERT(A Lite BERT)[29] δημοσιεύεται το 2020 με στόχο την επίλυσή τους, προτείνοντας δύο τεχνικές μείωσης των παραμέτρων:

- **Παραγοντοποίηση:** Οι ερευνητές του ALBERT διαπιστώνουν πως το μέγεθος που προκύπτει μετά τα embeddings, είναι πάντοτε ίσο με το μέγεθος του κρυφού επιπέδου, που σημαίνει πως όσο μεγαλώνει το μέγεθος του κρυφού επιπέδου θα ακολουθεί και το μέγεθος του κάθε embedding. Για να το αποφύγει αυτό, οι παράμετροι ενσωμάτωσης παραγοντοποιούνται, σχηματίζοντας δύο πίνακες μικρότερων διαστάσεων.
- **Διαμοιρασμός παραμέτρων:** Το ALBERT αποφασίστηκε να μοιράζει όλες τις παραμέτρους του μεταξύ των σταδίων attention και του feed forward δικτύου, σε κάθε στρώμα του.

Οι παραπάνω τροποποιήσεις οδήγησαν στη δημοσίευση 4 μοντέλων, με τις εξής αρχιτεκτονικές:

Model	Layers	Hidden nodes	Embedding	parameters
ALBERT base	12	768	128	12M
ALBERT large	24	1024	128	18M
ALBERT xlarge	24	2048	128	60M
ALBERT xxlarge	12	4096	128	235M

Εύκολα παρατηρεί κανείς τη μείωση στον αριθμό των παραμέτρων και το σταθερό μέγεθος των embeddings που δεν ακολουθεί το αύξον μέγεθος του κρυφού στρώματος.

DeBERTa

Το επόμενο μοντέλο της οικογένειας των BERT μοντέλων ονομάζεται DeBERTa (Decoding-enhanced BERT with Disentangled Attention) και δημοσιεύθηκε από τη Microsoft AI το 2021. Το DeBERTa χρησιμοποιεί δύο τεχνικές για να καταφέρει να ξεπεράσει στα περισσότερα benchmarks τα προηγούμενα μοντέλα:

- Διαφορετικός μηχανισμός attention: Εδώ κάθε λέξη εκπροσωπείται από δύο διαφορετικά διανύσματα, το ένα χρησιμοποιείται για την κωδικοποίηση του περιεχομένου και το άλλο για τη κωδικοποίηση της σχετικής θέσης, σε αντίθεση με το BERT, όπου τα δύο αυτά διανύσματα προστίθονταν και προέκυπτε το τελικό διάνυσμα.
- Απόλυτη θέση: Εκτός από τη σχετική θέση της κάθε λέξεως στο κείμενο εισόδου, το μοντέλο χρησιμοποιεί και την απόλυτη θέση του, ώστε να εκμεταλλευτεί τις διαφορές που μπορεί να έχει στην πρόβλεψη της λέξης, όπως για παράδειγμα αν μία λέξη χρησιμοποιείται σαν υποκείμενο ή αντικείμενο στην εξεταζόμενη πρόταση. Στην δημοσιευμένη έρευνα αναδεικνύεται η χρησιμότητα της απόλυτης θέσης, στην πρόταση "Ένα νέο μαγαζί άνοιξε στο νέο εμπορικό". Αν τοποθετηθεί μάσκα ταυτόχρονα στις λέξεις "μαγαζί" και "εμπορικό", τότε με τη παραπάνω λειτουργία το μοντέλο θα μπορεί να καταλάβει πως το υποκείμενο της πρότασης είναι η λέξη "μαγαζί" και να την τοποθετήσει στη σωστή θέση μέσα στη πρόταση.

2.5 ΕΡΓΑΛΕΙΑ ΚΕΙΜΕΝΟΥ

Για την ανάλυση συναισθημάτων, εκτός από τη χρήση κάποιου ισχυρού νευρωνικού δικτύου, αρκετές φορές χρησιμοποιούνται συμπληρωματικά και εργαλεία που βασίζονται κυρίως στην κατηγοριοποίηση του κειμένου με βάση κάποιο λεξικό. Το λεξικό αποτελείται από συναισθηματικά φορτισμένες λέξεις/φράσεις, οι οποίες όταν εντοπίζονται στο κείμενο, τους αποδίδεται η τιμή του λεξικού, που είναι ενδεικτική για το συναίσθημα που αντιπροσωπεύει. Στις αρχές της ανάλυσης συναισθήματος οι μέθοδοι που βασίζονται σε λεξικά είχαν υπερισχύσει, καθώς η υλοποίησή τους είναι ευκολότερη συγκριτικά με τα μεγάλα νευρωνικά δίκτυα. Ωστόσο είναι προφανές πως δε μπορούν να συγκριθούν στις επιδόσεις με τα δίκτυα βαθιάς μάθησης. Παρόλα αυτά, λόγω της ευκολίας ενσωμάτωσής τους, βρίσκουν εφαρμογή ακόμη και σήμερα, λειτουργώντας συμπληρωματικά στο μοντέλο μηχανικής μάθησης που εφαρμόζεται [30]. Η συνδυαστική χρήση ενός νευρωνικού δικτύου και ενός λεξικού ονομάζεται υβριδική προσέγγιση. Παρακάτω γίνεται αναφορά σε μερικά από τα εργαλεία που χρησιμοποιούνται ακόμη και σήμερα.

VADER: Είναι ένα εργαλείο ανάλυσης συναισθημάτων που βασίζεται σε λεξικό και προκαθορισμένους κανόνες, ειδικά σχεδιασμένο για κείμενο προερχόμενο από μέσα κοινωνικής δικτύωσης. Το VADER [31] είναι προεκπαιδευμένο και είναι ικανό να χειριστεί κείμενο ανεξαρτήτου γλώσσας. Ο τρόπος σχεδιασμού του του επιτρέπει να διαχειρίζεται αποτελεσματικά κείμενο που περιέχει ανεπίσημη γλώσσα, emoticons, ανολοκλήρωτες προτάσεις και slang. Εκτός από το να επιστρέψει ένα label για το κείμενο εισόδου, η έξοδος του εργαλείου συνοδεύεται από τιμές ποσοστών για κάθε συναίσθημα που καθορίζουν το πόσο σίγουρο είναι το εργαλείο για τη πρόβλεψή του.

TextBlob και NLTK βιβλιοθήκες: Η Python διαθέτει τις βιβλιοθήκες [TextBlob](#) και [NLTK](#), οι οποίες διευκολύνουν την επεξεργασία φυσικής γλώσσας. Ξεκινώντας με την βιβλιοθήκη TextBlob, μπορεί κανείς να προεπεξεργαστεί τα κείμενα εισόδου,

να κάνει ανάλυση συναισθήματος με το εργαλείο που περιέχει και να μεταφράσει τα κείμενα σε άλλες γλώσσες. Η TextBlob δημιουργήθηκε με στόχο να ενισχύσει την ήδη υπάρχουσα NLTK, προσδίδοντας της επιπλέον λειτουργίες και εστιάζοντας στην απλοτητα και στην ευκολία χρήσης.

Ας δούμε επιγραμματικά και μερικά γνωστά λεξικά.

SentiWordNet 3.0: Το SentiWordNet 3.0 [32] αποτελεί την τρίτη έκδοση της σειράς των SentiWordNet λεξικών. Για κάθε είσοδο το λεξικό αποδίδει μια σειρά από τιμές, οι οποίες δείχνουν τον βαθμό θετικότητας/αρνητικότητας/ουδετερότητας. Οι τιμές αυτές βρίσκονται στο διάστημα 0 εώς 1 και όσο μεγαλύτερες είναι εκφράζουν ισχυρότερο συναίσθημα. Το λεξικό συμπεριλαμβάνει λέξεις που δε περιορίζονται μόνο σε ουσιαστικά, αλλά επεκτείνονται και σε ρήματα, επίθετα και επιρρήματα.

SentiStrength: Και το SentiStrength [33] εστιάζει στην απόδοση τιμών συναίσθημάτος στο κείμενο εισόδου. Υποστηρίζει την εισαγωγή κειμένου σε διάφορες γλώσσες, πέραν των αγγλικών. Οι τιμές που επιστρέφει αναφέρονται στο θετικό και στο αρνητικό συναίσθημα, δεν εξετάζει δηλαδή την ουδετερότητα.

3

Ανάλυση συναισθήματος βάσει εικόνας

Το παρόν κεφάλαιο ακολουθεί παρόμοια δομή με το προηγούμενο, με τη θεμελιώδη διαφορά πως τώρα μελετούνται εκτενώς τα δεδομένα εικόνας. Αρχικά θα παρουσιαστούν τα βήματα προεπεξεργασίας των εικόνων, όπου πρέπει να λαμβάνεται υπόψη το μοντέλο το οποίο θα χρησιμοποιηθεί, το σετ δεδομένων που θα επεξεργαστεί και το task προς επίλυση.

Ακολουθώντας παρόμοια φιλοσοφία, θα συζητηθούν τα datasets εικόνων, τα οποία όμως ιστορικά δεν είχαν την ίδια επίδραση συγκριτικά με τα datasets κειμένου στην ανάλυση συναισθήματος, διότι οι εικόνες από μόνες τους εμφανίζουν δυσκολία στο να τους αποδοθεί το σωστό συναίσθημα.

Αμέσως μετά θα αναλυθούν τα σύγχρονα μοντέλα, ξεκινώντας την ανάλυση με τα συνελικτικά που χρησιμοποιήθηκαν εκτενώς τη προηγούμενη δεκαετία και προχωρώντας στα μοντέλα που βασίζονται σε Transformers, που έκαναν την εμφάνισή τους το 2021 και έχουν υπερισχύσει στην ερευνητική κοινότητα με την ικανότητά τους να προσαρμόζονται στα περισσότερα tasks με υψηλές επιδόσεις.

3.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΑΣ

Όπως τονίστηκε και στην περίπτωση του κειμένου, ετσί και εδώ, η προεπεξεργασία των εικόνων πριν εισαχθούν στο μοντέλο είναι καθοριστική για την ικανότητα εκμάθησης και προσαρμογής του στο επιλεγμένο task και dataset.

Αλλαγή ανάλυσης της εικόνας: Τα περισσότερα μοντέλα βαθιάς μάθησης απαιτούν συγκεκριμένο και κοινό μέγεθος εικόνας στο dataset, όπως ορίζεται στην αρχιτεκτονική και στον τρόπο εκπαίδευσής τους. Σε αυτές τις περιπτώσεις θα πρέπει οι εικόνες εισόδου να αποκτήσουν ένα σταθερό μέγεθος.

Κανονικοποίηση: Μία καλή τεχνική προεπεξεργασίας είναι η κανονικοποίηση των τιμών των pixels της εικόνας, συχνά στα διαστήματα $[0,1]$ ή $[-1,1]$. Η τεχνική αυτή βοηθάει στη σύγκλιση του μοντέλου και στην αμεταβλητότητά του για μικρές

ΚΕΦΑΛΑΙΟ 3. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΆΣΕΙ ΕΙΚΟΝΑΣ

αλλαγές στις τιμές των pixels (θόρυβος).

Μετατροπή έγχρωμης εικόνας σε ασπρόμαυρη ή το αντίστροφο: Σε σενάρια που το χρώμα της εικόνας δε χρησιμεύει στο task που χρησιμοποιείται το dataset, είναι θεμιτή η μετατροπή των εικόνων σε ασπρόμαυρες, ώστε να μειωθεί το υπολογιστικό κόστος των υπολογισμών και να βελτιωθεί η ακρίβεια του μοντέλου. Από την άλλη μεριά, υπάρχουν μοντέλα που αποδεδειγμένα λειτουργούν καλύτερα σε έγχρωμες εικόνες, όπου και χρειάζεται η μετατροπή από ασπρόμαυρες σε έγχρωμες.

Ενίσχυση της αντίθεσης (Contrast Enhancement): Η ενίσχυση της αντίθεσης μπορεί να βοηθήσει το μοντέλο, τονίζοντας χρήσιμα χαρακτηριστικά της εικόνας. Η ισορρόπηση ιστογράμματος είναι μία τεχνική που μπορεί να το επιτύχει αυτό [34].

Μείωση θορύβου: Όταν τα δεδομένα περιέχουν θόρυβο, μπορούν να εφαρμοσθούν φίλτρα για την καταπολέμησή του. Μερικά γνωστά φίλτρα αυτής της κατηγορίας είναι το γκαουσσιανό [35] και το Kalman [36].

Χειρισμός μεταδεδομένων: Σε μερικούς επιστημονικούς κλάδους, όπως για παράδειγμα στον ιατρικό, οι εικόνες μπορούν να συνοδεύονται από μεταδεδομένα που είναι χρήσιμα και πρέπει να ληφθούν υπόψη από τον ερευνητή.

Καθαρισμός: Υπάρχει η περίπτωση το dataset να μην είναι σωστά δομημένο και να περιέχει εικόνες άσχετες με το αντικείμενο της ταξινόμησης, ή εικόνες χαμηλής ποιότητας που θα μπέρδευαν το μοντέλο. Σε τέτοιες περιπτώσεις ο ερευνητής είναι ο υπεύθυνος για να αποφασίσει την διατήρηση ή όχι τους στο dataset.

Συμπίεση εικόνων: Όταν οι υπολογιστικοί πόροι είναι περιορισμένοι και δε μπορούν να υποστηρίξουν το μέγεθος του dataset, μπορεί να γίνει η ανταλλαγή μεταξύ απώλειας πληροφορίας και μείωσης του υπολογιστικού κόστους με τη συμπίεση των εικόνων.

Επαύξηση δεδομένων (Data augmentation): Η τεχνική της επαύξησης δεδομένων έχει αποδείξει πως βοηθάει το μοντέλο στο να γενικεύει καλύτερα σε εικόνες που δεν έχει δει κατά την εκπαίδευσή του, αυξάνοντας συνήθως την ακρίβειά του όταν καλείται να ταξινομήσει άγνωστες εικόνες. Στη πρόσφατη δημοσίευση [37] γίνεται μία εκτενής μελέτη της επίδρασης των μεθόδων επαύξησης στην εκπαίδευση μοντέλων. Επιγραμματικά οι εικόνες μπορούν να υποστούν τυχαία περιστροφή, αντικατοπτρισμό, κλιμάκωση, φιλτράρισμα, προσθήκη θορύβου και αλλαγή φωτεινότητας. Αργότερα, στο κεφάλαιο των πειραμάτων, αναλύεται η επίδραση εφαρμογής augmentations στο σύστημα της διπλωματικής εργασίας.

3.2 DATASETS ΕΙΚΟΝΑΣ

Όπως ήδη αναφέρθηκε, τα datasets που αποτελούνται μόνο από εικόνες και προορίζονται για ανάλυση συναισθήματος είναι περιορισμένα, αφού ποτέ δεν επικεντρώθηκε η επιστημονική κοινότητα στην ανάλυση συναισθήματος αυτουσίως από εικόνες. Έτσι στην ενότητα θα γίνει μια σύνοψη των συνόλων δεδομένων εικόνων, χωρίς να περιορίζονται για ανάλυση συναισθήματος.

CIFAR-10, CIFAR-100: Τα datasets έχουν παρόμοια λογική. Βρίσκονται στην [ιστοσελίδα του πανεπιστημίου του Τορόντο](#). Ο αριθμός υποδηλώνει τον αριθμό των ακλάσεων που περιέχει το κάθε σύνολο. Σε κάθε σετ υπάρχουν 60000 εικόνες που

3.3. ΜΟΝΤΕΛΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΌΡΑΣΗΣ/ ΜΟΝΤΕΛΑ ΕΙΚΟΝΑΣ

δείχνουν αντικείμενα (αεροπλάνο, γάτα, σκύλος, αυτοκίνητο). Ο στόχος του σετ είναι η κατηγοριοποίηση των εικόνων.

ImageNet: Το dataset αυτό [38] αποτελεί το μεγαλύτερο και διασημότερο dataset εικόνων. Χρησιμοποιείται από πολλά μοντέλα για την προεκμάθησή τους, καθώς έχει εκατομμύρια εικόνων γενικής χρήσης. Αποτελεί επίσης το κύριο μέτρο σύγκρισης μεταξύ των διαθέσιμων μοντέλων στην ταξινόμηση εικόνων.

MNIST: Το MNIST [39] θεωρείται το καταλληλότερο dataset εικόνων για την εισαγωγή του ερευνητή στην ταξινόμηση εικόνων, χάρη στο μέγεθος και τη δομή του. Αποτελείται από μια συλλογή χειρόγραφων ψηφίων (0-9) διαστάσεων 28x28.

3.3 ΜΟΝΤΕΛΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΌΡΑΣΗΣ/ ΜΟΝΤΕΛΑ ΕΙΚΟΝΑΣ

Σε συνέχεια του προηγούμενου κεφαλαίου, γίνεται μια αναφορά στα μοντέλα που μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση εικόνων. Τα μοντέλα αυτά γενικότερα είναι χρησιμά για tasks όλου του κλάδου της υπολογιστικής άρασης. Γενικεύοντας, τα μοντέλα που ακολουθούν μπορούν να χωριστούν σε δύο κατηγορίες: τα μοντέλα που βασίζουν την αρχιτεκτονική τους σε συνελικτικό νευρωνικό δίκτυο και τα μοντέλα που βασίζονται στη δομή των transformers.

Συνελικτικά Νευρωνικά Δίκτυα

Ένα συνελικτικό νευρωνικό δίκτυο είναι ένας αλγόριθμος βαθιάς μάθησης που λαμβάνει ως είσοδο εικόνες, τις οποίες καταφέρνει να ξεχωρίζει προσαρμόζοντας τις παραμέτρους του. Τα CNN δεν απαιτούν ιδιαίτερη προεπεξεργασία στην εικόνα εισόδου.

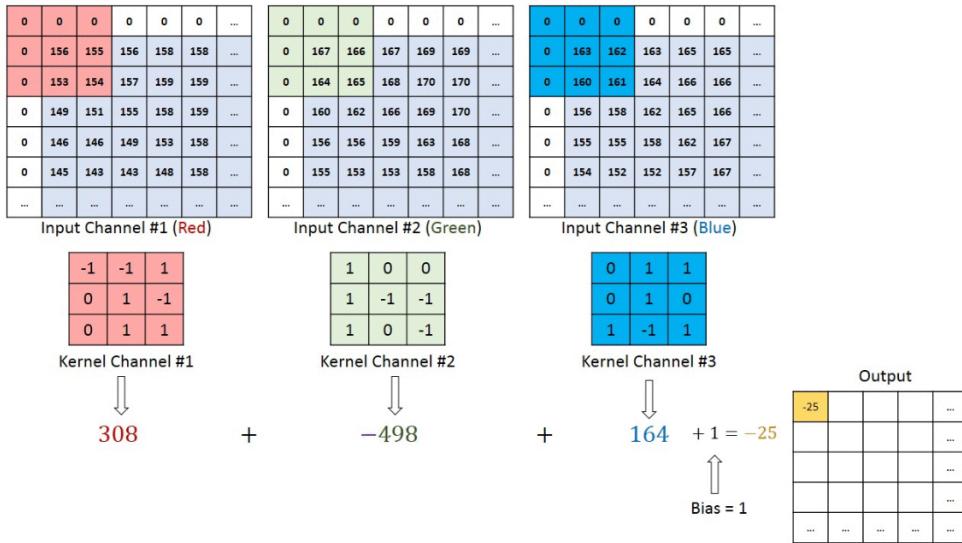
Το κύριο στρώμα τους είναι το στρώμα συνέλιξης. Η εικόνα εισόδου περνάει από το επιλεγμένο φίλτρο (πυρήνας). Η παρακάτω εικόνα ([σχήμα 3.1](#)) αναδεικνύει τη λειτουργία της συνέλιξης με πίνακα, δείχνοντας τον υπολογισμό της εξόδου του συνελικτικού στρώματος για ένα μόνο pixel. Η ίδια διαδικασία επαναλαμβάνεται, μετακινώντας τον πυρήνα σε όλη την εικόνα εισόδου, για όσα κανάλια αυτή έχει.

Έπειτα ακολουθεί το στρώμα συγκέντρωσης (pooling layer). Εδώ επιτυγχάνεται μείωση της διάστασης. Υπάρχουν δύο είδη pooling ([σχήμα 3.2](#)):

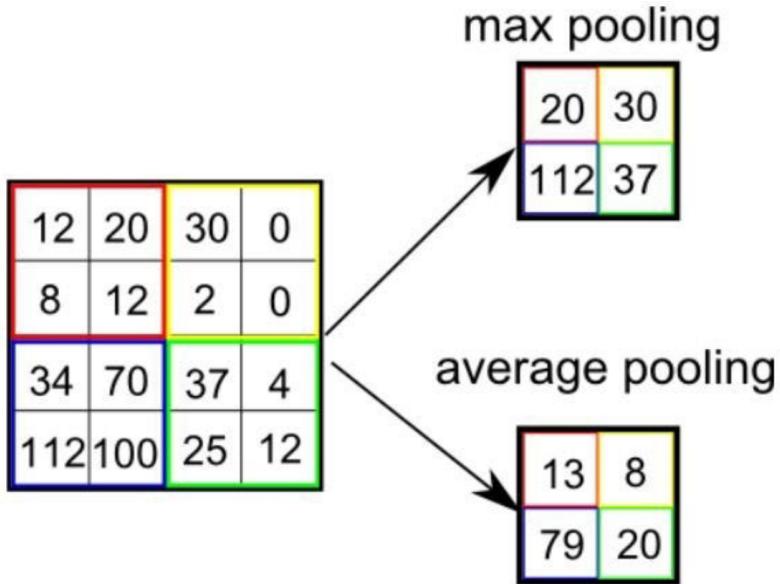
- **Μέγιστη συγκέντρωση (Maximum Pooling):** Από τον πίνακα εισόδου και με μάσκα συγκεκριμένου μεγέθους, επιλέγεται η μεγαλύτερη τιμή.
- **Μέση συγκέντρωση (Average Pooling):** Από τον πίνακα εισόδου και με μάσκα συγκεκριμένου μεγέθους, υπολογίζεται η μέση τιμή.

Στη συνέχεια και για την ολοκλήρωση της διαδικασίας ταξινόμησης της εικόνας εισόδου, ο πίνακας που προκύπτει μετά το στρώμα συγκέντρωσης, μετατρέπεται σε μονοδιάστατος και τροφοδοτεί το fully connected νευρωνικό δίκτυο το οποίο θα καθορίσει την τελική απόφαση, εφαρμόζοντας κάποια τεχνική στους κόμβους εξόδου του (softmax/argmax). Ολόκληρη η διαδικασία συνοφίζεται στο [σχήμα 3.3](#).

Αρχιτεκτονικές που βασίζονται στα CNN και μερικές από αυτές θα μελετηθούν παρακάτω, είναι οι ResNet, ResNext, VGGNet [40], DenseNet, AlexNet [41] και EfficientNet.



Σχήμα 3.1: Συνέλιξη με πίνακα για μέγεθος εικόνας $M \times N \times 3$ και πυρήνα $3 \times 3 \times 3$



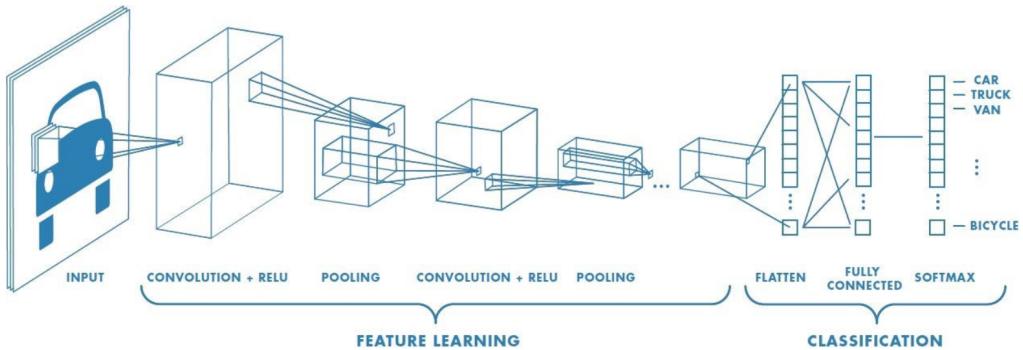
Σχήμα 3.2: Συγκέντρωση με πίνακα εισόδου 4×4 και μάσκα 2×2

ResNet

Καθώς χρησιμοποιούνταν ολοένα και μεγαλύτερα CNN, με περισσότερα στρώματα συνέλιξης, παρατηρήθηκε το φαινόμενο η κλίση να μηδενίζεται ή να απειρίζεται, με αποτέλεσμα η προσθήκη περισσότερων στρωμάτων να αυξάνει το σφάλμα αντί να το μειώνει.

Για την επίλυση του προβλήματος αυτού εμφανίσθηκε το ResNet (*Residual Network*) [42]. Στο ResNet εφαρμόζεται η τεχνική της προσπέρασης συνδέσεων. Η έξοδος από τη συνάρτηση ενεργοποίησης ενός στρώματος δε συνδέεται απευθείας με το επόμενο συνελικτικό στρώμα, αλλά προσπερνάει έναν αριθμό από αυτά. Στο [σχήμα 3.4](#)

3.3. ΜΟΝΤΕΛΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΌΡΑΣΗΣ/ ΜΟΝΤΕΛΑ ΕΙΚΟΝΑΣ



Σχήμα 3.3: Ταξινόμηση με CNN 2 στρωμάτων συνέλιξης

φαίνεται η αρχιτεκτονική του μοντέλου με 34 συνελικτικά στρώματα, καθώς και η λειτουργία που περιγράφτηκε σε αυτή τη παράγραφο.

Εκτός από το μοντέλο με τα 34 στρώματα, δημοσιεύτηκαν ταυτόχρονα και μοντέλα με 50,101 και 152 στρώματα.

