Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

How to Go Viral: Leveraging Graph and Semantic Counterfactual Algorithms

Diploma Thesis

by

Ioanna Kioura

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Πληροφορικής
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

How to Go Viral: Leveraging Graph and Semantic Counterfactual Algorithms

Diploma Thesis

by

Ioanna Kioura

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου, 2024.

.....	.....	.....
Γεώργιος Στάμου	Αθανάσιος Βουλόδημος	Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

.....
Ιωαννα Κιουρα
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παρούσα διατριβή εμβαθύνει στον πολύπλοκο τομέα του virality και παραγόντων που το καθορίζουν, εστιάζοντας στα βίντεο του YouTube. Τα εισαγωγικά κεφάλαια θέτουν τις βάσεις για την κατανόηση της προτεινόμενης μεθόδου, αναλύοντας βασικές έννοιες της θεωρίας γραφημάτων, της ανάλυσης συναισθήματος, του image captioning, της σημασιολογικής ομοιότητας και των εξηγήσεων με αντιπαράδειγμα, εργαλεία χρησιμοποιούνται για την κατασκευή ενός ολοκληρωμένου πλαισίου ανάλυσης του viral περιεχομένου. Πιο συγκεκριμένα, τα βασικά στοιχεία της θεωρίας γράφων παρέχουν το δομικό θεμέλιο, τονίζοντας πώς οι γράφοι μπορούν να αναπαραστήσουν πολύπλοκες σχέσεις και εξαρτήσεις εντός του περιεχομένου των βίντεο. Εξετάζονται τεχνικές ανάλυσης συναισθήματος για την κατανόηση του τρόπου με τον οποίο τα δεδομένα κειμένου γίνονται αντιληπτά από το κοινό και των συναισθηματικών αντιδράσεων που προκαλούν. Το κεφάλαιο σχετικά με το image captioning αναλύει την ενσωμάτωση της υπολογιστικής όρασης και της επεξεργασίας φυσικής γλώσσας για την αυτόματη παραγωγή περιγραφικών μεταδεδομένων για εικόνες από thumbnail βίντεο. Χρησιμοποιείται επίσης η σημασιολογική ομοιότητα προκειμένου να είναι δυνατή η σύγκριση κειμενικών δεδομένων και χρησιμοποιείται ένας σημασιολογικός αντιπαραθετικός αλγόριθμος για τον υπολογισμό των διαφορών και της απόστασης μεταξύ δύο γραφημάτων.

Στόχος της παρούσας διπλωματικής εργασίας είναι να εντοπίσει βασικούς παράγοντες που επαναλαμβάνονται σε όλα τα viral βίντεο και να κατασκευάσει ένα framework που θα παρέχει στους δημιουργούς περιεχομένου χρήσιμες συμβουλές για να αυξήσουν τις πιθανότητες τα βίντεό τους να γίνουν viral. Η προτεινόμενη μέθοδος περιλαμβάνει τη δημιουργία ενός προσαρμοσμένου συνόλου δεδομένων με βάση το YouTube Trending Video Dataset, τη μετατροπή των δεδομένων που σχετίζονται με βίντεο σε αναπαραστάσεις γράφων και τη χρήση αλγορίθμων αντιπαραβολής γράφων για τη σύγκρισή τους μεταξύ τους και τον εντοπισμό των βασικών συντελεστών που οδηγούν ένα βίντεο από non-viral σε viral κατάσταση. Διεξάγονται πειράματα, αποκλειστικά σε συγκεκριμένες κατηγορίες βίντεο καθώς και σε ένα μικτό σύνολο δεδομένων. Οι επικρατέστερες διαφορές μεταξύ non-viral και viral βίντεο αναδεικνύονται μέσω στατιστικής ανάλυσης και η ποιοτική ανάλυση προτείνει αλλαγές σε non-viral παραδείγματα βίντεο και διερευνά τα δυνατά και αδύνατα σημεία του framework. Συνολικά, η παρούσα διατριβή παρέχει ένα ισχυρό πλαίσιο για την κατανόηση και την ενίσχυση του virality του περιεχομένου στο YouTube, ενσωματώνοντας θεωρητικές γνώσεις με πρακτικές εφαρμογές για να προσφέρει πολύτιμες στρατηγικές για τους δημιουργούς περιεχομένου, τις επιχειρήσεις, τους influencers κ.α. που στοχεύουν στη μεγιστοποίηση της εμβέλειας και του αντίκτυπού τους.

Λέξεις-κλειδιά — Γράφοι Γνώσης, βίντεο Youtube, Εξηγήσεις με αντιπαράδειγμα, Virality

Abstract

This thesis explores the complex field of content virality and its determinant factors, specifically focusing on YouTube videos. The introductory chapters cover fundamental ideas in the essential concepts of graph theory, sentiment analysis, image captioning, semantic similarity, and counterfactual explanations, which are used to construct a comprehensive framework for understanding viral content. More specifically, graph theory basics provide the structural foundation, highlighting how graphs can represent complex relationships and dependencies present in video material. Sentiment analysis techniques are examined to understand the perception of and emotional response towards textual data. The chapter on image captioning demonstrates the integration of computer vision and natural language processing to automatically generate descriptive metadata for video thumbnails. Semantic similarity is also utilized in order to be able to compare textual data and a semantic counterfactual algorithm is issued to calculate the differences and distance between two graphs.

The objective of this thesis is to identify key factors that are recurrent across viral videos and construct a framework that will provide content creators with useful advice to increase their videos' chances for virality. The proposed method involves creating a customized dataset from YouTube Trending Video Dataset, transforming video-related data into graph representations, and employing graph counterfactual algorithms to compare them with one another and identify key elements that drive a video from non-viral to viral status. Experiments are conducted, in specific video categories and in a mixed dataset as well. The most prevalent differences between non-viral and viral videos are highlighted through statistical analysis and a qualitative analysis suggests changes to non-viral example videos and explores the framework's strengths and weaknesses. Overall, this thesis provides a robust framework for understanding and enhancing viral YouTube content, combining theoretical insights with practical applications to offer valuable advice for content creators, businesses, influencers, etc aiming to maximize their reach and impact.

Keywords — Knowledge Graphs, YouTube videos, Semantic Counterfactuals, Virality

Ευχαριστίες

Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη σε όλους εκείνους που συνέβαλαν στην ολοκλήρωση αυτού του έργου. Πρώτα απ' όλα, ευχαριστώ θερμά τον καθηγητή μου, κ. Γεώργιο Στάμου, για την πολύτιμη καθοδήγηση και υποστήριξή του ως επιβλέπων. Επίσης, ευχαριστώ τους Γεώργιο Φιλανδριανό και Κωνσταντίνο Θωμά, για την καθοδήγησή τους και για το γεγονός ότι μοιράστηκαν μαζί μου την έρευνά τους και τις ιδέες τους.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια και τους φίλους μου για την ψυχική στήριξη και την ενθάρρυνση που μου παρείχαν καθ' όλη τη διάρκεια αυτής της προσπάθειας. Η αγάπη και η υποστήριξή τους ήταν ανεκτίμητες για μένα.

Κιούρα Ιωάννα, Ιούλιος 2024

Contents

List of Figures

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Θεωρητικό υπόβαθρο

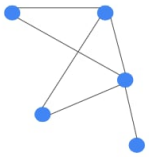
1.1.1 Βασικά στοιχεία θεωρίας γράφων

Ένας γράφος, που συμβολίζεται με G, είναι μια δομή που χρησιμοποιείται για τη μοντελοποίηση σχέσεων ζεύγων αντικειμένων. Ορίζεται ως ένα ζεύγος $G = (V, E)$, όπου:

- V είναι ένα σύνολο κορυφών ή κόμβων.
- E είναι ένα σύνολο ακμών ή συνδέσμων, represented as unordered pairs of vertices.

και αναπαρίσταται ως μη διατεταγμένα ζεύγη κορυφών. Εάν η σειρά των κορυφών σε κάθε ακμή έχει σημασία, ο γράφος ονομάζεται κατευθυνόμενος γράφος ή διγράφος και κάθε ακμή αναπαρίσταται ως ένα διατεταγμένο ζεύγος $e = (u, v)$. Εδώ, u και v είναι η ουρά και η κεφαλή της ακμής, αντίστοιχα. Οι κατευθυνόμενες ακμές μπορούν επίσης να αναφέρονται ως τόξα [99].

Στη βιβλιογραφία, ο όρος "γράφος" αναφέρεται συνήθως σε έναν απλό γράφο, ο οποίος έχει το πολύ μία ακμή μεταξύ δύο κορυφών, χωρίς να έχει αυτο-βρόχους.