ResNext

Το δίκτυο *ResNeXt* [43] διατηρεί τη λογική του *ResNet* με τις προσπεράσεις συνδέσεων. Ωστόσο εισάγει τον όρο *cardinality*, που αντιπροσωπεύει το μέγεθος του σετ μετασχηματισμών. Στο [σχήμα 3.5](#) συγκρίνεται ένα μπλοκ της αρχιτεκτονικής στα δίκτυα *ResNet* και *ResNeXt* (με *cardinality*=32). Στο εσωτερικό του μπλοκ, τα μεγέθη των φίλτρων που χρησιμοποιούνται είναι κοινά. Οι δύο αρχιτεκτονικές που δημοσιεύθηκαν είχαν 50 και 101 στρώματα.

DenseNet

Άλλο ένα CNN deep learning δίκτυο είναι το *DenseNet* [44]. Εισάγεται η έννοια του πυκνού μπλοκ (dense block). Αποτελείται από στρώματα ενός απλού συνελικτικού δικτύου, τα οποία όμως λαμβάνουν πληροφορία από την έξοδο των προηγούμενων στρωμάτων ([σχήμα 3.6](#)).

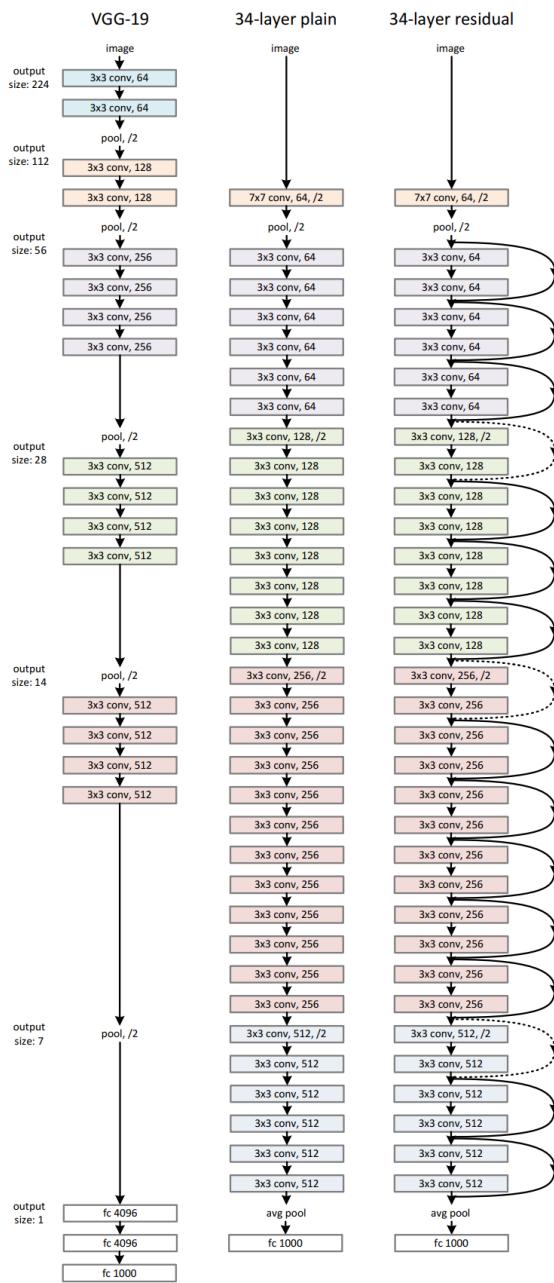
Συνολικά, η αρχιτεκτονική του *DenseNet* εμπεριέχει κάποια συνελικτικά στρώματα, στρώματα συγκέντρωσης (pooling) και τα πυκνά μπλοκ που αναφέρθηκαν. Αναπαράσταση του *DenseNet* δίνεται στο [σχήμα 3.7](#).

Το δίκτυο που αξιοποιεί τα πυκνά μπλοκ επιτυγχάνει την εύκολη οπισθοδιάδοση (backpropagation). Επιπρόσθετα μειώνει σημαντικά τον αριθμό των παραμέτρων και το υπολογιστικό κόστος, συγκριτικά με ένα *ResNet* δίκτυο.

Παρέχονται 4 αρχιτεκτονικές *DenseNet*, με αριθμούς στρωμάτων 121, 161, 169 και 201.

EfficientNet

Το τελευταίο δίκτυο που θα αναλυθεί από την οικογένεια των CNN είναι το *EfficientNet* [45]. Το δίκτυο εστιάζει στον τρόπο κλιμάκωσης του μοντέλου. Έτσι πραγματοποιεί μια σύνθετη κλιμάκωση (compound scaling), η οποία ονομάζεται σύνθετη διότι συνδυάζει ταυτόχρονα κλιμάκωση σε 3 άξονες, στο πλάτος του δικτύου, στο βάθος του

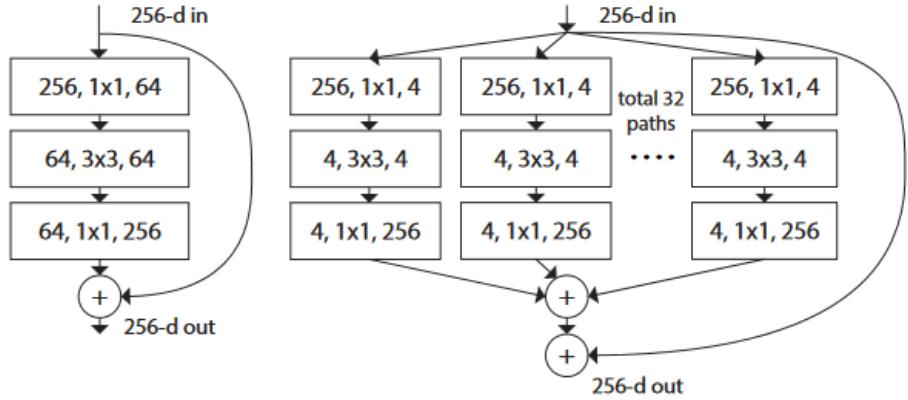


Σχήμα 3.4: Αριστερά: VGG-19 (παλιότερο μοντέλο), κέντρο: Αρχιτεκτονική ResNet χωρίς προσπέραση συνδέσεων, δεξιά: Αρχιτεκτονική ResNet

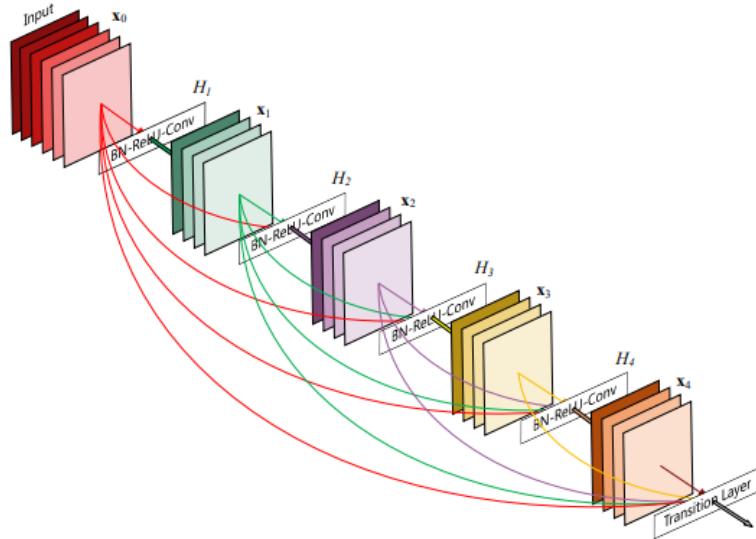
και στην ανάλυση της εικόνας εισόδου. Το πλάτος αναφέρεται στον αριθμό των κοναλιών που χρησιμοποιεί το δίκτυο σε κάθε στρώμα του, ενώ ως βάθος εννοείται ο αριθμός των στρωμάτων του. Η κλιμάκωση οδηγείται από έναν συντελεστή κλιμάκωσης φ.

Συνολικά δημοσιεύονται 8 αρχιτεκτονικές. Η βασική αρχιτεκτονική ονομάζεται EfficientNet-B0 και φαίνεται στο [σχήμα 3.8](#). Οι υπόλοιπες αρχιτεκτονικές προέκυψαν από την αρχική, αυξάνοντας την τιμή του συντελεστή φ ώστε να παραχθούν μεγαλύτερα δίκτυα. (EfficientNet-B1 ως και EfficientNet-B7).

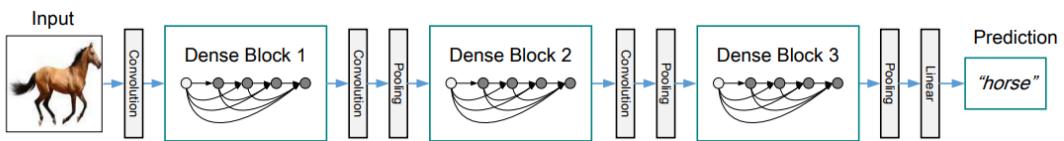
3.3. ΜΟΝΤΕΛΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΌΡΑΣΗΣ/ ΜΟΝΤΕΛΑ ΕΙΚΟΝΑΣ



Σχήμα 3.5: Αριστερά: Μπλοκ του ResNet, δεξιά: μπλοκ του ResNeXt

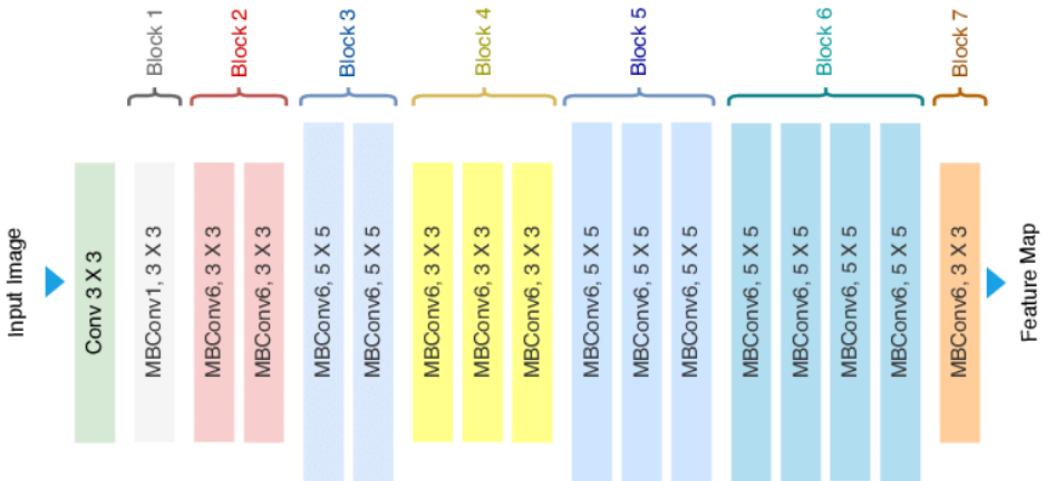


Σχήμα 3.6: Πυκνό Μπλοκ 5 στρωμάτων



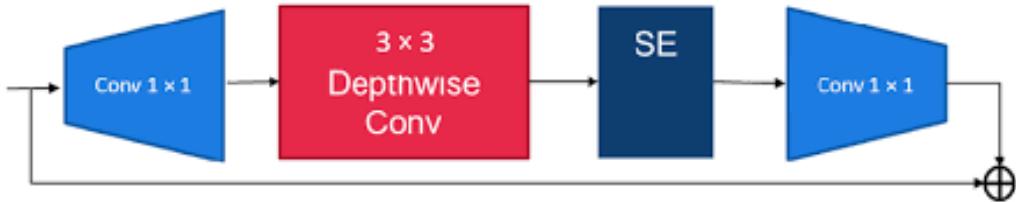
Σχήμα 3.7: Αρχιτεκτονική του DenseNet με 3 πυκνά μπλοκ

Στην αρχιτεκτονική αυτή χρησιμοποιείται σαν κύριο συνελικτικό μπλοκ μια παράλλαγη του απλού συνελικτικού στρώματος, η οποία ονομάζεται MBConv και προτάθηκε για πρώτη φορά στη δημοσίευση του MobileNetV2 [46]. Κάθε τέτοιο μπλοκ αποτελείται από μία σημειακή συνέλιξη για την αύξηση του αριθμού των καναλιών, μία κλασσική συνέλιξη, το SE μπλοκ και μία σημειακή συνέλιξη στο τέλος για την επαναφορά του αριθμού των καναλιών στην αρχική τους τιμή (σχήμα 3.9).



Σχήμα 3.8: Αρχιτεκτονική του EfficientNet-B0

Το SE (Squeeze and Excitation) χρησιμοποιείται ώστε να βοηθήσει το μοντέλο να εστιάζει στα σημαντικά χαρακτηριστικά.



Σχήμα 3.9: MBCConv μπλοκ

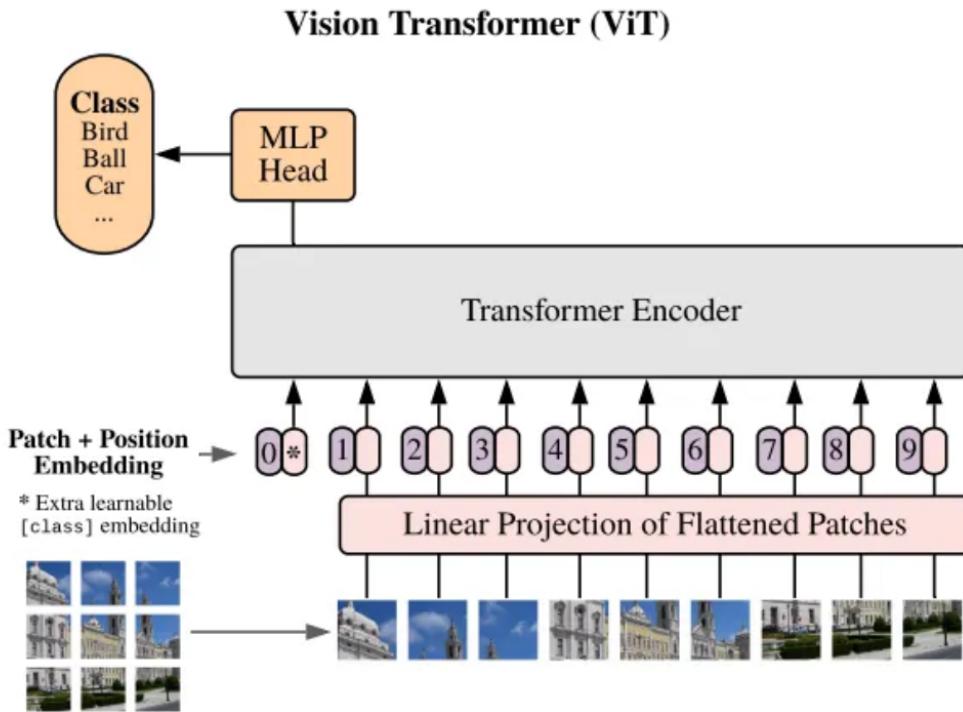
Μοντέλα εικόνας βασισμένα σε transformers

Προηγουμένως αναδείχθηκε η αξία των transformers για εφαρμογές της επεξεργασίας φυσικής γλώσσας. Ωστόσο η επιστημονικές εξελίξεις ανέδειξαν τις ικανότητες των transformers και στον τομέα της εικόνας.

ViT

Η δημοσίευση του ViT (*Vision Transformer*) [47] έφερε επανάσταση και στον κλάδο της υπολογιστικής όρασης, επιτρέποντας τη λειτουργία των transformers και σε εικόνες. Το ViT κατάφερε να ξεπεράσει σε επιδόσεις τα δίκτυα που βασιζόταν σε CNN. Τα ViT δίκτυα μπορούν να χρησιμοποιηθούν σε διάφορα tasks που βασίζονται σε εικόνες, όπως είναι η ταξινόμηση εικόνων, η αναγνώριση αντικειμένων, η τμηματοποίηση εικόνων και η κατανόηση χώρου.

Οι transformers χρειάζονται μία ακολουθία ως είσοδο. Για να πραγματοποιηθεί αυτό όταν η είσοδος είναι μία εικόνα, η εικόνα χωρίζεται σε τετραγωνικές περιοχές μεγέθους 16x16 pixels, οι οποίες ονομάζονται patches (σχήμα 3.10).



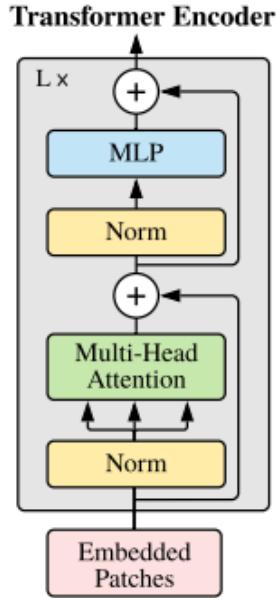
Σχήμα 3.10: Χωρισμός της εικόνας σε patches

Μετά τον χωρισμό σε patches, αυτά τοποθετούνται στη σειρά, σαν μια ακολουθία. Γίνεται η πράξη του εσωτερικού γινομένου για κάθε patch με έναν embedding πίνακα, σε αντιστοίχιση με τα embeddings που χρησιμοποιούνται στο BERT. Προκύπτει λοιπόν μια ακολουθία η οποία ολοκληρώνεται με την προσθήση της πληροφορίας θέσης. Έπειτα η τελική ακολουθία τροφοδοτεί τον κωδικοποιητή του transformer (σχήμα 3.11).

Αρχικά δημοσιεύτηκαν 4 μοντέλα ViT, διαφέροντας στο μέγεθός τους. Αυτά ονομάζονται ViT-small, ViT-base, ViT-large και ViT-huge.

Χρήσιμη για τη σύγκριση μεταξύ του ViT και των CNN αρχιτεκτονικών είναι η δημοσίευση από την Google Research [48]. Οι ερευνητές καταλήγουν στο συμπέρασμα πως ο τρόπος λειτουργίας των δικτύων παρουσιάζει σημαντικές διαφορές:

- Οι αναπαραστάσεις μεταξύ των "ρηχότερων" (αρχικών) και βαθύτερων στρωμάτων μοιάζουν περισσότερο στο ViT απότι στα CNN. Πιο συγκεκριμένα, στα CNN το κάθε στρώμα μπορεί να αναγνωρίσει κάποια χαρακτηριστικά της εικόνας. Ενδεικτικά, το πρώτο στρώμα εντοπίζει τα χαμηλού επιπέδου χαρακτηριστικά (χρώματα, ακμές), ενώ τα βαθύτερα στρώματα μπορούν να εντοπίσουν αντικείμενα και ολόκληρες σκηνές. Κάτι αντίστοιχο δε συμβαίνει στο ViT.
- Η μέθοδος της προσπέρασης συνδέσεων που χρησιμοποιείται στα ViT και σε μερικά CNN μοντέλα, όπως το ResNet, έχει εντονότερη επιρροή στα ViT.
- Τα ViT μοντέλα καταφέρνουν να διατηρούν περισσότερη χωρική πληροφορία.



Σχήμα 3.11: Κωδικοποιητής του transformer εικόνας

- Τα στρώματα του μηχανισμού προσοχής (attention) που περιλαμβάνει ο κωδικοποιητής ευνοούν το μοντέλο στο να λαμβάνει υπόψη την όλη την εικόνα και όχι μόνο τη γειτονιά ενός σημείου, όπως συμβαίνει στα CNN.

DINO

Ακολούθησε η δημοσίευση της Facebook AI Research το 2021, όπου προτείνεται ένα νέο μοντέλο το οποίο ακολουθεί την αυτο-εποπτευόμενη εκμάθηση (self-supervised learning) για την εκμάθηση του transformer, με την ονομασία *DINO* (*self-distillation with no labels*) [49].

Η διαδικασία της εκμάθησης ξεκινάει από δύο ViT μοντέλα, ίδιας αρχιτεκτονικής με διαφορετικές παραμέτρους, όπου το ένα ονομάζεται μαθητής και το άλλο καθηγητής. Τα βάρη του καθηγητή ανανεώνονται με βάση τα βάρη του μαθητή. Έτσι συμβολίζοντας ως θ_t το βάρος του καθηγητή και θ_s το βάρος του μαθητή, η ανανέωση γίνεται με τον τύπο:

$$\theta_t = \lambda\theta_t + (1 - \lambda)\theta_s$$

Το λ είναι μια παράμετρος με μορφή ημιτόνου, παίρνοντας τιμές από 0.996 εώς 1 και κάνοντας ανανέωση της τιμής όταν αλλάζει η εποχή κατά την εκπαίδευση.

Από κάθε εικόνα, αποκόπτονται τυχαία διάφορα τμήματα. Όσα έχουν μέγεθος μικρότερο από τη μισή εικόνα ονομάζονται τοπικές όψεις, ενώ όσα ξεπερνούν τη μισή εικόνα ονομάζονται ολικές όψεις. Κατά την εκπαίδευση, όλες οι τομές της εικόνας εκπαιδεύονται το μοντέλο-μαθητή, αλλά το μοντέλο-καθηγητή εκπαιδεύεται μόνο στις ολικές όψεις. Ο στόχος της εκπαίδευσης είναι το μοντέλο-μαθητή να μπορεί να ακολουθεί το μοντέλο-καθηγητή, παρόλο που βλέπει και εικόνες περιορισμένου μεγέθους συγκριτικά με την αρχική.

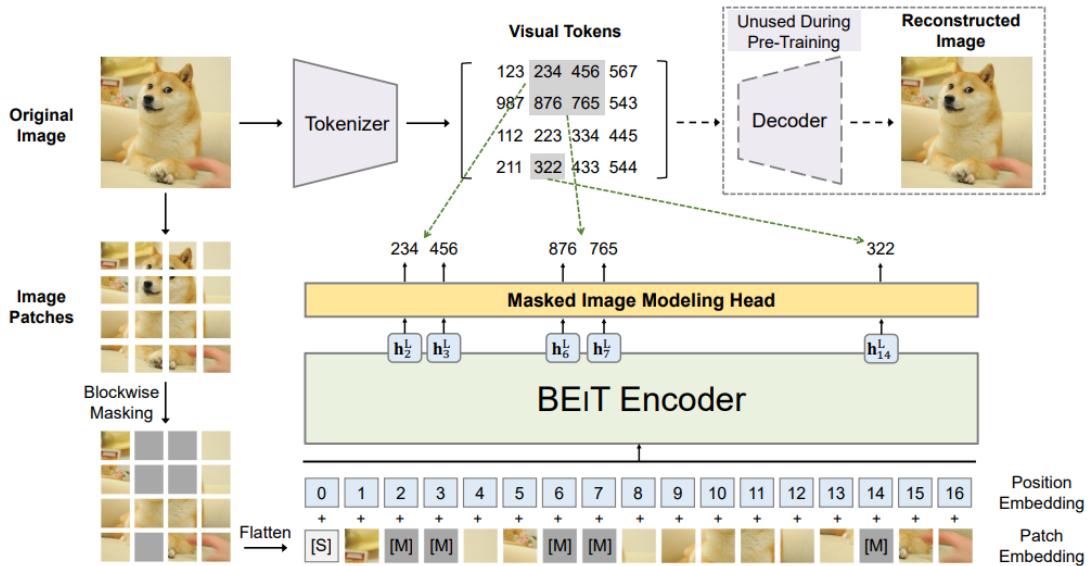
Δημοσιεύτηκαν 2 παραλλαγές του μοντέλου:

3.3. ΜΟΝΤΕΛΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΌΡΑΣΗΣ/ ΜΟΝΤΕΛΑ ΕΙΚΟΝΑΣ

1. dino-vits16: Χρησιμοποιεί το ViT-small κατά την εκμάθηση
2. dino-vitb16: Χρησιμοποιεί το ViT-base κατά την εκμάθηση

BEiT

Το *BEiT* (*Bidirectional Encoder representation from Image Transformers*) [50] δημοσιεύθηκε από την ερευνητική ομάδα Microsoft Research το 2022. Για την προεκμάθηση του μοντέλου, η κάθε εικόνα χωρίζεται σε patches, με τον ίδιο τρόπο που πραγματοποιείται και στο ViT. Εκτός αυτού, προκύπτει και ένας δεύτερος πίνακας δύο διαστάσεων, ο οποίος περιέχει τα tokens που υπολογίζονται από έναν tokenizer. Το κάθε token μπορεί να λάβει συγκεκριμένες τιμές, που προέρχονται από ένα λεξικό εικόνων. Άρα λοιπόν ο tokenizer ουσιαστικά αντιστοιχίζει τις τιμές από όλα τα κανάλια εισόδου σε έναν δισδιάστατο πίνακα, σύμφωνα με τις οδηγίες του χρησιμοποιούμενου λεξικού. Ο δισδιάστατος πίνακας που προκύπτει έχει ίδια διάσταση με το πλήθος των patches. Ο πίνακας των tokens δίνεται έπειτα ως είσοδος στον αποκωδικοποιητή, ο οποίος έχει ως στόχο να επαναδημιουργήσει την αρχική εικόνα.



Σχήμα 3.12: Διαδικασία προ-εκμάθησης του BEiT

Για την εκμάθηση του μοντέλου χρησιμοποιείται η τεχνική της μάσκας, όπως είχε αναφερθεί και στη περίπτωση του BERT. Εδώ επιλέγονται περίπου 40% από τα patches να καλυφθούν από κάποια μάσκα. Έτσι το μοντέλο κάνει προβλέψεις στα tokens των patches που λείπουν και ο στόχος της εκπαίδευσης είναι οι προβλέψεις αυτές να είναι όσο το δυνατόν πλησιέστερες στις πραγματικές.

Ο transformer που χρησιμοποιείται είναι το μοντέλο ViTBase. Συνολικά 4 μοντέλα είναι διαθέσιμα:

1. BeiT-base-224
2. BeiT-large-224

ΚΕΦΆΛΑΙΟ 3. ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΆΣΕΙ ΕΙΚΟΝΑΣ

3. Beit-base-384

4. Beit-large-384

Ο αριθμός που βρίσκεται στο όνομα του μοντέλου αναφέρεται στην ανάλυση των εικόνων στις οποίες εκπαιδεύτηκε.

4

Πολυτροπικό σύστημα ανάλυσης συναισθημάτων

Στα προηγούμενα κεφάλαια αναλύθηκαν εκτενώς τα δύο είδη πληροφορίας που καλύπτονται στη διπλωματική εργασία: το κείμενο και η εικόνα.