(a) Οπτική αναπαράσταση

Vertex	Adjacent to
1	2, 3
2	1, 3, 4
3	1, 2, 4
4	2, 3, 5
5	4

(b) Adjacency List

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

(c) Adjacency Matrix

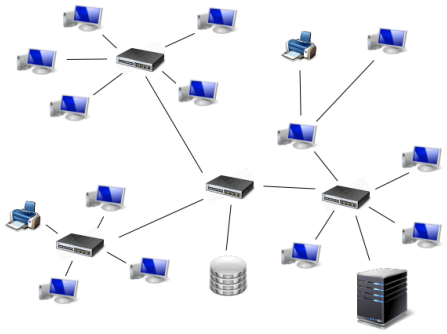
Figure 1.1.1: Representation of undirected graph [9].

Οι γράφοι μπορούν να περιγραφούν με τη χρήση μιας λίστας γειτνίασης ή ενός πίνακα γειτνίασης. Μία λίστα γειτνίασης απαριθμεί κάθε κορυφή ακολουθούμενη από τις γειτονικές της κορυφές. Ένας πίνακας γειτνίασης είναι ένας πίνακας $n \times n$ όπου η εγγραφή στη γραμμή i και στη στήλη j είναι 1 εάν υπάρχει ακμή μεταξύ των κορυφών v_i και v_j, και 0 διαφορετικά. Οι λίστες γειτνίασης είναι πιο αποδοτικές σε χώρο για αραιά γραφήματα, ενώ και οι δύο αναπαραστάσεις είναι παρόμοιες για πυκνά γραφήματα.

Ένα γράφημα του οποίου οι κορυφές είναι ονομασμένες ονομάζεται επισημασμένο γράφημα (labeled graph). Η γειτονιά $N(u)$ μιας κορυφής u είναι το σύνολο των γειτονικών κόμβων. Ο βαθμός μιας κορυφής είναι ο αριθμός των προσπίπτουσων ακμών.

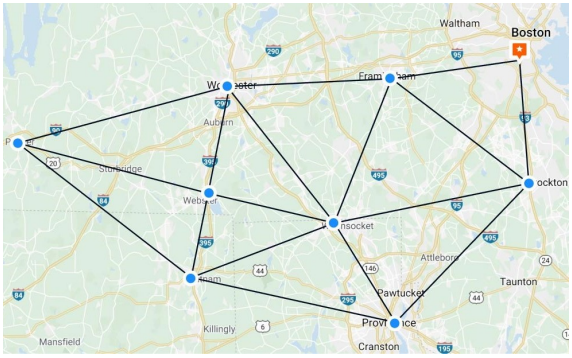
Ένα σταθμισμένο γράφημα έχει ακμές με αριθμητικές τιμές που ονομάζονται βάρη. Αυτά μπορεί να αντιπροσωπεύουν αποστάσεις, κόστη ή άλλα μέτρα. Σε έναν πίνακα γειτνίασης, τα βάρη αντικαθιστούν την τιμή 1 για να δηλώσουν την ύπαρξη μιας ακμής.

Οι γράφοι μοντελοποιούν πολύπλοκες σχέσεις σε τομείς όπως η επιστήμη των υπολογιστών, η μηχανική, η βιολογία και οι κοινωνικές επιστήμες. Χρησιμοποιούνται στην ανάλυση κοινωνικών δικτύων, στη βελτιστοποίηση μεταφορών και στη διαχείριση της δικτυακής κυκλοφορίας.



(a) Γράφος που αναπαριστά δίκτυο υπολογιστών

(b) Γράφος που αναπαριστά κονωνικό δίκτυο



(c) Γράφος που αναπαριστά δίκτυο μεταφοράς

Figure 1.1.2: Παραδείγματα αναπαραστάσεων γράφων

1.1.2 Διμερείς γράφοι και το Minimum Weight Full Matching πρόβλημα

Ένα γράφημα $G = (V, E)$ είναι διμερές αν το σύνολο κορυφών του V μπορεί να διαιρεθεί σε δύο ξένα υποσύνολα, U και W, έτσι ώστε καμία ακμή να μην συνδέει κορυφές εντός του ίδιου υποσυνόλου. Η ιδιότητα αυτή είναι χρήσιμη σε διάφορες εφαρμογές [63].

Τα διμερή γραφήματα χρησιμοποιούνται σε προβλήματα αντιστοίχισης, όπου ο στόχος είναι να αντιστοιχίσουμε κορυφές στο U με κορυφές στο W. Ένα πρόβλημα είναι το minimum weight full matching, το οποίο περιλαμβάνει την εύρεση του βέλτιστου τρόπου αντιστοίχισης στοιχείων δύο συνόλων έτσι ώστε το συνολικό κόστος να ελαχιστοποιείται.

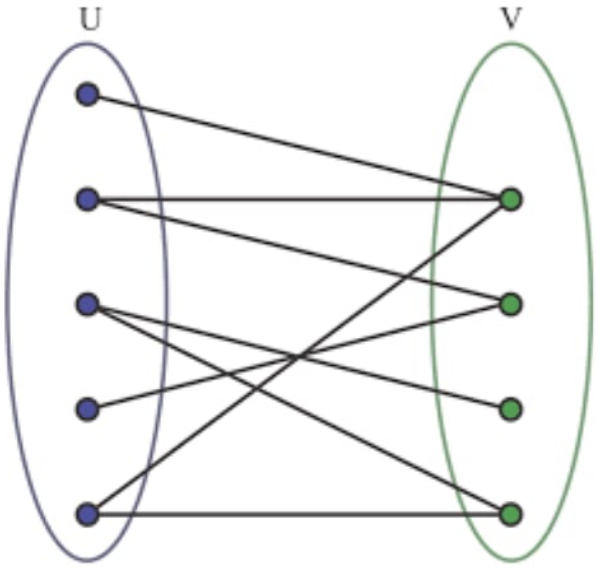


Figure 1.1.3: Ένα παράδειγμα διμερή γράφου [63]

Ορισμός Προβλήματος Minimum Weight Full Matching: Δεδομένων δύο συνόλων U και V με ακμές E, το καθένα με ένα βάρος $w(u, v)$, βρείτε ένα ταίριασμα $M \subseteq E$ που αντιστοιχίζει κάθε στοιχείο του U με ένα στο V, ελαχιστοποιώντας το άθροισμα των βαρών.

Μαθηματική διατύπωση:

$$\min \sum_{(u,v)\in M} w(u, v)$$

υπό την προϋπόθεση ότι:

$$|M| = \min(|U|, |V|)$$

Πολλοί αλγόριθμοι επιλύουν αυτό το πρόβλημα:

- Ο αλγόριθμος Hungarian (ή αλγόριθμος Kuhn-Munkres): Πολυωνυμική χρονική πολυπλοκότητα $O(n^3)$ [52, 65].
- Γραμμικός προγραμματισμός: [18].
- Αλγόριθμος δημοπρασίας: $O(n^3)$ [8].

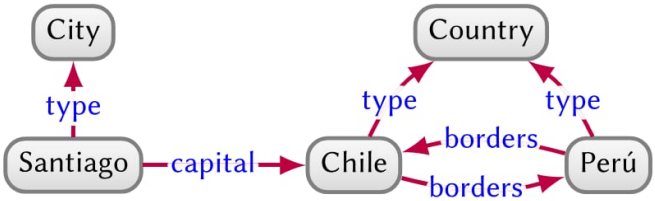
Ο αλγόριθμος του Karp [46] επιτυγχάνει αναμενόμενη χρονική πολυπλοκότητα $O(mn \log n)$ χρησιμοποιώντας ουρές προτεραιότητας, προσφέροντας σημαντικές βελτιώσεις για μεγάλα προβλήματα ανάθεσης.

1.1.3 Γράφοι γνώσης

Γενικές πληροφορίες

Ένας γράφος γνώσης είναι ένας γράφος δεδομένων που έχει σχεδιαστεί για να αναπαριστά τη γνώση σχετικά με τον πραγματικό κόσμο, όπου οι κόμβοι συμβολίζουν οντότητες και οι ακμές απεικονίζουν σχέσεις. Το σύγχρονο ενδιαφέρον για τους γράφους γνώσης ξεκίνησε με τον γράφο γνώσης της Google το 2012 [28], που αναπτύχθηκε από πηγές όπως η DBpedia και η Freebase [2, 11].