Μια παρόμοια συλλογιστική πορεία θα ακολουθήσουμε και εδώ, όπου θα εξερευνηθούν αρχικά οι μέθοδοι συνένωσης των δύο modalities και στη συνέχεια θα γίνει αναφορά σε datasets που περιλαμβάνουν ταυτόχρονα περισσότερα του ενός modalities.

Στη βιβλιογραφία υπάρχουν αρκετές δημοσιεύσεις που ερευνούν τα παραπάνω θέματα [51], δηλαδή την συνένωση των χαρακτηριστικών από τα διαφορετικά modalities και τα διαθέσιμα datasets.

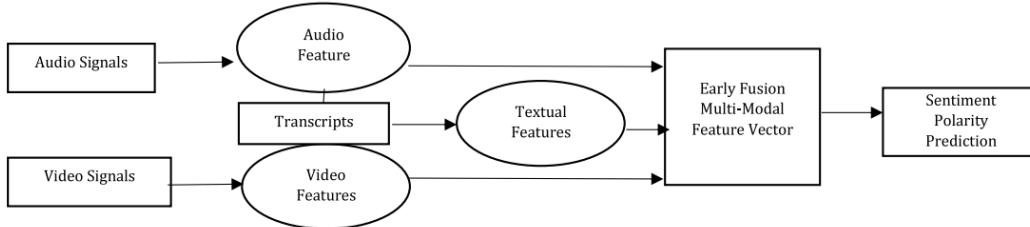
4.1 ΣΥΝΕΝΩΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Σε ένα πολυτροπικό σύστημα όπου ο ερευνητής καλείται να διαχειριστεί είσοδο από διάφορα modalities, οφείλει να βρει έναν τρόπο να ενοποιήσει τις πληροφορίες που εξάγει από το κάθε modality. Η διαδικασία αυτή στη βιβλιογραφία ονομάζεται *fusion* και εφεξής θα χρησιμοποιείται αυτός ο όρος στη διπλωματική εργασία. Η μέθοδος συνένωσης των χαρακτηριστικών παρουσιάζει αξιοσημείωτο ενδιαφέρον και υπάρχουν αρκετές επιλογές οι οποίες θα αναφερθούν παρακάτω.

Early Fusion/Fusion στο επίπεδο των χαρακτηριστικών: Στη κατηγορία αυτή, τα χαρακτηριστικά από τα διαφορετικά modalities συνενώνονται και σχηματίζουν έναν ενιαίο πίνακα χαρακτηριστικών. Αφού σχηματιστεί ο πίνακας, δίνεται σαν είσοδος σε έναν αλγόριθμο ταξινόμησης. Το πλεονέκτημα της συνένωσης στο επίπεδο των χαρακτηριστικών είναι πως συσχετίζει από νωρίς τα διαφορετικά modalities, οδηγώντας συνήθως σε υψηλότερες επιδόσεις, αφού διατηρείται όλη η πληροφορία τους. Το μεγάλο μειονέκτημα εδώ είναι πως οι μορφές των δεδομένων διαφέρουν

ΚΕΦΑΛΑΙΟ 4. ΠΟΛΥΤΡΟΠΙΚΟ ΣΥΣΤΗΜΑ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΩΝ

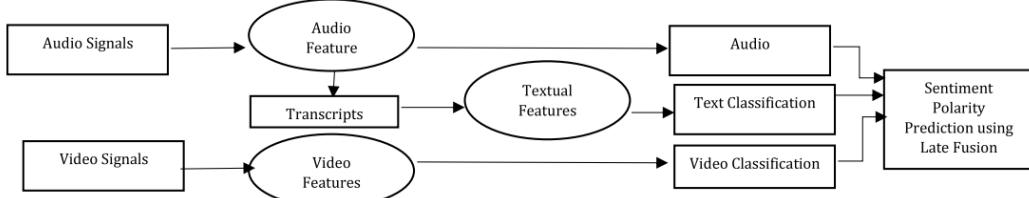
σημαντικά μεταξύ τους και θα πρέπει να γίνει ισχυρή προεπεξεργασία τους, προτού καταστεί εφικτή η συνένωσή τους. Το παρακάτω σχήμα αναδεικνύει την πρώιμη συνένωση σε ένα πολυτροπικό σύστημα εικόνας και ήχου ([σχήμα 4.1](#)).



Σχήμα 4.1: Early Fusion

Η δημοσίευση [52] είναι ενδεικτική της χρήσης του early fusion, όπου επικεντρώνεται στην ανάλυση της πειστικότητας της εισόδου κειμένου, εικόνας και ήχου.

Late Fusion/ Fusion στο επίπεδο της απόφασης: Η σημαντική διαφοροποίηση στην συνένωση στο επίπεδο της απόφασης, είναι πως η συνένωση γίνεται αφού υπολογιστούν οι προβλέψεις του κάθε modality. Προηγείται δηλαδή η ταξινόμηση και έπειτα η συνένωση των προβλέψεων. Οι περισσότεροι ερευνητές επιλέγουν τη συγκεκριμένη τεχνική, λόγω της ευκολίας της στην υλοποίηση. Ακόμη ένα πλεονέκτημα αυτής της τεχνικής είναι πως το μοντέλο για κάθε είδος πληροφορίας μπορεί να αντικατασταθεί εύκολα, χωρίς να επηρεάσει το υπόλοιπο σύστημα. Η τεχνική παρουσιάζει δυσκολία όταν χρειάζεται να χρησιμοποιηθούν διαφορετικοί classifiers για το κάθε modality. Εποπτικά το late fusion δίνεται στο [σχήμα 4.2](#).



Σχήμα 4.2: Late Fusion

Ενδεικτικά ο αναγνώστης μπορεί να παραπεμφθεί στο [53], όπου χρησιμοποιείται late fusion στη προσπάθεια ανάλυσης συναισθήματος από πολυτροπικά tweets.

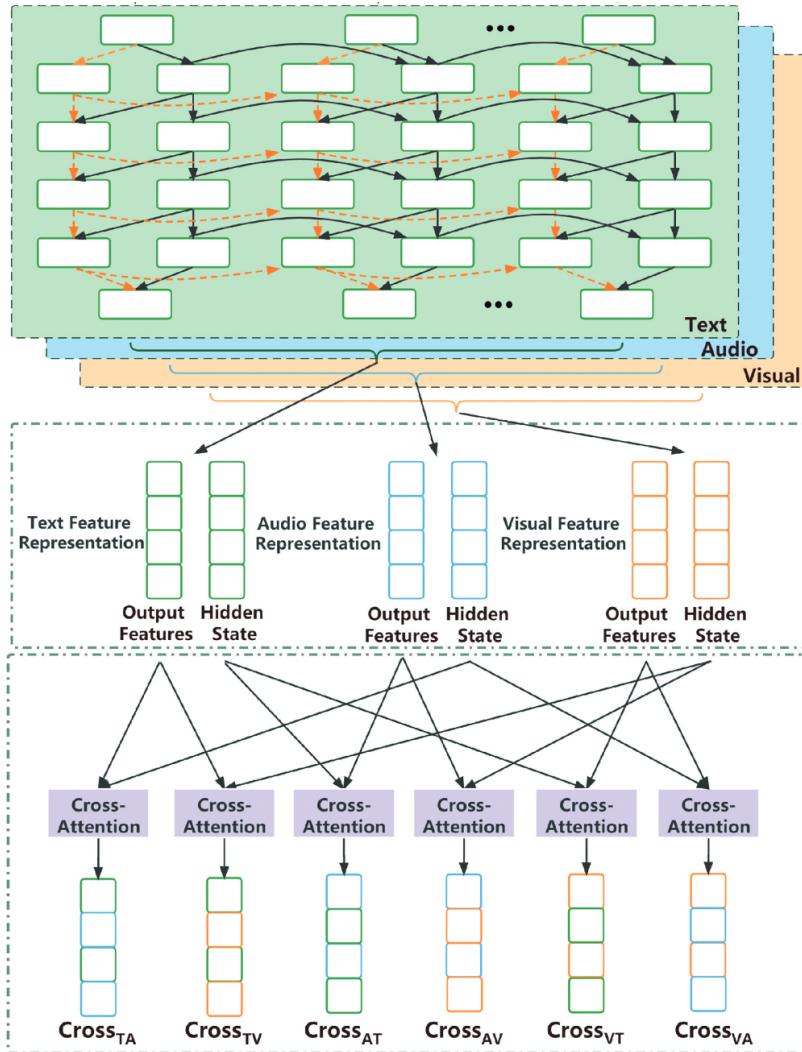
Hybrid Fusion: Η τεχνική λέγεται υβριδική καθώς συνδυάζει το early και το late fusion, με στόχο να εκμεταλλευτεί ταυτόχρονα τα οφέλη των δύο μεθόδων. Στη δημοσίευση [54] χρησιμοποιείται η υβριδική προσέγγιση σε πολυτροπικό σύστημα ανάλυσης συναισθήματος σε κριτικές ανεβασμένες στο YouTube.

Fusion στο επίπεδο του μοντέλου: Αναζητείται η συσχέτιση μεταξύ των δεδομένων των διαφορετικών modalities. Έπειτα κατασκευάζεται το μοντέλο που θα συνενώσει τα δεδομένα. Αυτή η τεχνική δεν είναι ιδιαίτερα διαδεδομένη. Στο [55] μπορεί και βρίσκεται εφαρμογή καθώς εμπλέκονται σαν modalities ο ήχος και το βίντεο, που έχουν άμεση συσχέτιση μεταξύ τους.

Tensor Fusion: Η τεχνική αναδεικνύεται στη δημοσίευση [56]. Γίνεται προσπάθεια να αξιοποιηθούν όλα τα modalities και να συνδυαστούν μεταξύ τους με

4.1. ΣΥΝΕΝΩΣΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

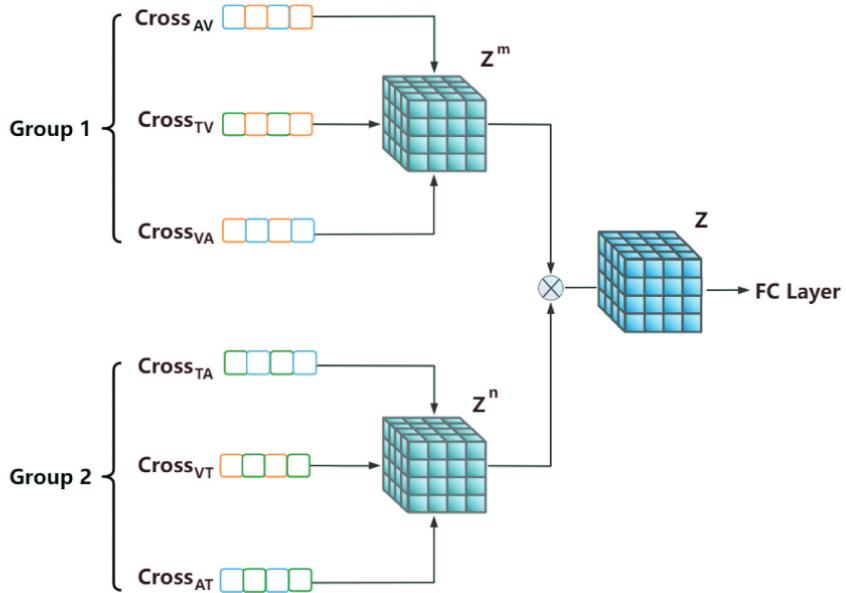
τον κάθε εφικτό τρόπο. Στη συγκεκριμένη δημοσίευση που αντιμετωπίζει 3 διαφορετικά modalities, το [σχήμα 4.3](#) δείχνει τον συνδυασμό των modalities αφού έχουν εξαχθεί τα χαρακτηριστικά τους, ενώ το [σχήμα 4.4](#) δείχνει σχηματικά τη λειτουργία του tensor fusion.



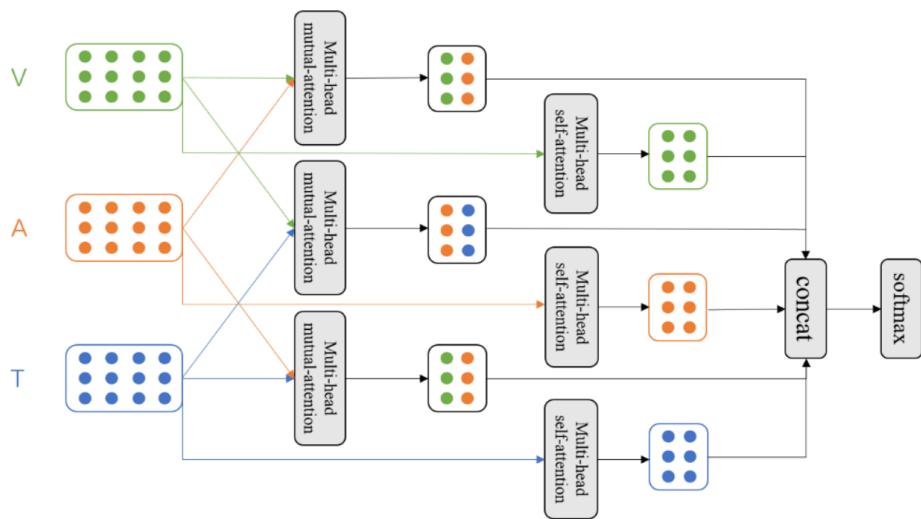
Σχήμα 4.3: Συνδυασμός των modalities

Fusion που βασίζεται στο μηχανισμό του attention: Ο μηχανισμός attention μπορεί να αξιοποιηθεί και στο fusion των διαφορετικών modalities ώστε να παροχθεί ένα καλύτερο αποτέλεσμα. Στο paper [57] χρησιμοποιείται η μέθοδος αυτή σε 3 modalities (video, audio, text). Εφαρμόζεται attention σε καθένα από τα 3 modalities αλλά και στους 3 συνδυασμούς που προκύπτουν παίρνοντας τα πιθανά ζεύγη τους, όπως φαίνεται στο [σχήμα 4.5](#).

Στη βιβλιογραφία γίνεται αναφορά και σε άλλες κατηγορίες τεχνικών fusion (fusion στο επίπεδο λέξεως, συνένωση ιεραρχίας) οι οποίες όμως δε θα αναλυθούν στη παρούσα διπλωματική εργασία λόγω της περιορισμένης χρήσης τους.



Σχήμα 4.4: Tensor Fusion



Σχήμα 4.5: Attention Fusion

4.2 ΠΟΛΥΤΡΟΠΙΚΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

Στη μέχρι τώρα ανάλυση εξερευνήθηκαν datasets κειμένου και datasets εικόνων. Υπάρχουν ωστόσο datasets που συνδυάζουν τα παραπάνω modalities, ενώ μπορεί να περιέχουν και δεδομένα ήχου, καθώς και δεδομένα βίντεο. Εδώ εστιάζουμε μόνο σε datasets που δημιουργήθηκαν για ανάλυση συναισθήματος.

1. **CMU-MOSEI [58]:** Ένα πολυτροπικό dataset ανάλυσης συναισθήματος, που περιέχει κείμενο, εικόνα και ήχο (3 modalities). Το CMU-MOSEI περιέχει βί-

ντεο ηθοποιών που εκφράζουν συναισθήματα. Περιέχει 3228 βίντεο, από 1000 διαφορετικούς ομιλητές, τα οποία λήφθηκαν από το youtube.

2. **MELD** [59]: Το dataset στοχεύει στην εκπαίδευση μοντέλου που θα είναι ικανό να αναγνωρίζει συναισθήματα σε μία συζήτηση. Επίσης περιέχει τα ίδια 3 modalities με το CMU-MOSEI, όπου οι διάλογοι συλλέχθηκαν από την τηλεοπτική σειρά "Τα Φιλαράκια".
3. **FACTIFY** [60]: Ο σκοπός του είναι η διασταύρωση πληροφοριών. Αποτελείται από 50000 δείγματα εικόνας και κειμένου. Οι πληροφορίες συλλέχθηκαν από το Twitter ειδησεογραφικών καναλιών της Ινδίας και της Αμερικής. Η πλειοψηφία των δεδομένων αναφέρονται σε κυβερνητικά και πολιτικά ζητήματα.
4. **MEMOTION 2** [61]: Αποτελεί συλλογή από memes, μεγέθους 10000. Κάθε δείγμα περιέχει εικόνα και κείμενο. Τα κύρια θέματα ενδιαφέροντος των memes είναι ο αθλητισμός, η πολιτική και η θρησκεία. Για κάθε meme υπάρχει ταξινόμηση ως προς το συναίσθημα (θετικό, αρνητικό, ουδέτερο) και το emotion (χιουμοριστικό, επιθετικό, σαρκαστικό, παρακινητικό).

5

Γλωποιήσεις

Προηγήθηκε η ανάλυση και επεξήγηση των μοντέλων και των τεχνικών που θα χρησιμοποιηθούν στις υλοποιήσεις της διπλωματικής που δίνονται στο κεφάλαιο αυτό. Αρχικά θα σημειωθεί το λογισμικό που χρησιμοποιήθηκε (εργαλεία, βιβλιοθήκες κτλ.) και στη συνέχεια το σύνολο των αρχιτεκτονικών που δοκιμάστηκαν.

5.1 ΛΟΓΙΣΜΙΚΟ

Η διπλωματική εργασία υλοποιήθηκε σχεδόν απόλυτα με χρήση της γλώσσας προγραμματισμού *Python*. Μόνο για την υλοποίηση της ιστοσελίδας χρησιμοποιήθηκαν επίσης οι γλώσσες *HTML/CSS* και περιορισμένα η γλώσσα προγραμματισμού *Javascript*.

Η ενότητα ενημερώνει τον αναγνώστη για το λογισμικό που χρησιμοποιήθηκε και για τους σκοπούς που εξυπηρετεί. Επίσης γίνεται μια σύντομη αναφορά στις βιβλιοθήκες της *Python* που αξιοποιήθηκαν, συνοδευόμενες από τον σύνδεσμο για το documentation τους.

- **Visual Studio Code:** Το Visual Studio Code είναι ένα πρόγραμμα επεξεργασίας κώδικα με υποστήριξη για λειτουργίες ανάπτυξης όπως τον εντοπισμό σφαλμάτων, την εκτέλεση εργασιών και τον έλεγχο έκδοσης. Στοχεύει να παρέχει μόνο τα εργαλεία που χρειάζεται ένας προγραμματιστής για έναν γρήγορο κύκλο δημιουργίας κώδικα-εντοπισμού σφαλμάτων και αφήνει πιο σύνθετες ροές εργασίας σε πληρέστερα IDE. Είναι γρήγορο, εύχρηστο και παρέχει υποστήριξη για τη μεγάλη πλειοφηφία των προγραμματιστικών γλωσσών με τις επεκτάσεις που διαθέτει. Το Visual Studio Code διατίθεται στα λειτουργικά συστήματα *macOs*, *Linux* και *Windows*.
- **Google Colab:** Λογισμικό που παρέχεται από το τμήμα έρευνας της Google (Google Research). Επιτρέπει την γραφή κώδικα σε *Python* υπό τη μορφή ση-

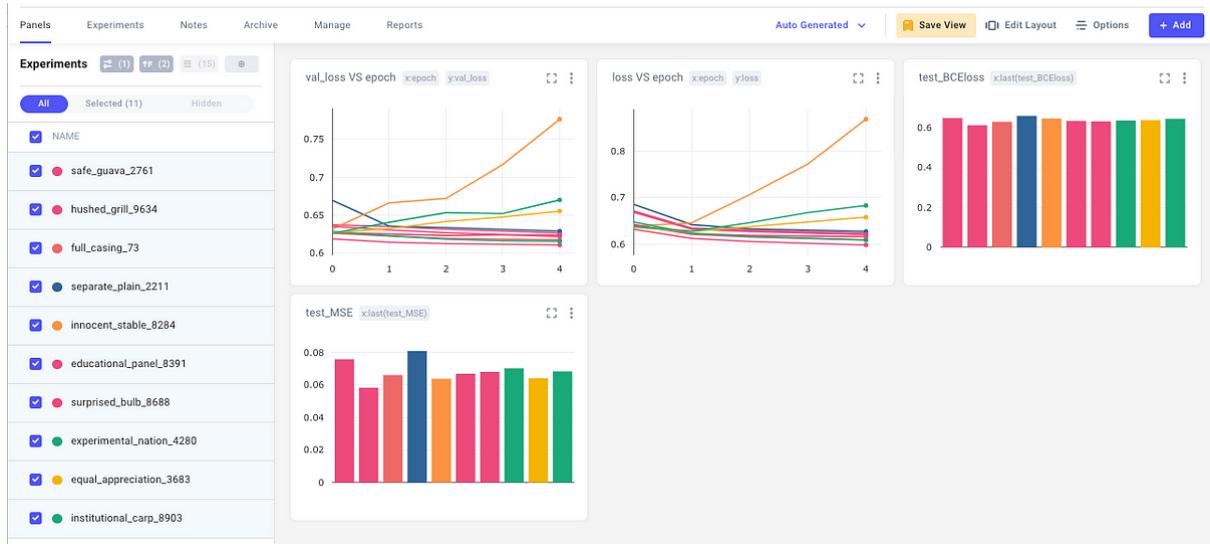
ΚΕΦΑΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

μειωματαρίου. Βρίσκει ευρεία εφαρμογή στη μηχανική μάθηση και στην ανάλυση δεδομένων για εκπαιδευτικούς σκοπούς. Η μορφή του σημειωματαρίου επιτρέπει στον χρήστη να γράφει τον κώδικα σε ξεχωριστά κελιά, βοηθώντας τον να διακριτοποιεί τα κομμάτια του κώδικα. Επιπλέον, ο χρήστης μπορεί να προσθέτει κείμενο ανάμεσα από τα κομμάτια κώδικα. Τα παραπάνω στοιχεία βοηθούν στη γραφή κατανοητού κώδικα και στη διαδικασία του debugging. Παρόλο που η δωρέαν έκδοση του λογισμικού έχει χρονικούς περιορισμούς, επιτρέπει το χρήστη να χρησιμοποιεί για τον κώδικα hardware της Google το οποίο ο ίδιος μπορεί να μη διαθέτει στον τοπικό υπολογιστή του (CPU, RAM, GPU).

- Kaggle Notebook: Η λειτουργία του είναι παρόμοια με το Google Colab. Παρέχεται και αυτό από την Google, ωστόσο τα δύο λογισμικά έχουν τα δικά τους πλεονεκτήματα και μειονεκτήματα. Το Colab είναι εύκολο στη χρήση του και δίνει στο χρήστη τη δυνατότητα να διασυνδέσει το notebook με το Google Drive του και το Github, για το κατέβασμα/ανέβασμα εγγράφων, δεδομένων. Το Kaggle από την άλλη βασίζεται περισσότερο στη κοινότητά του. Συνοδεύεται επίσης από πολλά datasets που μπορεί ο χρήστης να πειραματιστεί. Διατηρεί κιόλας ιστορικό από το notebook ώστε να μπορεί να γίνει μια αναδρομή σε προηγούμενες εκδόσεις.
- Hugging Face: Είναι μία **ιστοσελίδα** όπου βρίσκονται συγκεντρωμένα μοντέλα και datasets, τα οποία μπορούν να συνδεθούν απευθείας με την Python με τη χρήση κατάλληλης βιβλιοθήκης και να φορτωθούν στον κώδικα. Βασίζεται στη κοινότητά του, αφού ο κάθε χρήστης έχει τη δυνατότητα να ανεβάσει το δικό του μοντέλο. Τα διαθέσιμα μοντέλα δεν περιορίζονται στην ανάλυση συναισθήματος. Αντιθέτως, το hugging face διαθέτει μοντέλα για σχεδόν κάθε πτυχή της σύγχρονης τεχνητής νοημοσύνης (επεξεργασία φυσικής γλώσσας, υπολογιστική όραση, ενισχυτική μάθηση). Αντιστοίχως πλούσιο είναι και το περιεχόμενο σε datasets που διαθέτει.
- Comet ML: Ένα πολύ χρήσιμο εργαλείο για το logging πειραμάτων. Ο χρήστης μπορεί να διαχειρίζεται, να οπτικοποιεί και να βελτιστοποιεί μοντέλα, παρακολουθώντας την εκπαίδευσή τους και οτιδήποτε άλλο επιλέξει. Συνδέεται άμεσα με τον κώδικα μέσω αντίστοιχης βιβλιοθήκης. Το **σχήμα 5.1** αναδεικνύει τη χρησιμότητα της πλατφόρμας.
- PyTorch: είναι ένα **framework** μηχανικής μάθησης ανοιχτού κώδικα που βασίζεται στη γλώσσα προγραμματισμού Python και στη βιβλιοθήκη Torch. Προτιμάται συχνά για την έρευνα πάνω στη βαθιά μάθηση, ενώ οι ανταγωνιστές του είναι το **TensorFlow** και το **Keras**. Έχει σχεδιαστεί με στόχο την επιτάχυνση της διαδικασίας από τη δημιουργία ενός μοντέλου μέχρι την δημοσίευσή του.

Εκτός των παραπάνω λογισμικών/εργαλείων χρησιμοποιήθηκαν για την εκπόνηση της διπλωματικής το Google Drive και τα Google Sheets, τα οποία όμως είναι στοιχειώδη και δε θα αναλυθούν.

5.1. ΛΟΓΙΣΜΙΚΟ



Σχήμα 5.1: Σύγκριση πειραμάτων με το εργαλείο Comet ML

Η ανάλυση συνεχίζει με τις χρησιμοποιούμενες βιβλιοθήκες. Λόγω του όγκου της υλοποίησης χρειάστηκαν αρκετές βιβλιοθήκες, από τις οποίες θα αναλυθούν οι σημαντικότερες, ενώ οι υπόλοιπες θα αναφερθούν επιγραμματικά στον πίνακα 5.1.