Οι γράφοι γνώσης ενσωματώνουν πολύπλοκες σχέσεις δεδομένων, βελτιώνοντας την ακρίβεια της αναζήτησης και των εφαρμογών τεχνητής νοημοσύνης. Κωδικοποιούν σημασιολογικές σχέσεις, παρέχοντας κατανόηση του context που είναι απαραίτητη για εφαρμογές όπως τα chatbots και οι εικονικοί βοηθοί. Οι γράφοι γνώσης είναι επεκτάσιμοι και ευέλικτοι, γεγονός που τους καθιστά ιδανικούς για τον χειρισμό σύνθετων ερωτημάτων σε τομείς όπως η υγειονομική περίθαλψη και τα οικονομικά.



(a) Γράφος DEL

(b) Ετερογενής γράφος

Figure 1.1.4: Δεδομένα για πρωτεύουσες και χώρες σε ένα γράφημα del και ένα ετερογενές γράφημα [41].

Παρά τις δυνατότητές τους, η δημιουργία και η ενσωμάτωση γνώσης από πολλαπλές πηγές σε έναν συνεκτικό γράφο παραμένει πρόκληση. [71].

Resource Description Framework

Χρησιμοποιούμε το Resource Description Framework (μτφρ: πλαίσιο περιγραφής πόρων) ως μοντέλο δεδομένων για τους γράφους γνώσης. Το RDF, τυποποιημένο από το W3C, περιγράφει δεδομένα χρησιμοποιώντας τριπλέτες (υποκείμενο, κατηγόρημα, αντικείμενο), επιτρέποντας πολύπλοκες αναπαραστάσεις σχέσεων [71, 33, 27].

Παράδειγμα: Για την αναπαράσταση "John Smith created a webpage":

- Υποκείμενο: http://www.example.org/index.html
- Κατηγόρημα: http://purl.org/dc/elements/1.1/creator
- Αντικείμενο: http://www.example.org/people/JohnSmith

Το RDF χρησιμοποιεί URIs για τον μοναδικό προσδιορισμό των υποκειμένων και των κατηγορημάτων. Τα αντικείμενα μπορεί να είναι URIs ή κυριολεκτικά-literals (συμβολοσειρές, αριθμοί, ημερομηνίες). [57, 71, 27].

Κενοί κόμβοι: Οι κενοί κόμβοι αντιπροσωπεύουν πόρους χωρίς global αναγνωριστικό, και είναι χρήσιμοι για πολύπλοκες δομές δεδομένων [57, 71, 27].

Γράφος RDF: Μια συλλογή από τριπλέτες RDF σχηματίζει ένα γράφημα RDF, επιτρέποντας την ενσωμάτωση σύνθετων δεδομένων [57, 71, 27].

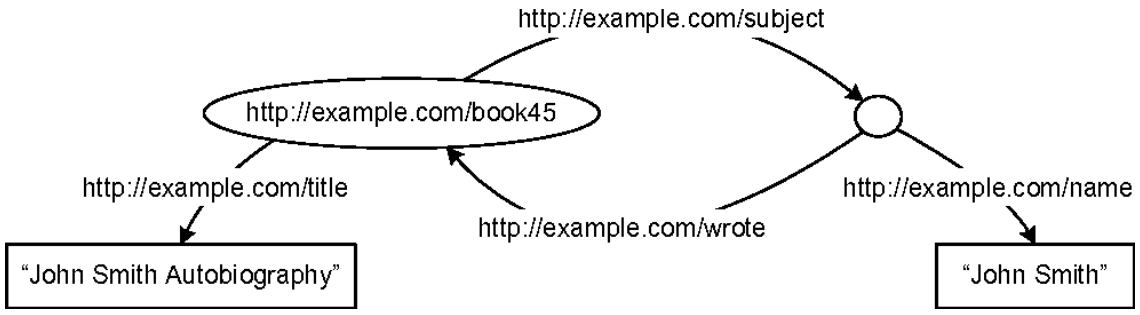


Figure 1.1.5: Παράδειγμα γράφου RDF

1.1.4 Εισαγωγή στην ανάλυση συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis), ή εξόρυξη γνώμης (opinion mining), είναι μια τεχνική NLP για τον εντοπισμό και την εξαγωγή υποκειμενικών πληροφοριών από κείμενο, ταξινομώντας το ως θετικό, αρνητικό ή ουδέτερο. Αναλύει τα συναισθήματα, τις απόψεις και την διάθεση, παρέχοντας πληροφορίες σχετικά με τη δημόσια αντίληψη και τη συναισθηματική αντίδραση απέναντι σε ένα κείμενο [79].