1. Pandas: Η βιβλιοθήκη Pandas είναι ένα ισχυρό εργαλείο για την ανάλυση και τον χειρισμό δεδομένων στη γλώσσα προγραμματισμού Python. Αυτή η βιβλιοθήκη παρέχει δομές δεδομένων όπως τα DataFrame και Series, που επιτρέπουν την αποθήκευση και την επεξεργασία δεδομένων σε μορφή πίνακα. Με την Pandas, μπορεί κανείς να διαχειριστεί δεδομένα από διάφορες πηγές, να φιλτράρει τα δεδομένα, και να τα αναλύσει. Χρησιμοποιείται ευρέως στον κλάδο της ανάλυσης δεδομένων. Είναι εύκολη στη χρήση και παρέχει πλούσια λειτουργικότητα για την εκτέλεση πολλών εργασιών στα δεδομένα. Στη διπλωματική βοηθάει στην οπτικοποίηση των κειμένων και στον καθαρισμό τους.
2. NumPy: Η NumPy (Numerical Python) είναι πανίσχυρη στο να διαχειρίζεται και να αναλύει πολυδιάστατους πίνακες, ενώ επίσης επιτρέπει υπολογισμούς γραμμικής άλγεβρας, προσθήκης πινάκων, και πολλά άλλα. Αποθηκεύει και επεξεργάζεται αποδοτικά δεδομένα και βρίσκει ευρεία εφαρμογή στους τομείς της μηχανικής μάθησης και ανάλυσης δεδομένων. Η NumPy αποτελεί το θεμέλιο για πολλές άλλες βιβλιοθήκες που ασχολούνται με αριθμητικούς υπολογισμούς και μπορεί να συνεργαστεί με τις περισσότερες βιβλιοθήκες, χωρίς να απαιτείται κάποια μετατροπή των δεδομένων σε λίστες της βασικής Python. Σε ένα μεγάλο ποσοστό υπολογισμών της διπλωματικής εργαζόμαστε με πίνακες της NumPy.
3. scikit-learn: Εύκολα μπορεί κανείς να εκπαιδεύσει και να αξιολογήσει μοντέλα μηχανικής μάθησης με τη χρήση της, καθώς περιλαμβάνει μια ευρεία γκάμα αλγορίθμων μηχανικής μάθησης (logistic regression, random forest, KNN κ.α.). Εκτός όμως από τους διάφορους classifiers, η βιβλιοθήκη περιέχει μετρικές,

ΚΕΦΆΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

όπως η ακρίβεια (accuracy), το F1 score και ο πίνακας σύγχυσης (confusion matrix) που αξιοποιούνται στη διπλωματική.

4. Pillow: Η Pillow επιτρέπει το άνοιγμα και την επεξεργασία εικόνων με εύκολο για τον χρήστη τρόπο. Βασικές λειτουργίες επεξεργασίας των εικόνων είναι η αλλαγή του μεγέθους, η περικοπή, η περιστροφή. Η διπλωματική εργασία αξιοποιεί τις δυνατότητές της βιβλιοθήκης για την προεπεξεργασία των εικόνων προτού μετατραπούν σε τένσορες, που είναι ο τύπος δεδομένων που απαιτείται για τα μοντέλα της PyTorch.
5. transformers: Αυτή η βιβλιοθήκη βασίζεται στην αρχιτεκτονική προσομοίωσης μετασχηματιστών (transformers) και παρέχει πρόσβαση σε προ-εκπαιδευμένα μοντέλα (παραδείγματα έχουν ήδη δωθεί σε προηγούμενα κεφάλαια). Με την Transformers, ο ερευνητής μπορεί να εκπαιδεύσει τα δικά του μοντέλα επεξεργασίας φυσικής γλώσσας και να τα εκπαιδεύσει πρακτικά σε οποιοδήποτε task επιθυμεί. Την χρησιμοποιούμε για την φόρτωση προεκπαιδευμένων μοντέλων από το Hugging Face.
6. torch: Βοηθάει στην ανάπτυξη βαθιών νευρωνικών δικτύων. Αποτελεί μια ανοιχτού κώδικα βιβλιοθήκη που διευκολύνει τη δημιουργία, την εκπαίδευση και την αξιολόγηση των νευρωνικών δικτύων. Η Torch παρέχει ένα πλούσιο περιβάλλον για τον προγραμματισμό νευρωνικών δικτύων, συμπεριλαμβανομένων των συνηθισμένων αρχιτεκτονικών όπως τα συνελικτικά δίκτυα (CNN)). Επιπλέον, η Torch υποστηρίζει την επιτάχυνση υπολογισμών μέσω GPU, πράγμα που την καθιστά κατάλληλη για εκπαιδευτικούς σκοπούς και ερευνητικές εφαρμογές στον τομέα της μηχανικής μάθησης και της βαθιάς μάθησης. Με την torch υλοποιήθηκαν τα μοντέλα που χρησιμοποιούνται στην εργασία και δίνονται αναλυτικά σε επόμενο κεφάλαιο.

Name	Version	Description
Pandas	1.4.1	Ευκολία στη ανάλυση και διαχείριση δεδομένων μέσω ειδικών δομών
h5py	3.9.0	Αποτελεσματική αποθήκευση μεγάλου όγκου πληροφορίας, εύκολη διασύνδεση με τη βιβλιοθήκη Numpy
torch	2.0.1	Διασύνδεση με το Pytorch Framework
torchvision	0.15.2	Επεξεργασία εικόνων για εφαρμογές υπολογιστικής όρασης
NumPy	1.25.2	Διαχείριση πινάκων, μαθηματικές πράξεις
matplotlib	3.5.1	Δημιουργία γραφημάτων κάθε μορφής
random	3.10.4	Αναπαραγωγή τυχαίων αριθμών
os	3.10.4	Πρόσβαση σε λειτουργίες του λειτουργικού συστήματος
sys	3.10.4	Πρόσβαση σε μεταβλητές και συναρτήσεις του interpreter
huggingface_hub	0.16.4	Διασύνδεση με το Hugging Face Framework
transformers	4.31.0	Φόρτωση και εκπαίδευση προεκπαιδευμένων μοντέλων από το Hugging Face Framework
tqdm	4.65.0	Βελτιωμένη απεικόνιση επαναληπτικών αλγορίθμων στη κονσόλα
scikit-learn	1.0.2	Αλγόριθμοι/συναρτήσεις μηχανικής μάθησης
re	3.10.4	Κανονικές εκφράσεις (Regular Expressions)
contractions	0.1.73	Επέκταση αποστρόφων που συνενώνουν δύο λέξεις, πχ γι' αυτό → για αυτό
Pillow	9.0.1	Επεξεργασία και διαχείριση εικόνων
opencv	4.5.5.64	Επεξεργασία εικόνων και tasks υπολογιστικής όρασης
zipfile	3.10.4	Διαχείριση αρχείων τύπου zip
vaderSentiment	3.3.2	Εργαλείο ανάλυσης συναισθήματος κειμένου Vader
comet_ml	3.33.11	Διασύνδεση με το comet-ML framework για το logging των πειραμάτων
deep_translator	1.11.4	Μετάφραση μεταξύ διαφορετικών γλωσσών

Πίνακας 5.1: Βιβλιοθήκες της Python που χρησιμοποιήθηκαν

5.2 ΕΠΙΛΕΓΜΕΝΟ DATASET: MVSA SINGLE

Ως κατάλληλο dataset επιλέχθηκε το MVSA-Single [62]. Το MVSA-Single αποτελείται από 4869 ζεύγη κειμένου-εικόνας, που λήφθηκαν από το Twitter με τη χρήση του Twitter API. Το συναίσθημα αποδίδεται σε 3 διαφορετικές κατηγορίες (αρνητικό, ουδέτερο και θετικό). Το σετ δεδομένων δημοσιεύθηκε με labels για το κείμενο και για την εικόνα, αλλά όχι και για τον συνδυασμό τους. Η διαδικασία της απόδοσης συναισθήματος έγινε από έναν μόνο annotator.

Στη συνέχεια δημοσιεύθηκε μια βελτιωμένη έκδοση του αρχικού dataset με την ονομασία MVSA-Multiple. Προστέθηκαν περισσότερα δεδομένα, αφού το νέο dataset αποτελείται από 19600 ζεύγη, διατηρήθηκαν οι 3 κατηγορίες συναισθημάτων, αλλά στη διαδικασία του annotation συμμετείχαν 3 διαφορετικοί annotators. Ωστόσο λόγω των δεσμεύσεων στους υπολογιστικούς πόρους στη διπλωματική εργασία εργαστήκαμε με το MVSA-Single, εκτός από την προεπεξεργασία που ακολουθεί στην επόμενη παράγραφο, όπου πραγματοποιήθηκε και στα δύο σύνολα δεδομένων.

MVSA-Single

Πριν την προεπεξεργασία, το dataset περιέχει 4869 ζεύγη κειμένου-εικόνας. Γίνεται έλεγχος αρχικά για το αν υπάρχουν κανονικά όλα τα δεδομένα σε κάθε είσοδο κειμένου και εικόνας. Δεν παρατηρείται κάποια απώλεια δεδομένων. Έπειτα πρέπει να αποδοθεί label συνολικά στην πολυτροπική είσοδο, μέσω των παρακάτω απλών κανόνων, ακολουθώντας την ίδια μεθοδολογία όπως έχει προταθεί σε προηγούμενες δημοσιεύσεις [63]:

1. Αν το label του κειμένου είναι ίδιο με το label της εικόνας, τότε και το συνολικό label θα είναι το ίδιο.
2. Αν ένα από τα δύο label είναι ουδέτερο και το άλλο είναι θετικό ή αρνητικό, τότε το συνολικό label θα είναι ίδιο με αυτό του μη ουδέτερου.
3. Αν το ένα label είναι θετικό και το άλλο αρνητικό, τότε το συνολικό label είναι άγνωστο και το δείγμα αφαιρείται από το dataset.

Φαινόμενο	Πλήθος
Αρχικό dataset	4869
Απώλεια δεδομένων	0
Αδυναμία απόδοσης label	358
Τελικό dataset	4511

Πίνακας 5.2: Προεπεξεργασία MVSA-Single

Το MVSA-Single δεν έχει κάποιο συγκεκριμένο περιεχόμενο, είναι γενικής χρήσης και περιλαμβάνει από θέματα πολιτικής μέχρι καθημερινά memes του διαδικτύου.

MVSA-Multiple

Για τον καθαρισμό του dataset ακολουθείται η ίδια διαδικασία με παραπάνω, με τη μόνη διαφορά πως σε αυτή τη περίπτωση υπάρχουν 3 annotations για κάθε label. Χρειάζεται λοιπόν πρώτα να βρεθεί με τη μέθοδο της πλειοφηφίας το κάθε label για κείμενο και εικόνα ξεχωριστά και στη συνέχεια να αποφασιστεί το πολυτροπικό label. Ο [πίνακας 5.3](#) συνοψίζει τη διαδικασία καθαρισμού του σετ δεδομένων:

Φαινόμενο	Πλήθος
Αρχικό dataset	19600
Απώλεια δεδομένων	3
Αδυναμία απόδοσης label	2576
Τελικό dataset	17024

Πίνακας 5.3: Προεπεξεργασία MVSA-Multiple

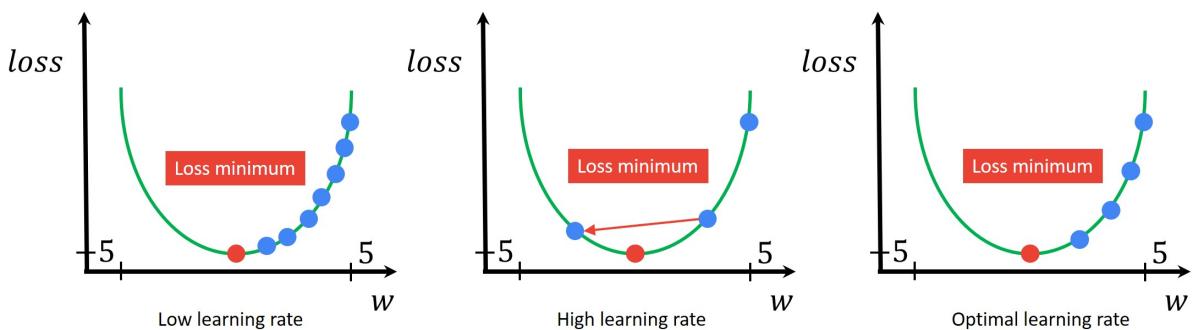
5.3 ΥΠΕΡΠΑΡΑΜΕΤΡΟΙ

Στην αμέσως επόμενη ενότητα θα δοθούν οι καλύτερες υλοποιήσεις για κάθε μοντέλο, όπου σε κάθε υλοποίηση θα προσδιορίζονται οι τιμές των παραμέτρων που εξηγούνται σε αυτό το κεφάλαιο. Οι υπερπαράμετροι στο πλαίσιο της μηχανικής μάθησης και της τεχνητής νοημοσύνης αναφέρονται στις ρυθμίσεις διαμόρφωσης ενός μοντέλου που δεν μαθαίνονται από τα δεδομένα αλλά έχουν οριστεί πριν από τη διαδικασία εκπαίδευσης. Αυτές οι παράμετροι διέπουν τη συμπεριφορά του αλγορίθμου εκμάθησης και μπορούν να επηρεάσουν σημαντικά την απόδοση ενός μοντέλου μηχανικής μάθησης [64].

- Epochs: Στη μηχανική μάθηση, οι εποχές παίζουν καθοριστικό ρόλο στην εκπαίδευση των μοντέλων. Μια εποχή αναφέρεται σε μία επανάληψη μέσω ολόκληρου του συνόλου των δεδομένων εκπαίδευσης κατά την εκπαίδευση του μοντέλου. Κατά τη διάρκεια κάθε εποχής, οι παράμετροι του μοντέλου ενημερώνονται χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης που βασίζεται σε κλίση, όπως η στοχαστική κάθισμας κλίσης (SGD) ή ο Adam, για την ελαχιστοποίηση της συνάρτησης απώλειας και τη βελτίωση της απόδοσης του μοντέλου. Ο αριθμός των εποχών είναι μια υπερπαράμετρος που καθορίζει πόσες φορές ολόκληρο το σύνολο δεδομένων εκπαίδευσης υποβάλλεται σε επεξεργασία από το μοντέλο. Η επιλογή του κατάλληλου αριθμού εποχών είναι απαραίτητη για την αποτελεσματική εκπαίδευση ενός μοντέλου. Πολύ λίγες εποχές μπορεί να οδηγήσουν σε υποπροσαρμογή, όπου το μοντέλο αποτυγχάνει να συλλάβει πολύπλοκα μοτίβα στα δεδομένα, ενώ πάρα πολλές εποχές μπορεί να οδηγήσουν σε υπερπροσαρμογή, όπου το μοντέλο εξειδικεύεται υπερβολικά στα δεδομένα εκπαίδευσης και δε μπορεί να αποδόσει σε άγνωστα δεδομένα.
- Learning rate: Στη μηχανική μάθηση, ο όρος ρυθμός μάθησης (learning rate) αναφέρεται στην τιμή που καθορίζει πόσο μεγάλα βήματα πρέπει να κάνει

ΚΕΦΑΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

ένα μοντέλο κατά την εκπαίδευση, όταν προσαρμόζει τις παραμέτρους του. Αυτή η τιμή επηρεάζει τον ρυθμό σύγκλισης του μοντέλου και μπορεί να έχει σημαντικό αντίκτυπο στην απόδοση του. Ένας υπερβολικά υψηλός ρυθμός μάθησης μπορεί να οδηγήσει σε αστάθεια στην εκπαίδευση, ενώ ένας υπερβολικά χαμηλός μπορεί να καθυστερήσει τη σύγκλιση του μοντέλου. Η επιλογή του κατάλληλου ρυθμού μάθησης απαιτεί πειραματισμό και προσαρμογή στην εκάστοτε εργασία μηχανικής μάθησης, και αποτελεί σημαντική πτυχή της διαδικασίας εκπαίδευσης των μοντέλων. Η πορεία της εκμάθησης ανάλογα με τη τιμή του ρυθμού εκμάθησης φαίνεται στο [σχήμα 5.2](#).

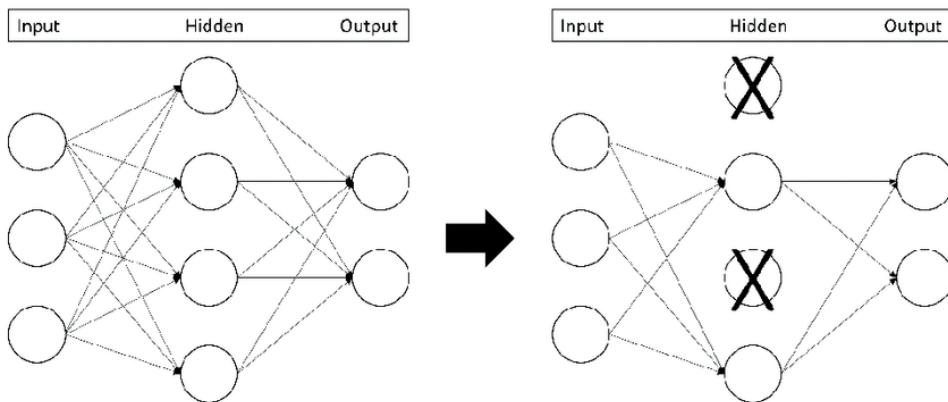


Σχήμα 5.2: Πορεία εκμάθησης για διαφορετικούς ρυθμούς εκμάθησης

- **Loss function:** Η συνάρτηση απώλειας ή αλλιώς συνάρτηση κόστους είναι απαραίτητη για την διαδικασία της εκμάθησης. Η συνάρτηση αυτή ποσοτικοποιεί τα σφάλματα του μοντέλου, μετρώντας την απόσταση μεταξύ της πρόβλεψης και της πραγματικής τιμής. Ο στόχος της συνάρτησης είναι να παρέχει μια μετρική για το πόσο καλά ανταπεξέρχεται το μοντέλο, οδηγώντας το στη βελτίωση των προβλέψεων του. Το μοντέλο προσπαθεί να μειώσει τη τιμή της συνάρτησης κόστους και ιδανικά να τη μηδενίσει. Στην επιβλεπόμενη μάθηση αυτό επιτυγχάνεται με την προσαρμογή των παραμέτρων του μοντέλου σε κάθε επανάληψη της διαδικασίας εκπαίδευσης. Στις υλοποιήσεις της διπλωματικής χρησιμοποιείται η συνάρτηση απωλειών cross-entropy. Για τον υπολογισμό της απόκλισης της πρόβλεψης από τη ζητούμενη τιμή, χρησιμοποιεί τις κατανομές πιθανότητας των προβλέψεων του μοντέλου και τον λογάριθμο.
- **batch size:** Το μέγεθος δέσμης καθορίζει τον αριθμό των δεδομένων εκπαίδευσης που χρησιμοποιούνται σε κάθε επανάληψη της διαδικασίας εκπαίδευσης. Αυτό επηρεάζει τόσο την ταχύτητα εκπαίδευσης όσο και τη χρήση μνήμης. Ένα μεγάλο μέγεθος δέσμης μπορεί να επιταχύνει την εκπαίδευση, αλλά μπορεί να απαιτήσει περισσότερη μνήμη. Αντίστροφα, ένα μικρό μέγεθος δέσμης μπορεί να μειώσει τη μνήμη που απαιτείται, αλλά μπορεί να επιβραδύνει την εκπαίδευση. Έτσι, η βέλτιστη επιλογή του batch size εξαρτάται από τον τύπο των δεδομένων εκπαίδευσης και τους υπολογιστικούς πόρους του συστήματος.
- **optimizer:** Ο βελτιωτής είναι υπεύθυνος για την ρύθμιση των παραμέτρων του μοντέλου με σκοπό τη μείωση της συνάρτησης κόστους και τη βελτίωση της

απόδοσης. Οι επιλογές περιλαμβάνουν αλγόριθμους όπως η στοχαστική καθοδήγηση κλίσης (stochastic gradient descent ή SGD), ο Adam και ο RMSprop, με κάθεναν από αυτούς να έχει τα δικά του πλεονεκτήματα και αδυναμίες. Η υπερπαράμετρος αυτή επηρεάζει την ταχύτητα σύγκλισης του μοντέλου και τον τρόπο με τον οποίο ενημερώνονται οι παράμετροι, επομένως η σωστή επιλογή του μπορεί να έχει σημαντική επίδραση στην απόδοση του μοντέλου. Στη δημοσίευση [65] γίνεται αρχικά μια σύγκριση μεταξύ των Adam και SGD optimizers. Στις υλοποιήσεις της διπλωματικής χρησιμοποιείται κατά κόρον ο AdamW βελτιωτής.

- **scheduler:** Ο προγραμματιστής (scheduler) αναφέρεται σε έναν μηχανισμό που ρυθμίζει τον ρυθμό μάθησης κατά τη διάρκεια της εκπαίδευσης ενός μοντέλου. Οι schedulers επιτρέπουν την αλλαγή του ρυθμού μάθησης με βάση διάφορες στρατηγικές, όπως η μείωση του ρυθμού κατά τη διάρκεια της εκπαίδευσης ή η αυξομείωσή του όταν η απόδοση στο σύνολο επικύρωσης (validation set) σταματάει να βελτιώνεται. Στη διπλωματική εργασία ο κύριος scheduler είναι ο εκθετικός, ο οποίος, προσαρμόζοντας την εκθετική παράμετρο, μπορεί να αυξάνει, να μειώνει ή να κρατάει σταθερό το ρυθμό εκμάθησης [66].
- **dropout:** Στα νευρωνικά δίκτυα χρησιμοποιείται συχνά το dropout ως τεχνική κανονικοποίησης [67]. Το dropout λειτουργεί αντιστρόφως από τον τρόπο που λειτουργούν τα κανονικά επίπεδα του δίκτυου, αφαιρώντας τυχαία νευρώνες από τη διαδικασία εκπαίδευσης. Αυτό το καθορισμένο ποσοστό απόπτωσης (τιμές από 0 για 0% εως 1 για 100%) ελέγχει πόσοι νευρώνες απενεργοποιούνται κατά τη διάρκεια κάθε επανάληψης της εκπαίδευσης, εμποδίζοντας έτσι το δίκτυο από το να εξαρτάται υπερβολικά από συγκεκριμένους νευρώνες και αποτρέποντας το overfitting στα δεδομένα εκπαίδευσης. Χρησιμοποιείται σχεδόν σε κάθε πιθανό μοντέλο της παρούσας διπλωματικής. Στην εικόνα που ακολουθεί φαίνεται η εφαρμογή του dropout στο κρυφό στρώμα ενός classifier, με πιθανότητα 50%. (σχήμα 5.3)



Σχήμα 5.3: Εφαρμογή dropout 0.5 στο κρυφό στρώμα

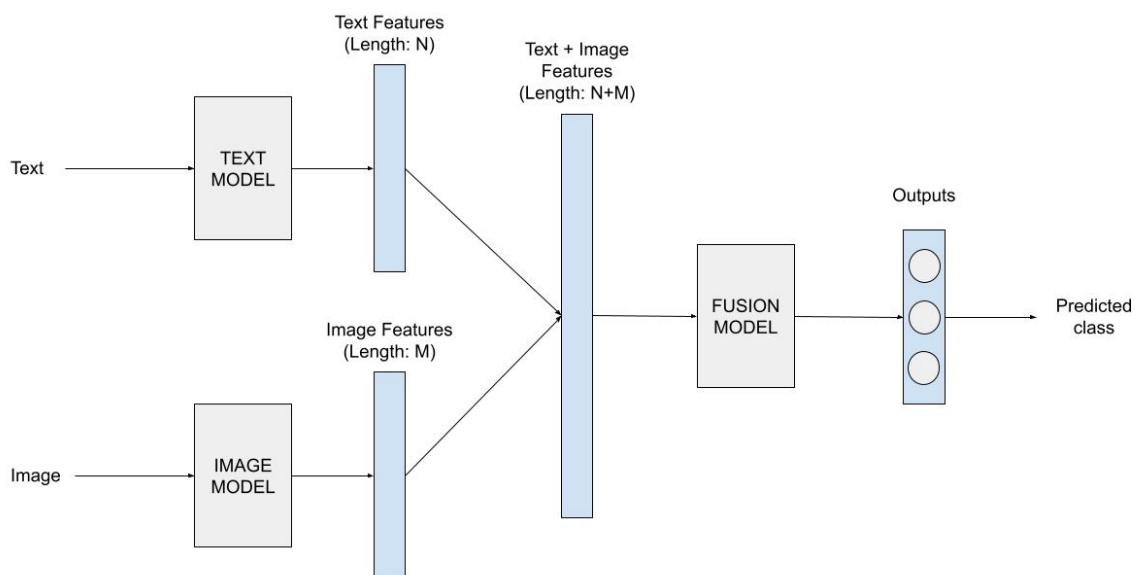
Στο επόμενο κεφάλαιο των πειραμάτων, αναφέρεται ο όρος αναζήτηση πλέγματος για την επιλογή των κατάλληλων υπερπαραμέτρων. Η **Αναζήτηση Πλέγματος**

ΚΕΦΑΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

(Grid Search) χρησιμοποιείται για τη συστηματική διερεύνηση διαφορετικών συνδυασμών υπερπαραμέτρων για την εύρεση της καλύτερης διαμόρφωσης για ένα μοντέλο. Η αναζήτηση πλέγματος λειτουργεί ορίζοντας ένα πλέγμα τιμών υπερπαραμέτρων και στη συνέχεια εκπαιδεύοντας και αξιολογώντας το μοντέλο για κάθε συνδυασμό τιμών. Η αναζήτηση πλέγματος καταφέρνει να εντοπίσει τον ιδανικότερο συνδυασμό παραμέτρων για τη συγκεκριμένη υλοποίηση, όμως ταυτόχρονα έχει το μειονέκτημα του υψηλού υπολογιστικού κόστους.

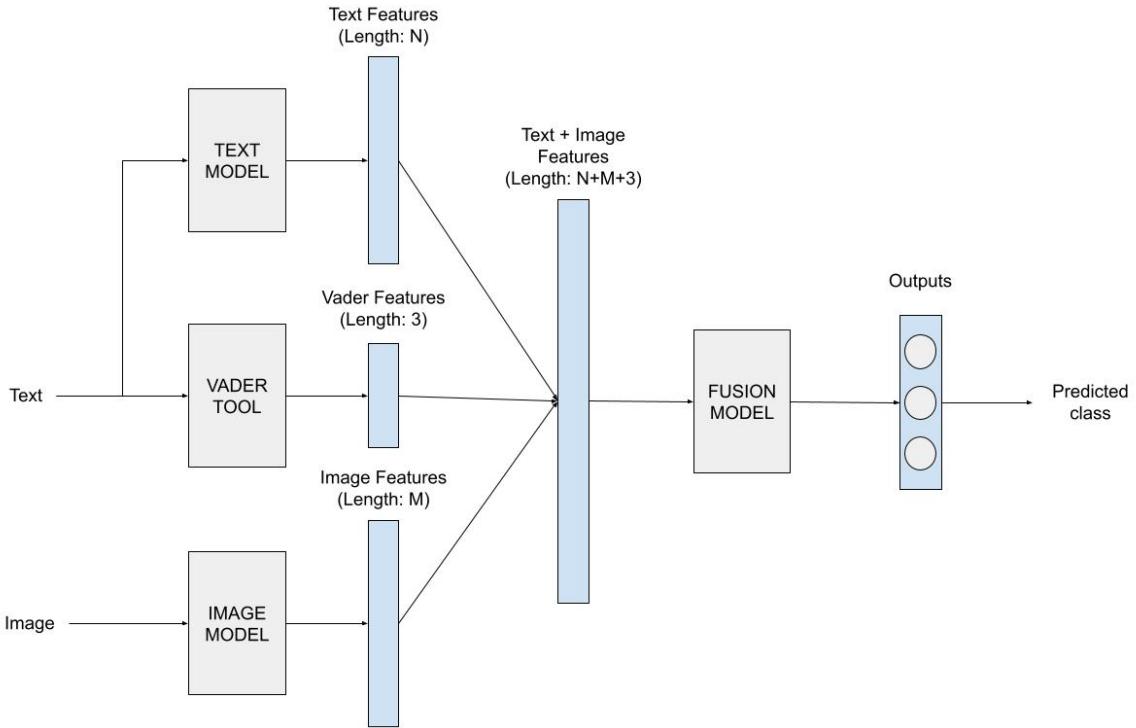
5.4 ΠΟΛΥΤΡΟΠΙΚΟ ΣΥΣΤΗΜΑ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΕΙΜΕΝΟΥ-ΕΙΚΟΝΑΣ

Η ολοκληρωμένη υλοποίηση του προτεινόμενου πολυτροπικού συστήματος ανάλυσης συναισθήματος φαίνεται στο [σχήμα 5.4](#). Κατά την εκπαίδευση, εκπαιδεύεται αρχικά το επιλεγμένο μοντέλο κειμένου. Έπειτα υπολογίζονται για κάθε είσοδο τα χαρακτηριστικά κειμένου. Η ίδια διαδικασία συμβαίνει και με το μοντέλο της εικόνας. Στο επόμενο βήμα, συνενώνονται τα χαρακτηριστικά κειμένου-εικόνας για το κάθε ζεύγος δεδομένων. Επιλέχθηκε η μέθοδος της πρώιμης συνένωσης (early fusion) ή συνένωση στο επίπεδο των χαρακτηριστικών ως η καταλληλότερη. Το μοντέλο συνένωσης εκπαιδεύεται για κάποιες εποχές και αφού τελειώσει η διαδικασία της εκπαίδευσής του, εξάγει τιμές για τις 3 πιθανές κλάσεις εξόδου (αρνητικό, ουδέτερο, θετικό) και από τις τιμές αυτές προκύπτει η τελική πρόβλεψη συναισθήματος. Δοκιμάστηκε επίσης να γίνει υβριδική προσέγγιση στην επεξεργασία του κειμένου, χρησιμοποιώντας και το εργαλείο Vader που είχε εξηγηθεί στο [υποκεφάλαιο 2.5](#). Το νεό σύστημα σε αυτή τη περίπτωση δίνεται στο [σχήμα 5.5](#).



Σχήμα 5.4: Πολυτροπικό σύστημα ανάλυσης συναισθήματος κειμένου και εικόνας

5.4. ΠΟΛΥΤΡΟΠΙΚΟ ΣΥΣΤΗΜΑ ΑΝΆΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΕΙΜΕΝΟΥ-ΕΙΚΟΝΑΣ



Σχήμα 5.5: Πολυτροπικό σύστημα ανάλυσης συναισθήματος κειμένου και εικόνας με τη προσθήκη του εργαλείου Vader

Για την εκπαίδευση του πολυτροπικού συστήματος ώστε να είναι έτοιμο να δεχτεί είσοδο κειμένου και εικόνας και να εξάγει το συναίσθημα στην έξοδο, η διαδικασία συνοψίζεται στα ακόλουθα βήματα:

1. Χωρισμός των δεδομένων σε train/validation/test σετ. Δίνεται προσοχή στο να διατηρηθούν τα σύνολα αυτούσια σε όλες τις εκπαιδεύσεις.
2. Προεπεξεργασία κειμένων.
3. Εκπαίδευση μοντέλου κειμένου.
4. Αποσύνδεση του ταξινομητή από το μοντέλο κειμένου και υπολογισμός των χαρακτηριστικών για ΟΛΑ τα κείμενα, ανεξαρτήτως σετ.
5. (Προαιρετικά) Εξαγωγή προβλέψεων από το εργαλείο Vader για όλα τα κείμενα.
6. Προεπεξεργασία εικόνων.
7. Εκπαίδευση μοντέλου εικόνων.
8. Αποσύνδεση του ταξινομητή από το μοντέλο εικόνων και υπολογισμός των χαρακτηριστικών για ΟΛΕΣ τις εικόνες.
9. Συνένωση των χαρακτηριστικών.

ΚΕΦΆΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

10. Εκπαίδευση μοντέλου συνένωσης.
11. Εξαγωγή προβλέψεων από το μοντέλο συνένωσης για το test set.

Προεπεξεργασία κειμένου

Τα βήματα της προεπεξεργασίας αποφασίστηκαν λαμβάνοντας υπόψη τα χαρακτηριστικά του dataset, το task της ανάλυσης συναισθήματος και προτάσεις από προηγούμενες δημοσιεύσεις πάνω στο ίδιο αντικείμενο. Προτού λοιπόν δοθούν τα δεδομένα κειμένου στο μοντέλο, περνούν από τα επόμενα στάδια επεξεργασίας:

1. Αφαίρεση των tags και γενικά των λέξεων που ξεκινούν με το σύμβολο '@'
2. Αφαίρεση των συμβόλων '#' διατηρώντας το περιεχόμενο
3. Αποκοπή επιπλέον κενών χαρακτήρων
4. Επέκταση των αποστρόφων

Επειδή το dataset σχηματίζεται από δεδομένα προερχόμενα από το Twitter, γίνεται ειδική μέριμνα για να αφαιρεθούν όλα τα tags (λέξεις που ξεκινούν με το σύμβολο '@') καθώς δε προσδίδουν νόημα στο κείμενο και δεν εξετάζεται πιθανή συσχέτιση μεταξύ διαφορετικών κειμένων του dataset. Αντίστοιχα το σύμβολο '#' αφαιρείται ώστε να μη μπερδεύεται το γλωσσικό μοντέλο με άγνωστες λέξεις που δεν έχει συναντήσει κατά την προεκπαίδευσή του.

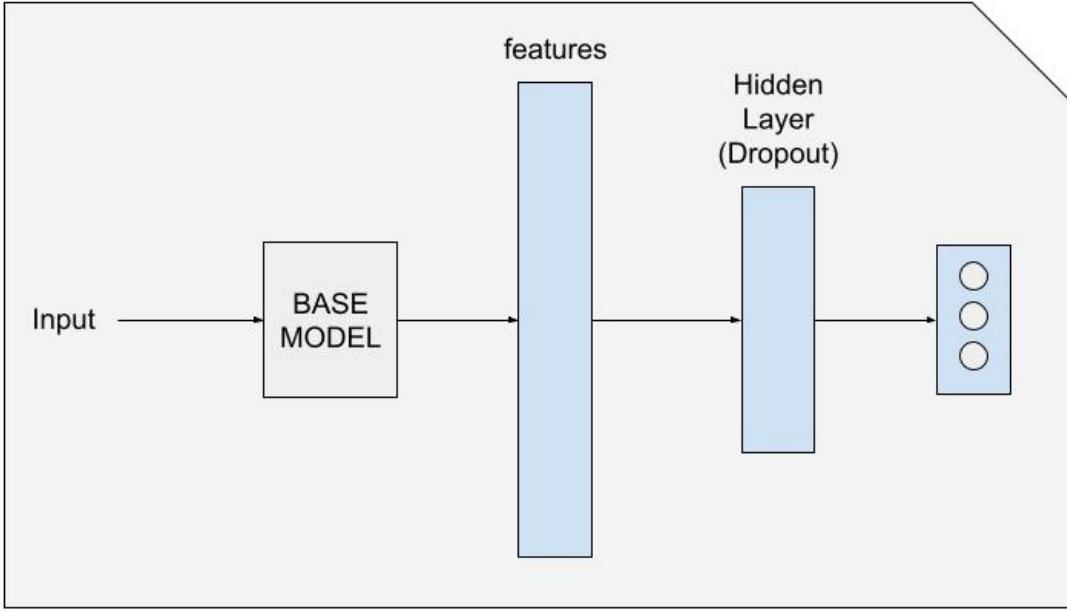
Εκτός των παραπάνω μεθόδων που σχετίζονται με το κείμενο εισόδου, τα κείμενα προτού δοθούν στο επιλεγμένο μοντέλο χρειάζεται να περάσουν από τον κατάλληλο tokenizer, αφού όλα τα υποψήφια μοντέλα βασίζονται στην αρχιτεκτονική του transformer.

Μοντέλο κειμένου

Τα μοντέλα που δοκιμάστηκαν και βελτιστοποιήθηκαν δοκιμάζοντας διαφορετικούς συνδυασμούς υπερπαραμέτρων είναι:

1. Bert base uncased
2. Bert base cased
3. roberta base
4. roberta large
5. albert base v2
6. microsoft deberta base

5.4. ΠΟΛΥΤΡΟΠΙΚΟ ΣΥΣΤΗΜΑ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΕΙΜΕΝΟΥ-ΕΙΚΟΝΑΣ



Σχήμα 5.6: Βασικό μοντέλο κειμένου/εικόνας για την διαδικασία της εκπαίδευσης

Στο [σχήμα 5.6](#) φαίνεται η δομή του μοντέλου για την εκπαίδευση. Το μοντέλο που χρησιμοποιείται σαν βάση είναι ένα προεκπαιδευμένο μοντέλο, που φορτώνεται από το Hugging Face. Για τον έλεγχο της πορείας της εκπαίδευσης δημιουργείται ο classifier που φαίνεται στο σχήμα, στον οποίο πειραματίζομαστε με το μέγεθος του κρυφού στρώματος και το ποσοστό του dropout (εφαρμόζεται στο κρυφό στρώμα). Στη περίπτωση όπου το μέγεθος του κρυφού στρώματος είναι ίσο με μηδέν, εννοείται πως πραγματοποιείται το πείραμα χωρίς αυτό το επιπλέον στρώμα.

Η διαδικασία επιλογής μοντέλου ήταν ιδιαίτερα απαιτητική, κυρίως λόγω των πολλών παραμέτρων και λόγω των απαιτήσεων σε υπολογιστικούς πόρους και χρόνο εκτέλεσης. Οι παράμετροι που δοκιμάστηκαν για κάθε μοντέλο συνοφίζονται στον ακόλουθο πίνακα ([πίνακα 5.4](#)), ενώ τα καλύτερα μοντέλα που προέκυψαν θα παρουσιαστούν στο επόμενο κεφάλαιο στα αποτελέσματα των πειραμάτων.

Παράμετρος	Εύρος τιμών
Batch size	[4, 32]
Learning Rate	$[10^{-6}, 10^{-3}]$
Epochs	[3, 10]
Dropout	[0, 0.8]
Hidden Layer Size	[0, 200]
Scheduler	[Descending, Ascending, Stable]

Πίνακας 5.4: Ύπερπαράμετροι μοντέλου κειμένου

Προεπεξεργασία εικόνας

Όπως και τα κείμενα, έτσι και οι εικόνες, χρειάζεται να περάσουν από ορισμένα βήματα επεξεργασίας πριν να τροφοδοτήσουν το μοντέλο για την εκπαίδευσή του. Η σημαντικότερη σχεδιαστική παράμετρος εδώ είναι η επιλογή των διάφορων augmentations που μπορούν να χρησιμοποιηθούν στην εικόνα. Η συνηθισμένη περίπτωση είναι το μοντέλο να εκπαιδεύεται και να αντιμετωπίζει καλύτερα άγνωστα δεδομένα, όταν έχουν εφαρμοσθεί τεχνικές διαφοροποίησης των εικόνων εισόδου. Παρόλα αυτά, στη περίπτωσή μας, φάνηκε πως το μοντέλο αναγνώριζε καλύτερα τα άγνωστα δεδομένα, όταν δε χρησιμοποιούνταν κανένα augmentation στις εικόνες εκπαίδευσης. Οι τεχνικές αλλαγής των εικόνων που δοκιμάστηκαν ήταν η αλλαγή μεγέθους, η τυχαία περιστροφή, η τυχαία αντικατάσταση της εικόνας με τη συμμετρική ως προς τον οριζόντιο άξονα (horizontal flip), η αλλαγή στη φωτεινότητα και την αντίθεση και η κανονικοποίηση. Τα πειράματα που αποδεικνύουν τον παραπάνω ισχυρισμό θα παρουσιαστούν στο επόμενο κεφάλαιο. Από όλες τις παραπάνω μετατροπές αναγκαστικά χρησιμοποιήθηκε η αλλαγή μεγέθους στην εικόνα, καθώς κατα κύριο λόγο εκπαιδεύονται μοντέλα που βασίζονται στην αρχιτεκτονική του transformer για εικόνες και απαιτούν συγκεκριμένες διαστάσεις εισόδου. Η κανονικοποίηση σε συγκεκριμένη μέση τιμή και διακύμανση καθώς και η αλλαγή μεγέθους καθορίζονται από το επιλεγμένο μοντέλο που φορτώνεται ως βάση. Επίσης οι εικόνες περνούν από τον image processor, ο οποίος είναι ειδικός επεξεργαστής των εικόνων για κάθε image transformer μοντέλο.

Μοντέλο εικόνας

Ο τρόπος εκπαίδευσης του μοντέλου της εικόνας είναι παρόμοιος με το μοντέλο του κειμένου, διαλέγοντας φυσικά ένα μοντέλο κατάλληλο για επεξεργασία δεδομένων εικόνας. Ωστόσο, υπάρχει σαφώς η διαφορά πως δοκιμάστηκαν και αρκετά μοντέλα τύπου συνελικτικού δικτύου, στα οποία δε χρησιμοποιήθηκε κάποιος ξεχωριστός classifier στην έξοδο τους, δηλαδή η εικόνα δινόταν στο μοντέλο και το μοντέλο αυτόματα προέβλεπε το αποτέλεσμα. Επειδή όμως, όπως θα φανεί και στο επόμενο κεφάλαιο, τα αποτελέσματα από τα συνελικτικά δίκτυα ήταν κατώτερα απ' ότι στα τύπου transformers δίκτυα, η ανάλυσή τους θα περιοριστεί στην ταξινόμηση εικόνων και δε θα αναφερθούν στο fusion. Τα μοντέλα που χρησιμοποιήθηκαν για τα πειράματα είναι:

- Μοντέλα CNN αρχιτεκτονικής:
 1. RESNET-34
 2. RESNET-50
 3. RESNET-101
 4. RESNEXT-50
 5. DENSENET-121
 6. DENSENET-161
 7. DENSENET-201

5.4. ΠΟΛΥΤΡΟΠΙΚΟ ΣΥΣΤΗΜΑ ΑΝΆΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΕΙΜΕΝΟΥ-ΕΙΚΟΝΑΣ

8. EFFICIENTNET-B0

9. EFFICIENTNET-B7

- Μοντέλα transformers αρχιτεκτονικής:

1. ViT
2. DINO-ViTb8
3. DINO-ViTb16
4. BEiT

Όπως και προηγουμένως, έτσι και εδώ υπάρχουν αρκετές παράμετροι που μελετήθηκαν για να βρεθούν τα μοντέλα με τις υψηλότερες επιδόσεις. Συνοπτικά δίνονται στον [πίνακα 5.5](#):

Παράμετρος	Εύρος τιμών
Batch size	[8, 64]
Learning Rate	[10^{-6} , 10^{-3}]
Epochs	[3, 20]
Dropout	[0, 0.8]
Hidden Layer Size	[0, 200]
Scheduler	[Descending, Ascending, Stable]

Πίνακας 5.5: Υπερπαράμετροι μοντέλου εικόνας

Fusion

Η πρώτη σκέψη και προσπάθεια για τη συνένωση ήταν να χρησιμοποιηθεί η τεχνική της συνένωσης στο επίπεδο των προβλέψεων (late fusion). Μετά τα πρώτα πειράματα όμως φάνηκε πως το μοντέλο δε μπορούσε να εξάγει σωστά συμπεράσματα συνδυάζοντας μόνο τις προβλέψεις από τα μοντέλα κειμένου και εικόνας. Έτσι, ως επόμενη λύση επιλέχθηκε να χρησιμοποιηθούν τα χαρακτηριστικά των δύο προηγούμενων μοντέλων. Τροφοδοτείται λοιπόν το μοντέλο συνένωσης από τους πίνακες χαρακτηριστικών και παράγει στην έξοδο τις τελικές του προβλέψεις, αποτέλεσμα των πληροφοριών από τα 2 modalities.

Για το μοντέλο της συνένωσης προτάθηκαν 4 αρχιτεκτονικές. Η πρώτη είναι ένα πλήρως συνδεδεμένο (fully connected) νευρωνικό δίκτυο με 4 κρυφά στρώματα. Δέχεται ως είσοδο τον πίνακα χαρακτηριστικών που έχει δημιουργηθεί για κάθε ζεύγος εικόνας-κειμένου. Το κάθε στρώμα αποτελείται από hidden size, hidden size, last hidden size και last hidden size αριθμό κόμβων. Οι τιμές hidden size και last hidden size είναι σχεδιαστικοί παράμετροι της αρχιτεκτονικής. Η δεύτερη αρχιτεκτονική είναι παρόμοια, αφαιρώντας ένα κρυφό στρώμα. Παραμένουν δηλαδή 3 κρυφά στρώματα, με μεγέθη hidden size, hidden size και last hidden size το καθένα. Ομοίως, οδηγούμαστε στη τρίτη αρχιτεκτονική αφαιρώντας ένα ακόμη κρυφό στρώμα, δηλαδή παραμένουν 2 κρυφά στρώματα μεγέθους hidden size. Τέλος, η

ΚΕΦΆΛΑΙΟ 5. ΥΛΟΠΟΙΗΣΕΙΣ

τέταρτη αρχιτεκτονική εκμεταλλεύεται τον μηχανισμό της προσοχής. Τα χαρακτηριστικά της εισόδου υπόκεινται σε attention με τον εαυτό τους (self attention) ταυτόχρονα σε πολλές κεφαλές (multi-head attention). Η έξοδος του attention συνεχίζει σε fully-connected νευρωνικό δίκτυο με 1 κρυφό στρώμα μεγέθους hidden size. Καθορίζεται ο αριθμός των κεφαλών με την παράμετρο number of heads. Συνολικά, οι παράμετροι ελέγχου για το fusion των modalities δίνονται στον [πίνακα 5.6](#):

Παράμετρος	Εύρος τιμών
Batch size	[4, 64]
Learning Rate	$[10^{-6}, 10^{-4}]$
Epochs	[5, 25]
Hidden Size	[200, 800]
Last Hidden Size	[50, 400]
Number of heads	[4, 8]

Πίνακας 5.6: Υπερπαράμετροι μοντέλου συνένωσης

6

Αποτελέσματα Πειραμάτων

Ήδη από το προηγούμενο κεφάλαιο έγιναν αρκετές αναφορές για τα αποτελέσματα των πειραμάτων. Σε αυτό λοιπόν το κεφάλαιο συλλέγουμε τα κυριότερα πειράματα και παρουσιάζουμε τα αποτελέσματά τους, εξάγοντας συμπεράσματα και παρατηρήσεις.

Το κεφάλαιο θα ξεκινήσει δείχνοντας τα πειράματα που σχετίζονται με το μοντέλο του κειμένου και θα αναδειχθεί το επιλεγμένο μοντέλο. Η ίδια διαδικασία θα ακολουθηθεί και για το μοντέλο της εικόνας και μετέπειτα για το μοντέλο της συνένωσης. Θα δομηθεί δηλαδή ξεκάθαρα το πολυτροπικό σύστημα που εισήχθηκε στο προηγούμενο κεφάλαιο.

Αξίζει να συμπεριληφθούν και πειράματα που πραγματοποιήθηκαν ώστε να διερευνηθεί η συμπεριφορά του συστήματος σε ιδέες που δεν είχαν δοκιμαστεί νωρίτερα, όπως είναι για παράδειγμα η συμπεριφορά του συστήματος στη σταδιακή προσθήκη/αφαίρεση τροποποιήσεων στην εικόνα και η ικανότητά του να εντοπίζει το συναίσθημα σε κείμενα της ελληνικής γλώσσας.

Τα αποτελέσματα προκύπτουν από 5-fold Cross Validation με χωρισμό 80/10/10 σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου αντίστοιχα. Συνολικά έγιναν fine-tuned 14 μοντέλα κειμένου και 30 μοντέλα εικόνας, εκ των οποίων 19 είναι συνελικτικά και 11 ανήκουν στην οικογένεια των transformers για εικόνες.

Το πολυτροπικό σύστημα επιτυγχάνει επίδοση που πλησιάζει τις πιο πρόσφατες δημοσιεύσεις που ασχολούνται με το ίδιο σύνολο δεδομένων, MVSA-Single, ενώ είναι το μόνο συγκριτικά με τα υπόλοιπα που χρησιμοποιεί image transformers για την εξαγωγή των χαρακτηριστικών των εικόνων. Οι δημοσιεύσεις που συγκρίνονται με τη δική μας υλοποίηση χρησιμοποιούν τα ίδια ποσοστά διαχωρισμού των δεδομένων σε υποσύνολα και τα ίδια βήματα για τον καθαρισμό του συνόλου δεδομένων, καταλήγοντας επίσης σε 4511 ζεύγη κειμένου-εικόνας με τις ίδιες ετικέτες. Στον [πίνακα 6.1](#) συνοψίζονται μερικές από τις πιο πρόσφατες δομές που έχουν δημοσιευθεί για το συγκεκριμένο dataset.

Σύστημα	Accuracy	F1-Score
MVAN [68]	72.98	72.98
Se-MLNN [69]	75.33	73.76
CLMLF [70]	75.33	73.46
VSA-PF [71]	76.11	73.69
ICCI [72]	79.33	77.51
Proposed	77.86	74.79

Πίνακας 6.1: Σύγκριση του προτεινόμενου συστήματος με προηγούμενες δημοσιεύσεις πάνω στο ίδιο σύνολο δεδομένων

6.1 ΠΙΝΑΚΕΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΕΙΡΑΜΑΤΩΝ

Η έλλειψη υπολογιστικών πόρων και ο μεγάλος χρόνος εκμάθησης των μοντέλων δεν επέτρεπε την αναζήτηση σε πλέγμα παραμέτρων (grid search), είναι δηλαδή πιθανό οι μετρικές που δίνονται στους πίνακες αυτής και των επόμενων ενοτήτων να μην είναι οι υψηλότερες εφικτές για το κάθε μοντέλο. Παρόλα αυτά όμως, είναι σίγουρα κοντά στις ιδανικές και ενδεικτικές για το μοντέλο και την ισχύ του συγκριτικά με το task και το dataset. Ακολουθούν λοιπόν πίνακες για το βέλτιστο αποτέλεσμα από κάθε μοντέλο και τις υπερπαραμέτρους που χρησιμοποιήθηκαν για να επιτευχθεί. Αρχικά βλέπουμε στον πίνακα 6.2 και στον πίνακα 6.3 τα μοντέλα BERT-base-cased, BERT-base-uncased, ALBERT-base, RoBERTa-base και DeBERTa-base που μπορούν να χρησιμοποιηθούν ως βάση για το κείμενο.

Αναφορικά με τις βέλτιστες παραμέτρους που φαίνονται στους πίνακες, η πρώτη παρατήρηση που μπορεί να γίνει είναι πως τα μοντέλα εκπαιδεύονται σε λίγες εποχές (μέχρι 6). Αυτό είναι λογικό γιατί τα μοντέλα είναι ήδη προεκπαίδευμένα και χρειάζονται απλά να προσαρμοστούν στο νέο dataset. Επιπλέον, καλύτερα λειτουργούν οι χαμηλοί ρυθμοί εκμάθησης, της τάξεως του 10^{-5} και 10^{-6} .

Για να αποφασιστεί το καλύτερο μοντέλο εστιάζουμε στις μετρικές της ακρίβειας και της τιμής του F1, που δίνονται στο δεύτερο μέρος των πινάκων. Επίσης για τη λήψη της απόφασης ελέγχεται και η πορεία της εκπαίδευσης με τις απώλειες στο σύνολο εκπαίδευσης και επικύρωσης. Η απώλεια στο σετ εκπαίδευσης δείχνει την πορεία της εκπαίδευσης και το κατά πόσο το μοντέλο έχει μάθει το σύνολο στο οποίο εκπαιδεύεται. Μία μεγάλη τιμή μπορεί να υποδείξει πως το μοντέλο δεν έχει εκπαιδευτεί κατάλληλα (underfitting), ενώ μια πολύ μικρή τιμή ενέχει τον κίνδυνο το μοντέλο να έχει προσαρμοστεί υπερβολικά στο σετ εκπαίδευσης (overfitting). Για την εξακρίβωση των παραπάνω φαινομένων χρειάζεται να παρακολουθείται παράλληλα και η απώλεια στο σετ επικύρωσης. Οι παρατηρήσεις είναι και εδώ οι ίδιες, με τη μεγάλη διαφορά πως η απώλεια αυτή αναφέρεται σε δεδομένα που το μοντέλο δεν έχει εκπαιδευτεί, σε άγνωστα δεδομένα. Σίγουρα λοιπόν είναι επιθυμητή η χαμηλότερη δυνατή τιμή απωλειών στο σετ επικύρωσης για την καλύτερη γενίκευσή του.

Για την μέτρηση των επιδόσεων του μοντέλου σε άγνωστα δεδομένα, υπολογίζεται η ακρίβεια και η μετρική του ζυγισμένου F1 σκορ. Είναι προφανές πως το ιδανικό είναι οι μετρικές να είναι όσο το δυνατόν μεγαλύτερες και πιο κοντά στη

6.1. ΠΙΝΑΚΕΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΕΙΡΑΜΑΤΩΝ

μονάδα. Σε αυτό το σημείο ας γίνει μία σύγκριση των μοντέλων και ας δοθούν τα σημαντικότερα πλεονεκτήματα/μειονεκτήματα τους όπως αναδείχθηκαν κατά τα διάφορα πειράματα. Ξεκινώντας από τα δύο BERT μοντέλα που δοκιμάστηκαν, υψηλότερες επιδόσεις επιτυγχάνει το μοντέλο που δέχεται και κεφαλαία γράμματα (BERT-cased). Τα κεφαλαία γράμματα σε κείμενα που προέρχονται από κοινωνικά δίκτυα θα μπορούσαν να εκφράσουν συναίσθημα οργής ή χαράς, επομένως είναι λογική η μικρή διαφορά. Παρόλα αυτά τα BERT μοντέλα είναι περιορισμένων δυνατοτήτων συγκριτικά με τα υπόλοιπα. Το ALBERT έχει το πλεονέκτημα να καταναλώνει λίγους υπολογιστικούς πόρους και να εκπαιδεύεται γρήγορα. Μπορεί επίσης να πιάσει υψηλές τιμές στις μετρικές, συνήθως κοντά με το RoBERTa. Το DeBERTa είναι το απαιτητικότερο μοντέλο σε υπολογιστικούς πόρους, με υψηλούς χρόνους εκμάθησης. Είναι ικανό όμως να φθάσει και στις υψηλότερες επιδόσεις συγκριτικά με τα υπόλοιπα. Το RoBERTa αποτελεί το μέσο μοντέλο συγκριτικά με τα προηγούμενα, όσον αφορά το χρόνο εκμάθησης και τις επιδόσεις του. Επιπρόσθετα δεν είναι τόσο επιφρεπές στις αλλαγές των τιμών των παραμέτρων όσο τα υπόλοιπα μοντέλα κειμένου.

Με βάση τις παραπάνω παρατηρήσεις, σε πρώτη εκτίμηση θα επιλεγεί το κατάλληλο μοντέλο ανάμεσα από τα ALBERT-base, RoBERTa-base και DeBERTa-base. Το **RoBERTa-base** μοντέλο έχει την υψηλότερη επίδοση και τις χαμηλότερες απώλειες στο σύνολο επικύρωσης, οπότε θα επιλεγεί ως το κατάλληλο μοντέλο κειμένου για τη συνέχεια της ανάλυσης.

Μοντέλο	BERT-base-cased	BERT-base-uncased	ALBERT-base
Εποχές	5	5	3
Learning rate	10^{-6}	10^{-6}	5×10^{-6}
Batch size	4	4	8
Dropout	0	0	0.5
Hidden Size	100	100	100
Scheduler	Stable	Stable	Stable
Accuracy score (Test)	71.62	65.71	71.62
F1 score (Test)	71.67	65.37	71.94

Πίνακας 6.2: Υπερπαράμετροι και μετρικές μοντέλων κειμένου (BERT-ALBERT)

Η ανάλυση συνεχίζεται στα υποψήφια μοντέλα εικόνας. Στους πρώτους πίνακες δίνονται τα μοντέλα που βασίζονται στην αρχιτεκτονική των CNN. Ξεκινώντας από τον [πίνακα 6.4](#), δοκιμάστηκαν resnet μοντέλα διαφορετικών μεγεθών. Τα μοντέλα παρουσιάζουν υψηλές τιμές απωλειών, κάνοντας εμφανή τη δυσκολία που έχουν να ταξινομήσουν σωστά τις εικόνες. Από τα 3 resnet μοντέλα καλύτερο αποδεικνύεται το resnet-50, ενώ τις χειρότερες επιδόσεις πετύχαινε το resnet-101, το οποίο σε πρώτη σκέψη δεν είναι αναμενόμενο καθώς είναι θεωρητικά το ισχυρότερο μοντέλο από τα τρία. Δοκιμάστηκε επίσης και το resnext-50 που αποτελεί βελτίωση του αρχικού, χωρίς όμως να καταφέρει να το ξεπεράσει.

Προχωρώντας στα μοντέλα της οικογένειας densenet ([πίνακα 6.5](#)), παρατηρείται και πάλι η δυσκολία γενίκευσης σε άγνωστα δεδομένα με υψηλά validation losses. Για μια ακόμη φορά το μοντέλο μέσου μεγέθους ξεπερνάει σε επιδόσεις

Μοντέλο	RoBERTa-base	DeBERTa-base
Εποχές	2	4
Learning rate	2×10^{-5}	2×10^{-5}
Batch size	16	8
Dropout	0.5	0.5
Hidden Size	100	100
Scheduler	Stable	Stable
Accuracy score (Test)	74.28	72.28
F1 score (Test)	74.35	72.34

Πίνακας 6.3: Υπερπαράμετροι και μετρικές μοντέλων κειμένου (ROBERTA-DEBERTA)

τα υπόλοιπα (densenet-161), μεγιστοποιώντας τις μετρικές και ελαχιστοποιώντας τις απώλειες. Επιπρόσθετα, προτιμώνται μεγαλύτερα batch sizes και υπερισχύει ο scheduler που μειώνει τον ρυθμό εκμάθησης.

Στον [πίνακα 6.6](#) ολοκληρώνονται οι CNN αρχιτεκτονικές με τα efficientnet μοντέλα. Αποτελούν την οικογένεια μοντέλων που πέτυχαν τις χαμηλότερες επιδόσεις συγκριτικά με τις προηγούμενες. Στα θετικά τους είναι πως κατάφεραν να μειώσουν τις απώλειες, χωρίς ωστόσο να διαπιστώνεται μείωση στις μετρικές. Εξακολουθεί να ισχύει και εδώ η παρατήρηση πως τα μεγαλύτερα μοντέλα δυσκολεύονται περισσότερο στο συγκεκριμένο σετ εικόνων, αφού είναι σημαντική η μείωση στην ακρίβεια και στο F1 στο efficientnet-B7 μοντέλο.

Η παρουσίαση των πιθανών μοντέλων εικόνας τελειώνει με τον [πίνακα 6.7](#), όπου δίνονται τα μοντέλα της οικογένειας των transformers. Εισέρχεται στις αρχιτεκτονικές και το κρυφό δίκτυο μεγέθους 100 κόμβων που χρησιμοποιείται ως classifier, όπου συνοδεύεται από 20% dropout σε κάθε μοντέλο. Τα ViT και BEiT απαιτούν παρόμοιο χρόνο εκπαίδευσης, ο οποίος είναι ελάχιστος συγκριτικά με τις απαιτήσεις του DINO. Ακόμη και το μέγεθος του DINO είναι σημαντικά μεγαλύτερο, δυσκολεύοντας τον ερευνητή να το χρησιμοποιήσει αν δε διαθέτει σοβαρούς υπολογιστικούς πόρους. Το ViT καταφέρνει να μειώσει σημαντικά τις απώλειες και στο σετ εκπαίδευσης και στο σετ επικύρωσης. Εκτός αυτού, ξεπερνάει στις μετρικές όλα τα προηγούμενα μοντέλα και λαμβάνοντας υπόψη τα παραπάνω, είναι το μοντέλο εικόνας που χρησιμοποιούμε από εδώ και στο εξής.

Στο [σχήμα 6.1](#) φαίνεται η πορεία της εκπαίδευσης των μοντέλων κειμένου. Στον οριζόντιο άξονα κάθε γραφήματος είναι οι εποχές (συνολικά 6). Το διάγραμμα δείχνει τις τιμές του F1 για το κάθε μοντέλο στο σύνολο επικύρωσης. Η εποχή στην οποία βρίσκεται η υψηλότερη τιμή του F1 στο validation set επιλέγεται ως η καταλληλότερη για το συγκεκριμένο μοντέλο. Αντιστοίχως, στο [σχήμα 6.2](#) βλέπουμε την πορεία εκπαίδευσης των μοντέλων εικόνας.

Με βάση τα παραπάνω αποτελέσματα επιλέχθηκαν λοιπόν τα μοντέλα βάσης για το κάθε modality. Στο κείμενο έχει επιλεγεί το RoBERTa του [πίνακα 6.3](#) και στην εικόνα το ViT του [πίνακα 6.7](#). Όπως αναφέρθηκε στο [υποκεφάλαιο 5.4](#), έχουν επιλεγεί 4 αρχιτεκτονικές για το μοντέλο που μπορεί να χρησιμοποιηθεί στη συνένωση. Τα καλύτερα αποτελέσματα πειραμάτων δίνονται στον [πίνακα 6.8](#). Στο

6.1. ΠΙΝΑΚΕΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΠΕΙΡΑΜΑΤΩΝ

Μοντέλο	RESNET-50	RESNEXT-50	RESNET-34	RESNET-101
Εποχές	3	5	3	3
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	64	16	32	32
Dropout	0	0	0	0
Hidden Size	-	-	-	-
Scheduler	Stable	Descending	Stable	Stable
Accuracy score (Test)	59.42	60.31	59.03	56.49
F1 score (Test)	57.5	59.4	57.17	56.21

Πίνακας 6.4: Υπερπαράμετροι και μετρικές των RESNET μοντέλων εικόνας

Μοντέλο	DENSENET-121	DENSENET-161	DENSENET-201
Εποχές	5	4	2
Learning rate	5×10^{-5}	7×10^{-5}	5×10^{-5}
Batch size	64	48	64
Dropout	0	0	0
Hidden Size	-	-	-
Scheduler	Descending	Descending	Descending
Accuracy score (Test)	58.66	59.87	57.93
F1 score (Test)	56.53	57.47	57.96

Πίνακας 6.5: Υπερπαράμετροι και μετρικές των DENSENET μοντέλων εικόνας

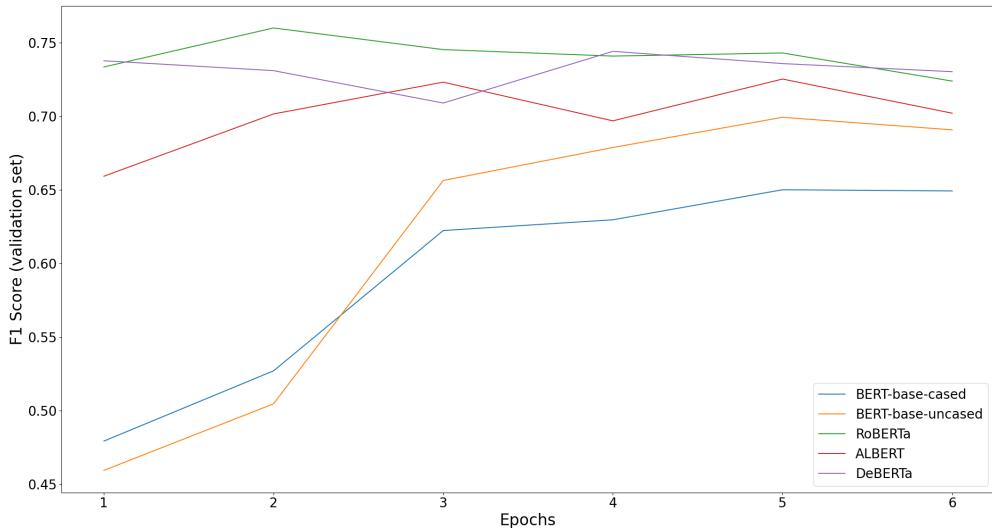
Μοντέλο	EFFICIENTNET-B1	EFFICIENTNET-B7
Εποχές	5	4
Learning rate	5×10^{-5}	5×10^{-5}
Batch size	64	24
Dropout	0	0
Hidden Size	-	-
Scheduler	Descending	Stable
Accuracy score (Test)	58.76	56.34
F1 score (Test)	57.55	55.68

Πίνακας 6.6: Υπερπαράμετροι και μετρικές των EFFICIENTNET μοντέλων εικόνας

σύνολο των πειραμάτων, που πραγματοποιήθηκαν με grid search αφού το επέτρεπαν οι υπολογιστικές απαιτήσεις του μοντέλου συνένωσης, χρησιμοποιήθηκαν 25 εποχές. Είναι φανερό πως οι αρχιτεκτονικές συνένωσης δε παρουσιάζουν έντονες διαφορές μεταξύ τους, με καταλληλότερη να είναι εκείνη με τα δύο κρυφά στρώματα. Σημειώνεται πως το μοντέλο της συνένωσης δέχεται στην είσοδό του στο σύνολο 1536 χαρακτηριστικά (768 χαρακτηριστικά από την έξοδο του μοντέλου κειμένου και 768 χαρακτηριστικά από την έξοδο του μοντέλου εικόνας). Πειρά-

Μοντέλο	ViT	BEiT	DINO-ViTB8
Εποχές	3	3	2
Learning rate	2×10^{-5}	8×10^{-6}	5×10^{-5}
Batch size	16	16	8
Dropout	0.2	0.2	0.2
Hidden Size	100	100	100
Scheduler	Stable	Stable	Stable
Accuracy score (Test)	66.08	58.09	64.3
F1 score (Test)	65.48	58.12	63.04

Πίνακας 6.7: Υπερπαράμετροι και μετρικές των transformer μοντέλων εικόνας

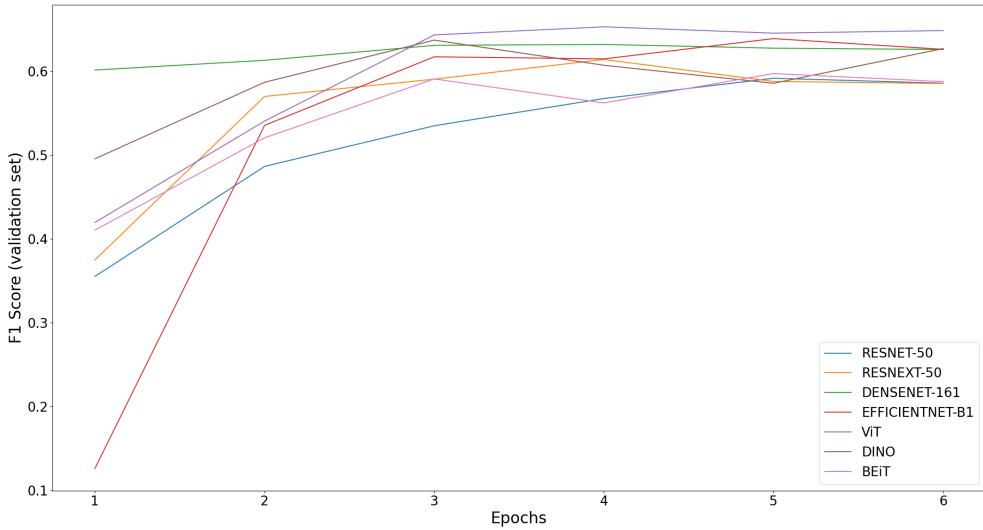


Σχήμα 6.1: Fine-tuning των σημαντικότερων μοντέλων κειμένου για 6 εποχές

ματα με τη χρήση του Vader δε θα παρουσιαστούν, καθώς η επίδραση του στη συνολική ακρίβεια είναι ελάχιστη, αυξάνοντας τις μετρικές κατά περίπου 0.1%.

Ο πίνακας των μοντέλων συνένωσης αναδεικνύει τη χρησιμότητα των πολυτροπικών συστημάτων. Οι μετρικές που προκύπτουν κατά την συνένωση βρίσκονται πάνω από τις μετρικές του κάθε modality ξεχωριστά. Η ισχύς του κάθε μοντέλου όπως αυτή φαίνεται στον πίνακα, ίσχυε και γενικότερα στο μεγαλύτερο εύρος των πειραμάτων, χωρίς ιδιαίτερες αποκλίσεις με την αλλαγή των υπερπαραμέτρων. Ο μηχανισμός της προσοχής φαίνεται να λειτουργεί αποτελεσματικά στη περίπτωσή μας επιτυγχάνοντας λίγο μικρότερες επιδόσεις συγκριτικά με τα κλασσικά πλήρως συνδεδεμένα δίκτυα, παρόλο που τα χαρακτηριστικά προέρχονται από διαφορετικά είδη πληροφορίας. Η πορεία εκπαίδευσης των διαφορετικών αρχιτεκτονικών φαίνεται στο σχήμα 6.3

6.2. ΠΕΙΡΑΜΑΤΙΣΜΟΣ ΜΕ ΤΗΝ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΕΙΚΟΝΩΝ



Σχήμα 6.2: Fine-tuning των σημαντικότερων μοντέλων εικόνας για 6 εποχές

Μοντέλο	4 hidden layers	3 hidden layers	2 hidden layers	Attention
Epochs	5	13	6	12
Learning Rate	3×10^{-5}	10^{-5}	3×10^{-5}	3×10^{-5}
Hidden Size	800	400	800	800
Last Hidden Size	400	300	-	-
Number of heads	-	-	-	8
Scheduler	Descending	Descending	Descending	Descending
Accuracy score (Test)	74.89	74.09	75.01	73.58
F1 score (Test)	74.73	74.16	74.79	73.86

Πίνακας 6.8: Υπερπαράμετροι και μετρικές των μοντέλων συνένωσης

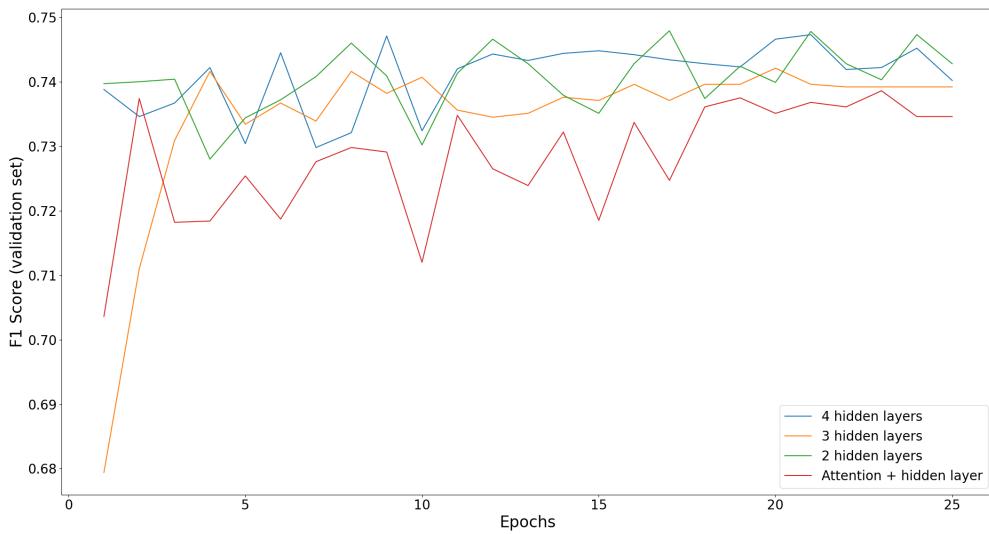
6.2 ΠΕΙΡΑΜΑΤΙΣΜΟΣ ΜΕ ΤΗΝ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΕΙΚΟΝΩΝ

Τα παραπάνω πειράματα πραγματοποιήθηκαν χωρίς κάποια ειδική επεξεργασία των εικόνων προτού δοθούν στο αντίστοιχο μοντέλο. Εν συνεχείᾳ, κρατώντας το ισχυρότερο μοντέλο εικόνας όπως επιλέχθηκε στο προηγούμενο υποκεφάλαιο, ερευνήθηκε η διαφορά του συστήματος στην ικανότητα γενίκευσης σε άγνωστα δεδομένα, επιλέγοντας σε κάθε πείραμα διαφορετικά augmentations των εικόνων. Τα αποτελέσματα των πειραμάτων συνοψίζονται στον πίνακα 6.9.

Πείραμα 0: Καμία προεπεξεργασία

Είναι η πειραματική διάταξη που χρησιμοποιήθηκε και στη προηγούμενη ενότητα. Δε γίνεται κανένα augmentation στις εικόνες.

Πείραμα 1: Κανονικοποίηση των εικόνων



Σχήμα 6.3: Εκπαίδευση των αρχιτεκτονικών συνένωσης για 25 εποχές

Η μόνη προεπεξεργασία που πραγματοποιείται είναι η κανονικοποίηση των εικόνων, χρησιμοποιώντας ως μέση τιμή και τυπική διακύμανση τις τιμές του ViT processor.

Πείραμα 2: Προσθήκη τυχαιότητας

Εκτός από τη κανονικοποίηση, οι εικόνες αλλάζουν μέγεθος, καθρεφτίζονται τυχαία ως προς τον οριζόντιο άξονα και υπόκεινται σε τυχαία περιστροφή.

Πείραμα 3: Πλήρης επεξεργασία των εικόνων

Κρατώνται οι τυχαιότητες που προστέθηκαν στο πείραμα 2 και επιπλέον οι εικόνες αλλάζουν στη φωτεινότητα, στην αντίθεση των χρωμάτων, στον κορεσμό τους και στην μικρή αλλαγή στη τιμή των χρωμάτων. Επιπροσθέτως εφαρμόζεται ένας affine μετασχηματισμός.

Αριθμός πειράματος	0	1	2	3
Image Accuracy score	66.08	62.53	66.3	63.64
Image F1 score	65.48	62.02	65.16	62.45
Multimodal Accuracy score	75.01	71.62	72.28	73.94
Multimodal F1 score	74.79	71.26	71.96	73.5

Πίνακας 6.9: Πειράματα διαφορετικών μεθόδων προεπεξεργασίας των εικόνων

Ενώ γενικότερα στη βιβλιογραφία η προσθήκη τυχαιότητας στην προεπεξεργασία των εικόνων βοηθάει το μοντέλο στο να γενικεύει και να προβλέπει αποτελεσματικότερα, φαίνεται στη περίπτωσή μας πως το μοντέλο δυσκολεύεται υπερβολικά

6.2. ΠΕΙΡΑΜΑΤΙΣΜΟΣ ΜΕ ΤΗΝ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΕΙΚΟΝΩΝ

και οι επιδόσεις του μειώνονται. Καθώς προχωράμε στα πειράματα και άρα αυξάνει η τυχαιότητα που προστίθεται στο μοντέλο, σταδιακά η ακρίβεια και το F1 σκορ στο test set μειώνονται, με εξαίρεση το τελικό πείραμα που υπάρχει μια μικρή αύξηση αλλά και πάλι δε μπορεί να φτάσει τις επιδόσεις του αρχικού πειράματος.

7

Επεκτάσεις

Το μοντέλο τα έχει πάει εξαιρετικά στο dataset στο οποίο εκπαιδεύτηκε. Το dataset περιείχε κείμενα όμως μόνο στα αγγλικά. Στο κεφάλαιο αυτό θα γίνει έρευνα σχετικά με το αν το μοντέλο είναι ικανό να χρησιμοποιηθεί και σε κείμενα άλλων γλωσσών, πέραν των αγγλικών.

Συνεχίζοντας, δημιουργείται, για τους σκοπούς της διπλωματικής εργασίας και τον έλεγχο της γενίκευσης σε εντελώς άγνωστα δεδομένα, ένα ελληνικό dataset.

Εκτός των παραπάνω, δημιουργείται ιστοσελίδα στην οποία μπορεί να δοθεί είσοδος (κείμενο, εικόνα ή και τα δύο) και χρησιμοποιώντας τα εκπαιδευμένα μοντέλα που επιλέχθηκαν, να επιστραφεί στο χρήστη η ανάλυση συναίσθηματος για την είσοδο που έδωσε.

7.1 ΔΟΚΙΜΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΕ ΔΙΑΦΟΡΕΤΙΚΕΣ ΓΛΩΣΣΕΣ

Σε μια ακόμη προσπάθεια γενίκευσης του προτεινόμενου συστήματος, ελέγχθηκε η ικανότητά του να προβλέπει σωστά το συναίσθημα όταν η είσοδος είναι σε άλλη γλώσσα, πέραν των αγγλικών. Για τον σκοπό αυτό μεταφράσαμε το αρχικό dataset σε διαφορετικές γλώσσες, με χρήση του μεταφραστή της Google μέσω του API του, από τη βιβλιοθήκη `deep_translator` της Python.

Οι γλώσσες στις οποίες έγινε η μετάφραση είναι τα ελληνικά, τα γερμανικά, τα ρωσικά, τα ισπανικά και τα ινδικά (`hindi`). Το μοντέλο κλήθηκε αρχικά να αναγνωρίσει το συναίσθημα όταν η είσοδος είναι διαφορετική από τη γλώσσα στην οποία εκπαιδεύτηκε. Στον [πίνακα 7.1](#) φαίνονται τα αποτελέσματα του πειράματος στο οποίο το μοντέλο κειμένου εκπαιδεύτηκε στα αγγλικά αλλά αναγνώρισε το συναίσθημα κειμένων διαφορετικής γλώσσας. Χρησιμοποιήθηκε το μοντέλο `xlm-roberta-base` [73], που είναι το αντίστοιχο γλωσσικό μοντέλο του `roberta-base`, προεκπαιδευμένο σε κείμενα διαφόρων γλωσσών.

Η πρώτη παρατήρηση που μπορεί να γίνει είναι πως το πολυτροπικό σύστημα

ΚΕΦΑΛΑΙΟ 7. ΕΠΕΚΤΑΣΕΙΣ

Γλώσσα	Multimodal Accuracy score %	Multimodal F1 Score %
Αγγλικά	73.84	73.49
Ελληνικά	72.28	71.95
Γερμανικά	72.51	72.2
Ισπανικά	70.51	70.08
Ρωσικά	72.28	71.91
Ινδικά	70.95	70.27

Πίνακας 7.1: Εκπαίδευση στα αγγλικά κείμενα, πρόβλεψη σε όλες τις γλώσσες

αποφασίζει ορθότερα στη γλώσσα στην οποία έγινε fine-tuning το μοντέλο κειμένου. Παρόλα αυτά όμως, είναι ικανό να δώσει προβλέψεις σχεδόν εξίσου καλές, με μια μικρή μείωση στις μετρικές της τάξης του 1-3%, και για τις υπόλοιπες γλώσσες. Το επόμενο πείραμα ελέγχει αν αυτό μπορεί να συμβεί με οποιαδήποτε γλώσσα επιλεγεί για την αρχική εκπαίδευση. Τώρα λοιπόν, στον [πίνακα 7.2](#), το μοντέλο κειμένου εκπαιδεύεται στα ελληνικά και καλείται να αναγνωρίσει το συναίσθημα στις υπόλοιπες γλώσσες. Κατά την εκπαίδευση, η ακρίβεια στα ελληνικά είναι χαμηλότερη απότι ήταν στα αγγλικά κείμενα.

Γλώσσα	Multimodal Accuracy score %	Multimodal F1 Score %
Αγγλικά	71.39	71.06
Ελληνικά	72.51	72.21
Γερμανικά	71.4	71.12
Ισπανικά	71.84	71.32
Ρωσικά	70.29	69.9
Ινδικά	71.84	71.17

Πίνακας 7.2: Εκπαίδευση στα ελληνικά κείμενα, πρόβλεψη σε όλες τις γλώσσες

Οι παρατηρήσεις που έγιναν προηγουμένως, ισχύουν και σε αυτή τη πειραματική διάταξη. Όπως είναι αναμενόμενο, την υψηλότερη επίδοση έχει το μοντέλο στα ελληνικά, δηλαδή τη γλώσσα στην οποία εκπαιδεύτηκε. Παρόλα αυτά, και οι υπόλοιπες γλώσσες βρίσκονται στο εύρος 1-3% πιο κάτω, μια αποδεκτή απώλεια λόγω της διαφορετικής γλώσσας. Το τρίτο και τελευταίο πείραμα προς αυτή τη κατεύθυνση χρησιμοποιεί ταυτόχρονα τα αγγλικά και ελληνικά κείμενα για την εκπαίδευση του xlm-roberta. Πιο συγκεκριμένα, το σετ εκπαίδευσης αποτελείται από το σετ εκπαίδευσης των αγγλικών και της μεταφράσης τους στα ελληνικά, έχει δηλαδή διπλάσιο μέγεθος συγκριτικά με τα προηγούμενα πειράματα. Τα αποτελέσματα του βρίσκονται στον [πίνακα 7.3](#).

Τα αποτελέσματα του πειράματος έχουν ιδιαίτερο ενδιαφέρον. Τα κείμενα των αγγλικών και των ελληνικών πετυχαίνουν τις μεγαλύτερες επιδόσεις τους αλλά συμπεριλαμβάνονται και στα δεδομένα εκπαίδευσης. Με αυτό το τρόπο μάλιστα τα αγγλικά και ελληνικά κείμενα ξεπερνούν τις επιδόσεις που είχαν στο βασικό σύστημα που έχουμε προτείνει. Οι υπόλοιπες γλώσσες φαίνεται να μην επηρεάζονται θετικά ή αρνητικά από τον διπλασιασμό του συνόλου εκπαίδευσης.

7.2. ΠΕΙΡΑΜΑΤΙΣΜΟΣ ΜΕ ΤΟ ΕΛΛΗΝΙΚΟ BERT

Γλώσσα	Multimodal Accuracy score %	Multimodal F1 Score %
Αγγλικά	74.78	73.77
Ελληνικά	75.66	74.72
Γερμανικά	71.39	70.85
Ισπανικά	70.51	69.87
Ρωσικά	72.06	71.63
Ινδικά	72.51	71.83

Πίνακας 7.3: Εκπαίδευση και στα αγγλικά και στα ελληνικά κείμενα, πρόβλεψη σε όλες τις γλώσσες

7.2 ΠΕΙΡΑΜΑΤΙΣΜΟΣ ΜΕ ΤΟ ΕΛΛΗΝΙΚΟ BERT

Στη παρούσα ενότητα θα χρησιμοποιηθεί το *greek bert* [74], έκδοση του bert εκπαιδευμένη σε ελληνικά δεδομένα. Το **ελληνικό BERT** είναι διαθέσιμο στο Hugging Face όπως και τα προηγούμενα προεκπαιδευμένα μοντέλα που χρησιμοποιήθηκαν στην εργασία. Για τη σύγχριση εκτελούνται 4 πειράματα και συγκρίνεται με το BERT μοντέλο κειμένου bert-base-uncased, καθώς θέλουμε να εστιάσουμε στη λειτουργικότητά του απέναντι στην αρχική έκδοση και κατά πόσο αυξάνει την απόδοση σε ελληνική είσοδο. Τα πειράματα συνοπτικά είναι:

- Πείραμα 0: Εκπαίδευση του bert με αγγλικό κείμενο
- Πείραμα 1: Εκπαίδευση του bert με ελληνικό κείμενο
- Πείραμα 2: Εκπαίδευση του greek bert με ελληνικό κείμενο
- Πείραμα 3: Εκπαίδευση του greek bert με αγγλικό κείμενο

Τα αποτελέσματα των πειραμάτων συνοψίζονται στον [πίνακα 7.4](#). Από το πείραμα 1 είναι εμφανής η δυσκολία του bert να αναγνωρίσει ελληνικό κείμενο, αφού δεν έχει εκπαιδευτεί σε αυτή τη γλώσσα. Η χαμηλή απόδοση του μοντέλου κειμένου επηρεάζει και ολόκληρο το πολυτροπικό σύστημα μειώνοντας την ακρίβεια περίπου στο 66%, ενώ όταν η είσοδος είναι στα αγγλικά η ακρίβεια έχει τιμή ίση με 69.14%. Δίνοντας τις δύο γλώσσες στο greek bert, καταφέρνει να εκπαιδευτεί σωστά και στις δύο περιπτώσεις (πειράματα 2 και 3), επιτυγχάνοντας παρόμοιες επιδόσεις και στο συνολικό σύστημα.

Αριθμός πειράματος	0	1	2	3
Text Accuracy score	64.75	50.33	56.42	53.67
Text F1 score	62.95	44.39	56.46	59.2
Multimodal Accuracy score	69.62	66.96	71.4	69.4
Multimodal F1 score	69.14	66.14	70.66	68.36

Πίνακας 7.4: Σύγχριση των μοντέλων κειμένου bert και greek bert

7.3 ΔΗΜΙΟΥΡΓΙΑ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Το σύστημα καταφέρνει να γενικεύσει σε διαφορετικές γλώσσες. Εστιάζοντας στην ελληνική γλώσσα, δημιουργούμε ένα νέο dataset που αποτελείται από ζεύγη κειμένου-εικόνας, προερχόμενα από το ελληνικό Twitter. Το μέγεθος του σετ είναι μικρό (260 ζεύγη) καθώς ο στόχος του είναι να χρησιμοποιηθεί μόνο ως σετ ελέγχου στα ήδη εκπαιδευμένα μοντέλα. Συλλέχθηκε χειροκίνητα από το ελληνικό Twitter και περιλαμβάνει tweets δημοσιευμένα από τον Απρίλιο εώς τον Σεπτέμβριο του 2023. Οι κύριες περιοχές ενδιαφέροντος του dataset είναι η πολιτική, ο αθλητισμός και η επικαιρότητα.

Έπειτα από τη συλλογή των δεδομένων, ακολουθήθηκε η διαδικασία της επισήμανσης (labeling). Το labeling είναι μια απαιτητική διαδικασία, καθώς υπάρχουν αρκετές περιπτώσεις που είναι δύσκολο για τον άνθρωπο ακόμη να ξεχωρίσει το συναίσθημα που προσπαθεί να αποδώσει το κείμενο ή η εικόνα. Τέτοια φαινόμενα μπορούν να παρατηρηθούν σε περιπτώσεις ειρωνείας, χιούμορ που είναι πολύ συχνά στα κοινωνικά δίκτυα. Επιπρόσθετα, ο άνθρωπος που αποδίδει τα labels πρέπει να είναι όσο το δυνατόν πιο αντικειμενικός και να προσεγγίζει τα δεδομένα όχι με τη προσωπική του αντίληψη, αλλά με τον τρόπο που αποδόθηκαν από τον συγγραφέα τους. Οι δημοσιεύσεις [75], [76] και [77] αναδεικνύουν το πρόβλημα και προσπαθούν να προτείνουν λύσεις και κατευθυντήριες γραμμές για τους annotators. Επιπρόσθετα έχουν δημιουργηθεί εργαλεία που βασίζονται σε κανόνες ή στη μηχανική μάθηση για αυτόματο labelling [78], ειδικά όταν το dataset περιέχει πολλά δεδομένα και η διαδικασία είναι υπερβολικά χρονοβόρα.

Στο labelling συμμετείχαν δύο άνθρωποι, οι οποίοι κλήθηκαν να ταξινομήσουν το συναίσθημα ανάμεσα στις κατηγορίες θετικό, αρνητικό και ουδέτερο, όπως ακριβώς και στο αρχικό σετ δεδομένων, δίνοντας labels ξεχωριστά για το κείμενο, την εικόνα και συνολικά για το ζεύγος. Γρήγορα αναδείχθηκε το πρόβλημα που αναφέρθηκε παραπάνω, καθώς ακόμη και ανάμεσα σε δύο μόνο annotators υπήρξαν πολλές διαφορετικές απόψεις. Στον πίνακα 7.5 φαίνονται το πλήθος των conflicts για κάθε κατηγορία. Ο αριθμός των ζευγών στα οποία δεν προκύπτει κάποια σύγκρουση απόψεων είναι 179 από τα 260, δηλαδή σε ποσοστό είναι 68.85%.

	Κείμενο	Εικόνα	Κείμενο + Εικόνα
Αριθμός conflicts	35	47	40
Ποσοστό %	13.46	18.08	15.38

Πίνακας 7.5: Συγκρούσεις στα labels μεταξύ των 2 annotators

Οι συγκρούσεις απόψεων μεταξύ των annotators βρίσκεται ανάμεσα στις κατηγορίες ουδέτερο και μία εκ των θετικού και αρνητικού συναίσθηματος. Δεν υπάρχει δηλαδή κάποιο δείγμα του συνόλου δεδομένων όπου ο ένας annotator σημειώνει θετικό και ο άλλος αρνητικό συναίσθημα. Ένα ενδεικτικό παράδειγμα φαίνεται στο σχήμα 7.1, όπου δείχνει την εικόνα και στη περιγραφή βρίσκεται το κείμενο που την συνοδεύει. Οι ετικέτες που αποδίδει ο κάθε annotator φαίνεται στον πίνακα 7.6, όπου στη τελευταία γραμμή δίνεται και το τελικό annotation που προκύπτει με

7.3. ΔΗΜΙΟΥΡΓΙΑ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

τη συμπλήρωση σύμφωνα με τους κανόνες που χρησιμοποιήθηκαν και στο MVSA-Single.



Σχήμα 7.1: Κείμενο: "...οι μόνοι που έχουμε το προνόμιο να λέμε τον ουρανό "ουρανό" και την θάλασσα "θάλασσα", όπως την έλεγαν ο Ομηρος και ο Πλάτωνας πριν από δίαμισυ χιλιάδες χρόνια"..., Οδυσσέας Ελύτης, Τελευταία δύση του Αυγούστου

	Κείμενο	Εικόνα	Κείμενο + Εικόνα
Annotator 1	Ουδέτερο	Θετικό	Θετικό
Annotator 2	Ουδέτερο	Ουδέτερο	Ουδέτερο
Final Annotation	Ουδέτερο	Θετικό	Θετικό

Πίνακας 7.6: Παράδειγμα σύγκρουσης απόψεων μεταξύ των δύο annotators

Εξαιτίας των πολλών συγκρούσεων, ιρίθηκε απαραίτητο να γίνει ένα τελικό labelling. Στη πρώτη προσέγγιση, ένας από τους δύο annotators ανέλαβε να επανελέγξει όλα τα δεδομένα που περιλάμβαναν conflicts. Στον πίνακα 7.7 γίνεται για πρώτη φορά φανερό το πρόβλημα γενίκευσης στις νέες εικόνες. Η ανικανότητα του μοντέλου να ταξινομήσει τις εικόνες επηρεάζει συνολικά το πολυτροπικό σύστημα, αφού η έξοδος του επίσης αποτυγχάνει (είναι ελάχιστα πάνω από τη τυχαία επιλογή).

Η αμέσως επόμενη σκέψη ήταν να εξαιρεθούν τελείως τα conflicts από το dataset, οδηγώντας σε ένα σετ δεδομένων αρκετά μικρότερο. Η κίνηση αυτή παραδόξως

ΚΕΦΑΛΑΙΟ 7. ΕΠΕΚΤΑΣΕΙΣ

	Κείμενο	Εικόνα	Κείμενο + Εικόνα
Accuracy	72.31	42.31	43.85
F1	72.05	36.39	35.39

Πίνακας 7.7: Χρήση τρίτου annotator

βοήθησε μόνο το κείμενο, αφού η αύξηση στην εικόνα είναι ελάχιστη, όπως φαίνεται στον [πίνακα 7.8](#).

	Κείμενο	Εικόνα	Κείμενο + Εικόνα
Accuracy	75.42	44.69	44.69
F1	74.88	40.79	36.95

Πίνακας 7.8: Εξαίρεση των δεδομένων που προκαλούν συγκρούσεις από το σετ δεδομένων

Η τελευταία προσέγγιση που δοκιμάστηκε ήταν τα conflicts να συμπληρωθούν όπως ακριβώς και στον καθαρισμό του MVSA-Single dataset, με τα αποτελέσματα να βρίσκονται στον [πίνακα 7.9](#). Για υπενθύμιση του αναγνώστη, τα βήματα για τον καθαρισμό έιναι τα εξής:

1. Αν το label του κειμένου είναι ίδιο με το label της εικόνας, τότε και το συνολικό label θα είναι το ίδιο.
2. Αν ένα από τα δύο label είναι ουδέτερο και το άλλο είναι θετικό ή αρνητικό, τότε το συνολικό label θα είναι ίδιο με αυτό του μη ουδέτερου.
3. Αν το ένα label είναι θετικό και το άλλο αρνητικό, τότε το συνολικό label είναι άγνωστο και το δείγμα αφαιρείται από το dataset.

Πράγματι, η συγκεκριμένη προσέγγιση βελτίωσε το μοντέλο της εικόνας, αλλά και πάλι οι μετρικές είναι πολύ χαμηλές.

	Κείμενο	Εικόνα	Κείμενο + Εικόνα
Accuracy	71.21	47.86	47.86
F1	70.93	43.4	40.58

Πίνακας 7.9: Συμπλήρωση των δεδομένων που προκαλούν συγκρούσεις όπως στο MVSA-Single

7.4 ΔΗΜΙΟΥΡΓΙΑ ΙΣΤΟΣΕΛΙΔΑΣ

Για την επέκταση των δυνατοτήτων και χρήσεων του πολυτροπικού συστήματος που αναπτύχθηκε, αποφασίστηκε η δημιουργία ιστοσελίδας. Η ιστοσελίδα δημιουργήθηκε με τέτοιο τρόπο ώστε να μπορεί να δέχεται σαν είσοδο είτε κείμενο, είτε

εικόνα, είτε και τα δύο modalities ταυτόχρονα. Επιπλέον, αναγνωρίζει αυτόματα την γλώσσα εισόδου και αποδέχεται τα ελληνικά και τα αγγλικά.

Σε αντιστοίχιση με το υποκεφάλαιο 5.1 οι βιβλιοθήκες και τα εργαλεία που αξιοποιηθηκαν φαίνονται στον πίνακα 7.10. Το Flask [79] είναι framework της Python, γνωστό για την απλότητα και την ευελιξία που προσφέρει στον προγραμματιστή. Χρησιμοποιείται για τη δημιουργία διαδικτυακών εφαρμογών και APIs. Το Flask επιτρέπει στον χρήστη να οργανώσει τον κώδικα με τον τρόπο που εκείνος επιθυμεί, αφού του παρέχει τα εργαλεία και τις βιβλιοθήκες που χρειάζεται, χωρίς να επιβάλλει συγκεκριμένη δομή. Το Flask μπορεί να χρησιμοποιηθεί για δημιουργία διαδρομών (routing), επικοινωνία με HTTP αιτήματα και αποκρίσεις (Handle HTTP Requests/Responses) και για διαχείριση βάσεων δεδομένων. Στη διπλωματική εργασία το Flask δημιουργήθηκε για να διαχειρίζεται τα αιτήματα πρόσβασης στην ιστοσελίδα και ανάλογα με τις εισόδους την επιστροφή των κατάλληλων HTML ιστοσελίδων.

Name	Version	Description
Flask	2.3.3	Δημιουργία διαδικτυακών εφαρμογών και APIs
io	3.10.4	Εργαλεία για εργασία με streams πληροφορίας
uuid	3.10.4	Δημιουργία μοναδικών IDs
json	3.10.4	Κωδικοποίηση και αποκωδικοποίηση JSON
apscheduler	3.10.4	Εκτέλεση περιοδικά jobs στο παρασκήνιο του server
werkzeug	2.3.7	Χρήσιμα εργαλεία για δημιουργία ιστοσελίδας

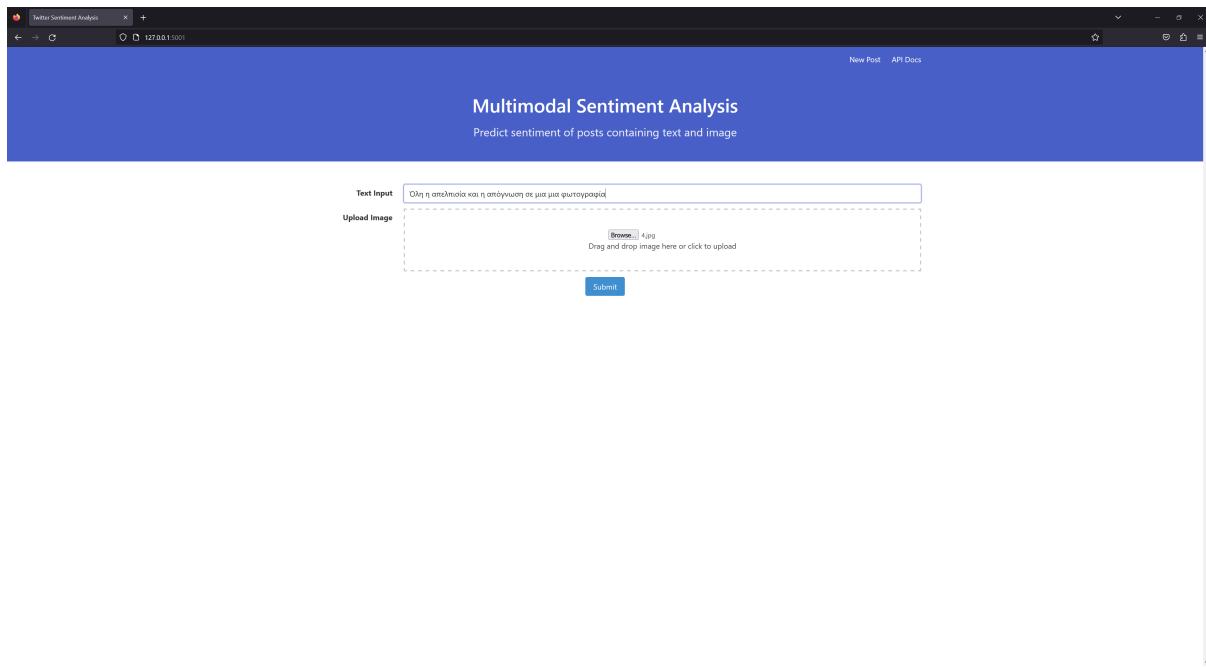
Πίνακας 7.10: Βιβλιοθήκες της Python που χρησιμοποιήθηκαν για την υλοποίηση της ιστοσελίδας

Εκτός των βιβλιοθηκών που φαίνονται στον πίνακα, χρησιμοποιούνται και αρκετές από τις βιβλιοθήκες που είχαν αναλυθεί στην υπόλοιπη διπλωματική για τη δημιουργία και έλεγχο του πολυτροπικού συστήματος. Για το σχεδιασμό της ιστοσελίδας χρησιμοποιήθηκαν οι γλώσσες HTML/CSS, ενώ το CSS ενισχύθηκε με το Bulma Framework [80]. Τέλος, ελάχιστα χρησιμοποιήθηκε η γλώσσα προγραμματισμού Javascript, για τη δημιουργία συναρτήσεων που αλλάζουν τη μορφή της ιστοσελίδας δυναμικά. Παρακάτω δίνονται διάφορα screenshots που αναδεικνύουν τη χρήση της ιστοσελίδας.

Η ιστοσελίδα αποτελείται από δύο σελίδες. Στην αρχική σελίδα, δίνεται ως είσοδος ένα ζεύγος κειμένου-εικόνας από το custom dataset που δημιουργήθηκε (σχήμα 7.2).

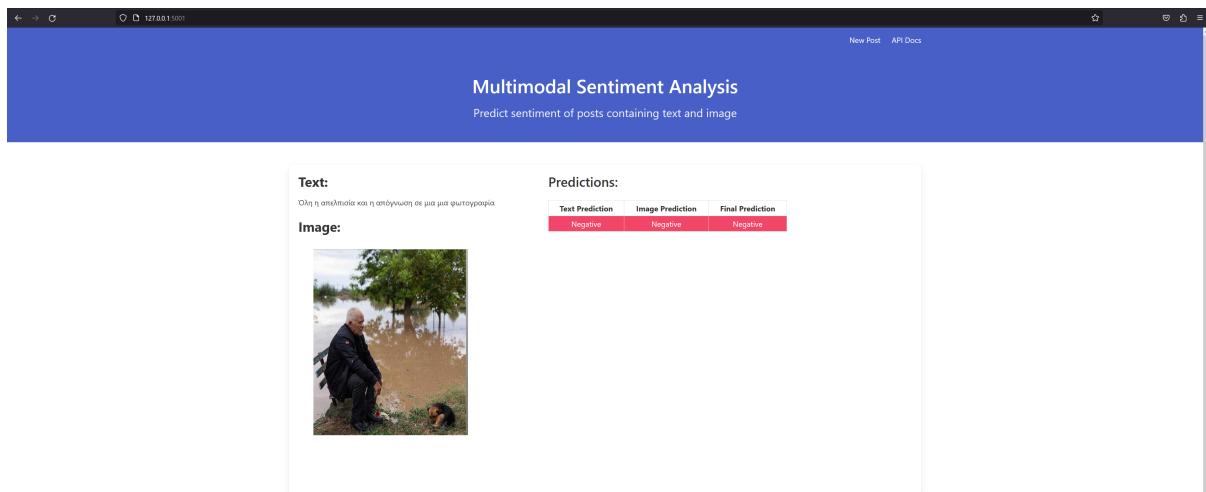
Έπειτα, η σελίδα ανανεώνεται και επιστρέφει στο χρήστη το συναίσθημα, ξεχωριστά για το κείμενο, για την εικόνα και το ζεύγος πληροφορίας. Στην αριστερή στήλη φαίνονται ξανά τα δεδομένα που δόθηκαν σαν είσοδος. Φαίνεται στο

ΚΕΦΑΛΑΙΟ 7. ΕΠΕΚΤΑΣΕΙΣ



Σχήμα 7.2: Βασική σελίδα της ιστοσελίδας. Ο χρήστης εισάγει το κείμενο και την εικόνα που επιθυμεί.

σχήμα 7.3 πως το σύστημα αποφάσισε πως και το κείμενο και η εικόνα και ο συνδυασμός τους έχουν αρνητικό συναίσθημα.

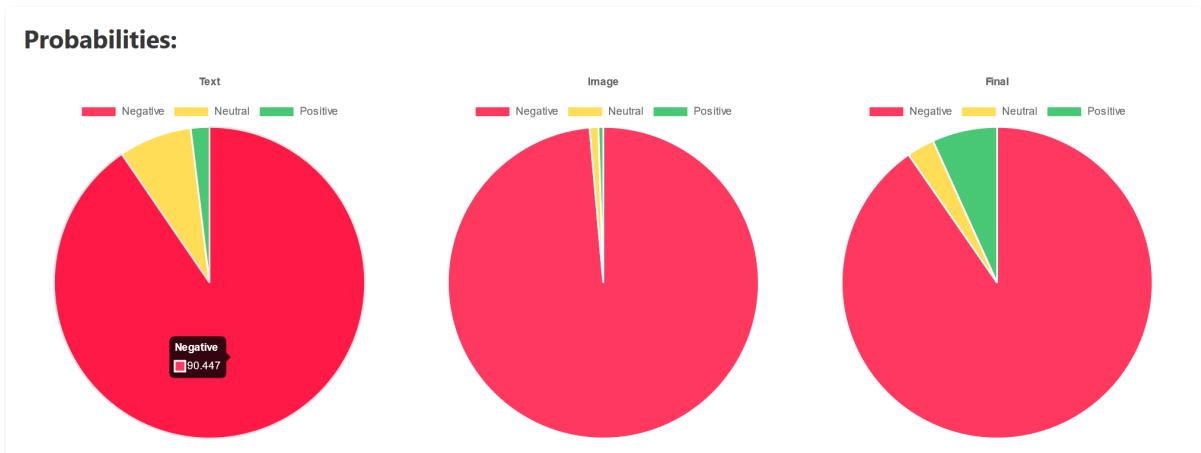


Σχήμα 7.3: Τα αποτελέσματα της ανάλυσης συναίσθημάτος στα δεδομένα εισόδου

Εκτός όμως από τις προβλέψεις, η ιστοσελίδα δείχνει στον χρήστη και διαγράμματα σε μορφή πίτας, που περιέχουν αναλυτικά τις πιθανότητες της κάθε πρόβλεψης (σχήμα 7.4).

Όπως αναφέραμε, η ιστοσελίδα μπορεί να διαχειριστεί κείμενα και της αγγλικής γλώσσας, καθώς και την απουσία κάποιου modality. Στο σχήμα 7.5 φαίνεται πως προσαρμόζονται τα αποτελέσματα στην ιστοσελίδα στη περίπτωση που ο χρήστης δεν επιθυμεί να συνοδέψει το κείμενο με κάποια εικόνα.

7.4. ΔΗΜΙΟΥΡΓΙΑ ΙΣΤΟΣΕΛΙΔΑΣ

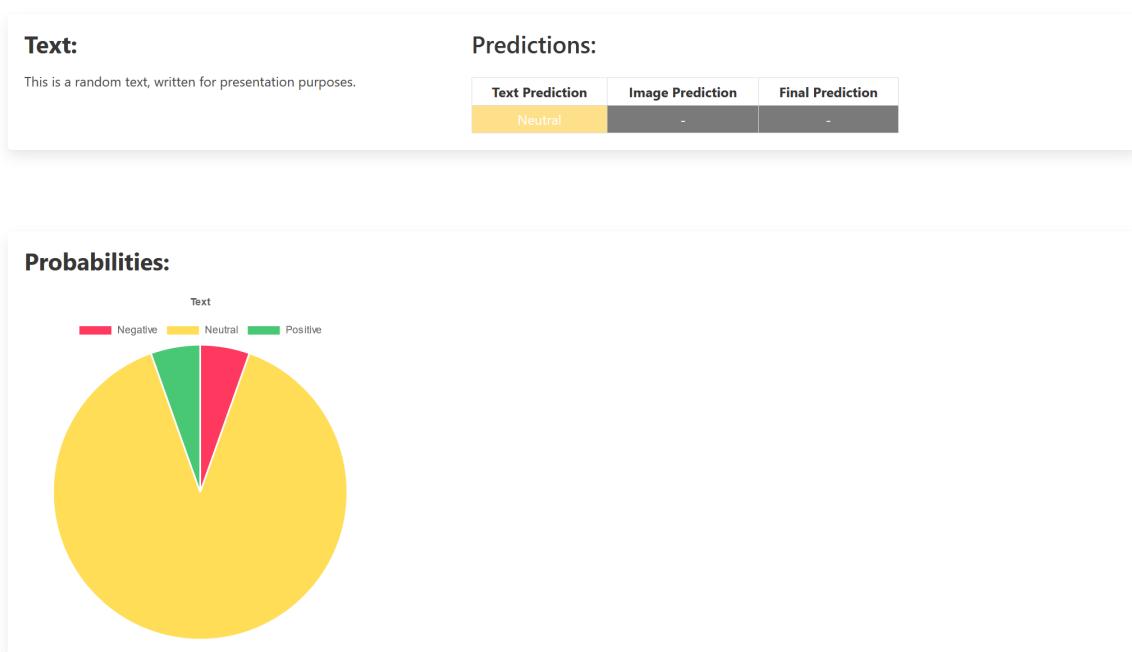


Σχήμα 7.4: Γραφήματα με τις πιθανότητες της κάθε πρόβλεψης

New Post API Docs

Multimodal Sentiment Analysis

Predict sentiment of posts containing text and image



Σχήμα 7.5: Αποτελέσματα ανάλυσης συναισθήματος σε είσοδο κειμένου, απουσία εικόνας

8

Συμπεράσματα

Στα προηγούμενα κεφάλαια αναλύθηκαν όλα τα διαθέσιμα μοντέλα, παρουσιάστηκαν αποτελέσματα πειραμάτων, προτάθηκε και υλοποιήθηκε το πολυτροπικό σύστημα ανάλυσης συναίσθημάτων και έπειτα αναζητήθηκαν και συζητήθηκαν επεκτάσεις του. Στο κεφάλαιο αυτό γίνεται μια γενική ανακεφαλαίωση στα παραπάνω θέματα, κάνοντας μια σύνοψη στα κυριότερα συμπεράσματα που εξάγονται από ολόκληρη την ανάλυση και εξερεύνηση. Επιπλέον συζητώνται τα διάφορα προβλήματα που αντιμετωπίσθηκαν στη πορεία και ορισμένα που εξακολουθούν να παραμένουν άλυτα.

8.1 ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Θα δοθούν συμπεράσματα για τις κυριότερες κατηγορίες της υλοποίησης, συνοψίζοντας διάφορα πειράματα των προηγουμένων κεφαλαίων:

- **Μοντέλο κειμένου:** Τα μοντέλα που μπορούν να χρησιμοποιηθούν για τη ταξινόμηση των εισόδων κειμένου είναι παρόμοια και ανήκουν στη οικογένεια των BERT μοντέλων. Επιλέχθηκε ως καταλληλότερο (όχι δεσμευτικά) το μοντέλο RoBERTa, ενώ έγινε φανερό πως υπάρχουν διαθέσιμα μοντέλα για οποιαδήποτε γλώσσα δοθεί σαν είσοδος, αρκεί το μοντέλο να έχει προεκπαιδευτεί και προσαρμοστεί σε αυτήν. Ακόμη και να μην έχει προσαρμοστεί στη γλώσσα εισόδου, το μοντέλο θα καταφέρει να αναλύσει το συναίσθημα, έχοντας μία μικρή επίπτωση στην ακρίβεια του.
- **Χρήση λεξικού:** Η χρήση λεξικού ή αντίστοιχου εργαλείου για την ενίσχυση των προβλέψεων, όσον αφορά την είσοδο κειμένου, φαίνεται πως δε βοηθάει ιδιαίτερα το σύστημα, ανεβάζοντας τις επιδόσεις του σε αμελητέο βαθμό. Παρόλο λοιπόν που δοκιμάστηκε η υβριδική προσέγγιση, επιλέχθηκε τελικά να παραλειφθεί οποιοδήποτε λεξικό που λειτουργεί βοηθητικά.

- **Προεπεξεργασία εικόνας:** Παραδόξως φάνηκε πως τα μοντέλα εικόνας λειτουργούν καλύτερα με ελάχιστη προεπεξεργασία στην εικόνα, εφαρμόζοντας μόνο την απαραίτητη για τη τροφοδοσία τους στα μοντέλα της οικογένειας των transformers.
- **Μοντέλο εικόνας:** Δόθηκε έμφαση στην επιλογή του μοντέλου εικόνας, με εκτενείς πειραματισμούς, χρησιμοποιώντας μοντέλα που βασίζονται στην αρχιτεκτονική των συνελικτικών δικτύων αλλά και μοντέλα που βασίζονται στους transformers. Το κυριότερο συμπεράσμα είναι πως το ViT μοντέλο ξεπέρασε σημαντικά όλα τα υπόλοιπα που συμμετείχαν στα πειράματα. Επίσης φάνηκε από τα αποτελέσματα των πειραμάτων πως τα μεγάλα και πιο σύνθετα μοντέλα δε κατάφερναν να ενισχύσουν τις επιδόσεις του μοντέλου, ενώ σε αρκετές περιπτώσεις μείωναν την ακρίβεια του.
- **Μοντέλο συνένωσης:** Για τη συνένωση των διαφορετικών modalities αποφασίστηκε να χρησιμοποιηθεί η τεχνική early fusion. Από τις διαθέσιμες αρχιτεκτονικές αποδείχθηκε ισχυρότερη η υλοποίηση ενός πλήρως συνδεδεμένου νευρωνικού δικτύου με 2 κρυφά επίπεδα. Ολοκληρώνεται λοιπόν έτσι το πολυτροπικό σύστημα ανάλυσης συναισθήματος, καταφέρνοντας να πετύχει ακρίβεια περίπου ίση με 74.79% σε άγνωστα δεδομένα του dataset.

8.2 ΠΡΟΒΛΗΜΑΤΑ

Η διπλωματική εργασία εκτάθηκε σε διάφορους τομείς της σύγχρονης βαθιάς μάθησης. Κατά τη διάρκεια της διπλωματικής προέκυψαν προβλήματα τα οποία λύθηκαν και προβλήματα που εξακολουθούν να χρήζουν επίλυσης. Στο υποκεφάλαιο αυτό βλέπουμε τα προβλήματα των παραπάνω κατηγοριών, ενώ στο επόμενο κεφάλαιο προτείνονται πιθανές λύσεις και μελλοντικές επεκτάσεις γενικότερα για τη συνέχιση και εξέλιξη των υλοποιήσεων.

- **Υπολογιστικοί πόροι:** Τα χρησιμοποιούμενα μοντέλα απαιτούν υψηλή υπολογιστική ισχύ. Είναι αναγκαία η χρήση GPU για την παραλληλοποίηση των εργασιών. Ακόμη όμως και με χρήση GPU, για παράδειγμα από τις διαθέσιμες στο Google Colab, υπάρχει θέμα με τη μνήμη που καταναλώνεται, είτε αυτή είναι μνήμη της κάρτας γραφικών είτε του συστήματος. Επιπρόσθετα, οι περιορισμένοι πόροι απέτρεψαν τις όποιες προσπάθειες επέκτασης και σε άλλα σετ δεδομένων (για παράδειγμα το MVSA-Multiple δοκιμάστηκε), καθώς ο κώδικας δε μπορούσε να τερματίσει λόγω υπερφόρτωσης της μνήμης. Σε προσπάθεια επίλυσης του ζητήματος έγιναν αρκετές βελτιστοποιήσεις στον κώδικα ώστε να είναι όσο το δυνατόν γρηγορότερος και χαμηλής υπολογιστικής πολυπλοκότητας.
- **Χρόνος εκμάθησης:** Οι χρόνοι εκμάθησης των μοντέλων είναι υψηλοί, ακόμη και για λίγες εποχές. Επιπλέον, το Google Colab εισάγει χρονικούς περιορισμούς στη χρήση των GPU που διαθέτει. Τα παραπάνω δυσκολεύουν τη διαδικασία της επιλογής των καταλληλότερων παραμέτρων για το κάθε μοντέλο. Προς αντιμετώπιση του προβλήματος, πειράματα γινόταν και στη πλατφόρμα

του Kaggle, απαιτώντας όμως καλύτερη διαχείριση των αρχείων, με κοινή χρήση μεταξύ λογαριασμών ή αποθήκευσή τους σε cloud υπηρεσίες και τον έλεγχο των εκδόσεων μεταξύ των διαφορετικών πλατφόρμων (version control).

- Δημιουργία του νέου dataset: Επισημάνθηκαν ήδη τα προβλήματα που αντιμετωπίστηκαν κατά τη δημιουργία των σετ δεδομένων. Ιδιαίτερο πρόβλημα υπήρξε στο labelling, όπου φάνηκε η υποκειμενικότητα του συναισθήματος στο πως το εκλαμβάνει ξεχωριστά ο κάθε άνθρωπος, όσο αντικειμενικός και να προσπαθεί να είναι.
- Γενίκευση: Στα πειράματα γενίκευσης που βασιζόταν σε χρήση διαφορετικών γλωσσών στην είσοδο το πολυτροπικό σύστημα ανταπεξήλθε ικανοποιητικά. Ωστόσο όταν κλήθηκε να αναλύσει συναίσθημα σε τελείως καινούριο σετ δεδομένων είχε πρόβλημα γενίκευσης σε άγνωστα δεδομένα, λόγω των διαφορών που εντοπίζονται στην αποτύπωση του συναισθήματος της εικόνας στα δύο datasets.
- Χρήση state of the art τεχνολογιών: Η διπλωματική εργασία ακολουθεί τις τελευταίες εξελίξεις στους τομείς της επεξεργασίας φυσικής γλώσσας και της υπολογιστικής όρασης. Αυτόματα το γεγονός αυτό εισάγει προκλήσεις κατά το implementation των μοντέλων στον κώδικα.

9

Μελλοντικές επεκτάσεις

Στο παρόν κεφάλαιο θα αναφερθούν αρχικά κάποιες προτεινόμενες λύσεις για τα άλυτα προβλήματα που αναφέρθηκαν στην προηγούμενη ακριβώς ενότητα. Ξεκινώντας από το κυριότερο πρόβλημα των υπολογιστικών πόρων, η προφανής και ευκολότερη λύση είναι η αξιοποίηση κάποιου server που να επιτρέπει τη χρήση μεγαλύτερης υπολογιστικής ισχύος. Εκεί θα μπορεί να ελεγχθεί και να επαναδημιουργηθεί η προτεινόμενη υλοποίηση και σε διαφορετικά dataset.

Όσον αφορά το πρόβλημα του νέου dataset, μία πρώτη προσέγγιση θα ήταν να συμπεριληφθούν περισσότεροι άνθρωποι στο annotation του, για παράδειγμα 5, ώστε να απαλείφονται προβλέψεις που βασίζονται στην υποκειμένικη άποψη και αντίληψη.

Από το πολυτροπικό σύστημα στο σύνολό του, το μοντέλο της εικόνας φάνηκε να υστερεί συγκριτικά με την υπόλοιπη δομή. Αξίζει λοιπόν να πραγματοποιηθεί ξανά εκτενής έρευνα στο καταλληλότερο μοντέλο, λαμβάνοντας υπόψη και τις νέες εξελίξεις της τεχνολογίας. Κατά τη διάρκεια της διπλωματικής συνεχώς δημοσιεύονταν επιστημονικά άρθρα και έρευνες με νέες προσεγγίσεις και ιδέες πάνω στο αντικείμενο της ανάλυσης συναισθημάτων. Επιπρόσθετα, διαρκώς εμφανίζονται νέες αρχιτεκτονικές μοντέλων στην επιστημονική κοινότητα που θα μπορούσαν να δώσουν λύση στο πρόβλημα του εντοπισμού συναισθήματος στην εικόνα.

Πέραν της εικόνας, το μοντέλο της συνένωσης δε καταφέρνει να ανεβάσει σε σημαντικό βαθμό τις μετρικές, συγκριτικά με το κάθε modality μεμονωμένα. Προτείνεται λοιπόν η περαιτέρω αναζήτηση για αρχιτεκτονικές που μπορούν να χρησιμοποιηθούν στο fusion.

Ακόμη μία ενδιαφέρουσα επέκταση στην παρούσα υλοποίηση είναι η ενσωμάτωση και άλλων modalities, πέραν του κειμένου και της εικόνας. Για παράδειγμα στο [υποκεφάλαιο 4.2](#) ονομάστηκαν ενδεικτικά ορισμένα datasets που περιέχουν βίντεο και ήχο.

Στο κομμάτι της ιστοσελίδας, μπορούν να γίνουν βελτιώσεις σχετικά με την απεικόνιση της πληροφορίας. Επίσης, καλή θα ήταν η δημιουργία ενός API, που θα

ΚΕΦΆΛΑΙΟ 9. ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΆΣΕΙΣ

παρέχει τη δυνατότητα το πολυτροπικό σύστημα που δημιουργήθηκε να δέχεται αιτήματα από άλλες εφαρμογές/ιστοσελίδες.

Βιβλιογραφία

- [1] Rachael E Jack, Roberto Caldara, and Philippe G Schyns. “*Internal representations reveal cultural diversity in expectations of facial expressions of emotion.*“. *Journal of Experimental Psychology: General*, 141(1):19, 2012.
- [2] Peter Norvig Stuart J. Russell. “*Artificial Intelligence: A Modern Approach*“. Pearson Education, 2009.
- [3] Min Chen, Francisco Herrera, and Kai Hwang. “*Cognitive computing: architecture, technologies and intelligent applications*“. *Ieee Access*, 6:19774–19783, 2018.
- [4] Simon JD Prince. “*Computer vision: models, learning, and inference*“. Cambridge University Press, 2012.
- [5] Batta Mahesh. “*Machine learning algorithms-a review*“. *International Journal of Science and Research (IJSR). [Internet]*, 9(1):381–386, 2020.
- [6] Jeannette Lawrence. “*Introduction to neural networks*“. California Scientific Software, 1993.
- [7] K. R. Chowdhary. “*Natural Language Processing*“, pages 603–649. Springer India, New Delhi, 2020. ISBN 978-81-322-3972-7.
- [8] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. “*More than a Feeling: Accuracy and Application of Sentiment Analysis*“. *International Journal of Research in Marketing*, 40(1):75–87, 2023. ISSN 0167-8116.
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “*Attention is All you Need*“. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, “*Advances in Neural Information Processing Systems*“, volume 30. Curran Associates, Inc., 2017.
- [10] Thang Luong, Hieu Pham, and Christopher D. Manning. “*Effective Approaches to Attention-based Neural Machine Translation*“. In “*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*“, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [11] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. “*Listen, attend and spell: A neural network for large vocabulary conversational speech recognition*“. In “*2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*“, pages 4960–4964, 2016.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [12] Wouter Kool, Herke van Hoof, and Max Welling. “*Attention, Learn to Solve Routing Problems!*”, 2019.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*”, 2019.
- [14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “*Language Models are Few-Shot Learners*”, 2020.
- [15] OpenAI. “*GPT-4 Technical Report*”, 2023.
- [16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. “*LLaMA: Open and Efficient Foundation Language Models*”, 2023.
- [17] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. “*PaLM: Scaling Language Modeling with Pathways*”, 2022.
- [18] Lili Hao and Lizhu Hao. “*Automatic Identification of Stop Words in Chinese Text Classification*”. In “*2008 International Conference on Computer Science and Software Engineering*”, volume 1, pages 718–722, 2008.
- [19] Amal Alajmi, E Mostafa Saad, and RR Darwish. “*Toward an ARABIC stop-words list generation*”. International Journal of Computer Applications, 46(8): 8–13, 2012.
- [20] Hakan Ayral and Sirma Yavuz. “*An automated domain specific stop word generation method for natural language text classification*”. In “*2011 International Symposium on Innovations in Intelligent Systems and Applications*”, pages 500–503, 2011.

- [21] V.J. Hodge and J. Austin. “*A comparison of standard spell checking algorithms and a novel binary neural approach*“. IEEE Transactions on Knowledge and Data Engineering, 15(5):1073–1081, 2003.
- [22] Julian McAuley and Jure Leskovec. “*Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text*“. In “*Proceedings of the 7th ACM Conference on Recommender Systems*“, RecSys ’13, page 165–172, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324090.
- [23] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. “*Learning word vectors for sentiment analysis*“. In “*Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*“, pages 142–150, 2011.
- [24] Nabiha Asghar. “*Yelp Dataset Challenge: Review Rating Prediction*“, 2016.
- [25] Alec Go, Richa Bhayani, and Lei Huang. “*Twitter sentiment classification using distant supervision*“. CS224N project report, Stanford, 1(12):2009, 2009.
- [26] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. “*Recursive deep models for semantic compositionality over a sentiment treebank*“. In “*Proceedings of the 2013 conference on empirical methods in natural language processing*“, pages 1631–1642, 2013.
- [27] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. “*Cross-Lingual Sentiment Quantification*“. IEEE Intelligent Systems, 35(3):106–114, 2020.
- [28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “*RoBERTa: A Robustly Optimized BERT Pretraining Approach*“, 2019.
- [29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. “*ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*“, 2020.
- [30] Li Yang, Ying Li, Jin Wang, and R. Simon Sherratt. “*Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning*“. IEEE Access, 8:23522–23530, 2020.
- [31] C. Hutto and Eric Gilbert. “*VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*“. Proceedings of the International AAAI Conference on Web and Social Media, 8(1):216–225, May 2014.
- [32] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. “*Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining*“. In “*Lrec*“, volume 10, pages 2200–2204, 2010.
- [33] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. “*Sentiment strength detection for the social web*“. Journal of the American Society for Information Science and Technology, 63(1):163–173, 2012.

- [34] M. Abdullah-Al-Wadud, Md. Hasanul Kabir, M. Ali Akber Dewan, and Oksam Chae. “*A Dynamic Histogram Equalization for Image Contrast Enhancement*“. IEEE Transactions on Consumer Electronics, 53(2):593–600, 2007.
- [35] G. Deng and L.W. Cahill. “*An adaptive Gaussian filter for noise reduction and edge detection*“. In “*1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*“, pages 1615–1619 vol.3, 1993.
- [36] Dan Simon. “*Kalman filtering*“. Embedded systems programming, 14(6):72–79, 2001.
- [37] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. “*A Comprehensive Survey of Image Augmentation Techniques for Deep Learning*“. Pattern Recognition, 137:109347, 2023. ISSN 0031-3203.
- [38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “*ImageNet: A large-scale hierarchical image database*“. In “*2009 IEEE Conference on Computer Vision and Pattern Recognition*“, pages 248–255, 2009.
- [39] Li Deng. “*The mnist database of handwritten digit images for machine learning research [best of the web]*“. IEEE signal processing magazine, 29(6):141–142, 2012.
- [40] Karen Simonyan and Andrew Zisserman. “*Very Deep Convolutional Networks for Large-Scale Image Recognition*“, 2015.
- [41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “*ImageNet Classification with Deep Convolutional Neural Networks*“. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, “*Advances in Neural Information Processing Systems*“, volume 25. Curran Associates, Inc., 2012.
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “*Deep Residual Learning for Image Recognition*“. In “*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*“, June 2016.
- [43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. “*Aggregated Residual Transformations for Deep Neural Networks*“, 2017.
- [44] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. “*Densely Connected Convolutional Networks*“. In “*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*“, July 2017.
- [45] Mingxing Tan and Quoc Le. “*EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*“. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, “*Proceedings of the 36th International Conference on Machine Learning*“, volume 97 of “*Proceedings of Machine Learning Research*“, pages 6105–6114. PMLR, 09–15 Jun 2019.
- [46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “*MobileNetV2: Inverted Residuals and Linear Bottlenecks*“. In “*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*“, June 2018.

- [47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*”, 2021.
- [48] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. “*Do Vision Transformers See Like Convolutional Neural Networks?*”. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, “*Advances in Neural Information Processing Systems*”, volume 34, pages 12116–12128. Curran Associates, Inc., 2021.
- [49] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. “*Emerging Properties in Self-Supervised Vision Transformers*”. In “*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*”, pages 9650–9660, October 2021.
- [50] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. “*BEiT: BERT Pre-Training of Image Transformers*”, 2022.
- [51] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. “*Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions*”. *Information Fusion*, 91:424–444, 2023. ISSN 1566-2535.
- [52] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. “*Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia*”. *ACM Trans. Interact. Intell. Syst.*, 6(3), oct 2016. ISSN 2160-6455.
- [53] Guoyong Cai and Binbin Xia. “*Convolutional neural networks for multimedia sentiment analysis*”. In “*Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4*”, pages 159–167. Springer, 2015.
- [54] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. “*YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context*”. *IEEE Intelligent Systems*, 28(3):46–53, 2013.
- [55] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. “*Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification*”. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.
- [56] Shengyi Jiang, Xueming Yan, Haiwei Xue and Ziang Liu. “*Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling*”. *Applied Artificial Intelligence*, 36(1):2000688, 2022.

- [57] Chen Xi, Guanming Lu, and Jingjie Yan. “*Multimodal Sentiment Analysis Based on Multi-Head Attention Mechanism*“. In “*Proceedings of the 4th International Conference on Machine Learning and Soft Computing*“, ICMLSC ’20, page 34–39, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376310.
- [58] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. “*Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph*“. In “*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*“, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [59] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. “*MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*“, 2019.
- [60] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. “*Factify: A multi-modal fact verification dataset*“. In “*Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DEFACTIFY)*“, 2022.
- [61] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. “*Memotion 2: Dataset on sentiment and emotion analysis of memes*“. In “*Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*“, 2022.
- [62] Teng Niu, Shuai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. “*Sentiment Analysis on Multi-View Social Data*“. In “*MultiMedia Modeling*“, page 15–27, 2016.
- [63] Nan Xu and Wenji Mao. “*MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis*“. In “*Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*“, CIKM ’17, page 2399–2402, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185.
- [64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. “*Deep learning*“. MIT press, 2016.
- [65] Zijun Zhang. “*Improved Adam Optimizer for Deep Neural Networks*“. In “*2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*“, pages 1–2, 2018.
- [66] Zhiyuan Li and Sanjeev Arora. “*An Exponential Learning Rate Schedule for Deep Learning*“, 2019.
- [67] Pierre Baldi and Peter J Sadowski. “*Understanding Dropout*“. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, “*Advances*

in *Neural Information Processing Systems*“, volume 26. Curran Associates, Inc., 2013.

- [68] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. “*Image-Text Multimodal Emotion Classification via Multi-View Attentional Network*“. IEEE Transactions on Multimedia, 23:4014–4026, 2021.
- [69] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. “*A Fair and Comprehensive Comparison of Multimodal Tweet Sentiment Analysis Methods*“. In “*Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*“, MMPT ’21, page 37–45, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385305.
- [70] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. “*CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection*“. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, “*Findings of the Association for Computational Linguistics: NAACL 2022*“, pages 2282–2294, Seattle, United States, July 2022. Association for Computational Linguistics.
- [71] Junyu Chen, Jie An, Hanjia Lyu, and Jiebo Luo. “*Improving Visual-textual Sentiment Analysis by Fusing Expert Features*“, 2022.
- [72] Jieyu An and Wan Mohd Nazmee Wan Zainon. “*Integrating color cues to improve multimodal sentiment analysis in social media*“. Engineering Applications of Artificial Intelligence, 126:106874, 2023. ISSN 0952-1976.
- [73] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “*Unsupervised Cross-lingual Representation Learning at Scale*“, 2020.
- [74] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. “*GREEK-BERT: The Greeks Visiting Sesame Street*“. In “*11th Hellenic Conference on Artificial Intelligence*“, SETN 2020, page 110–117, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388788.
- [75] Saif Mohammad. “*A practical guide to sentiment annotation: Challenges and solutions*“. In “*Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*“, pages 174–179, 2016.
- [76] Milagros Miceli, Martin Schuessler, and Tianling Yang. “*Between subjectivity and imposition: Power dynamics in data annotation for computer vision*“. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–25, 2020.
- [77] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. “*Two contrasting data annotation paradigms for subjective NLP tasks*“. arXiv preprint arXiv:2112.07475, 2021.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [78] Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. “*A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts*“. PLoS One, 15(11):e0242050, 2020.
- [79] Miguel Grinberg. “*Flask web development: developing web applications with python*“. ” O'Reilly Media, Inc.”, 2018.
- [80] Jeremy Thomas, Oleksii Potekhin, Mikko Lauhakari, Aslam Shah, and Dave Berning. “*Creating interfaces with Bulma*“. Bleeding Edge Press Santa Rosa, 2018.