1.1.5 Εφαρμογές της ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος εφαρμόζεται σε διάφορους τομείς [55]:

Επιχειρήσεις: Οι επιχειρήσεις χρησιμοποιούν την ανάλυση συναισθήματος για να αναλύσουν τα σχόλια των πελατών, τις κριτικές των προϊόντων τους και τις τάσεις της αγοράς, να παρακολουθούν τη φήμη της μάρκας τους και να αξιολογούν την ικανοποίηση των πελατών. Εξετάζοντας τα μέσα κοινωνικής δικτύωσης και τις διαδικτυακές κριτικές, μπορούν να αξιολογήσουν το κοινό αίσθημα και να προσαρμόσουν τις στρατηγικές μάρκετινγκ τους.

Πολιτική: Στην πολιτική, η ανάλυση συναισθήματος συμβάλλει στην αντίληψη της κοινής γνώμης για πολιτικές, πολιτικά πρόσωπα και γεγονότα. Βοηθά τους αναλυτές να κατανοήσουν το αίσθημα των ψηφοφόρων, να παρακολουθήσουν τις αλλαγές της κοινής γνώμης και να σχεδιάσουν στρατηγικές. Η ανάλυση των μέσων κοινωνικής δικτύωσης και των ειδησεογραφικών άρθρων μπορεί να αποκαλύψει πληροφορίες για τη συμπεριφορά των ψηφοφόρων και να προβλέψει τα εκλογικά αποτελέσματα.

Υγειονομική περίθαλψη: Η ανάλυση συναισθήματος βελτιώνει τη φροντίδα των ασθενών αναλύοντας τα σχόλια από έρευνες, μέσα κοινωνικής δικτύωσης και φόρουμ. Βοηθά τους παρόχους υγειονομικής περίθαλψης να εντοπίζουν τομείς προς βελτίωση και να κατανοούν τις εμπειρίες των ασθενών. Παρακολουθεί επίσης το κοινό αίσθημα σχετικά με τις πολιτικές και τις θεραπείες υγείας.

1.1.6 Τεχνικές ανάλυσης συναισθήματος

Στην ανάλυση συναισθήματος χρησιμοποιούνται διάφορες προσεγγίσεις [79]:

Μέθοδοι που βασίζονται στην χρήση λεξικού: Οι συγκεκριμένες μέθοδοι βασίζονται σε προκαθορισμένα λεξικά λέξεων αντιστοιχισμένων σε συναισθήματα, αποδίδοντας βαθμολογία συναισθήματος σε λέξεις ή φράσεις για τον προσδιορισμό του συνολικού συναισθήματος. Ωστόσο, μπορεί να δυσκολευτούν με τα συμφραζόμενα και τις γλωσσικές αποχρώσεις, όπως ο σαρκασμός και οι ιδιωματισμοί.

Μέθοδοι που χρησιμοποιούν μηχανική μάθηση: Αυτές οι μέθοδοι περιλαμβάνουν την εκπαίδευση αλγορίθμων σε σύνολα δεδομένων με ετικέτες με σκοπό την πρόβλεψη του συναισθήματος. Καταγράφουν σύνθετα μοτίβα και λαμβάνουν υπόψη το context. Οι συνήθεις προσεγγίσεις περιλαμβάνουν:

Επιβλεπόμενη μάθηση: Αλγόριθμοι όπως οι Naive Bayes, SVM και νευρωνικά δίκτυα εκπαιδεύονται σε σύνολα δεδομένων με ετικέτες.

Μη επιβλεπόμενη μάθηση: Αυτές οι μέθοδοι προσδιορίζουν το συναίσθημα χωρίς επισημασμένα δεδομένα, χρησιμοποιώντας τεχνικές ομαδοποίησης ή μοντελοποίησης θεμάτων.

Υβριδικές μέθοδοι: Συνδυάζουν προσεγγίσεις βασισμένες σε λεξικό και στη μηχανική μάθηση και έτσι βελτιώνουν την ακρίβεια αξιοποιώντας τα πλεονεκτήματα και των δύο.

1.1.7 Προκλήσεις στην ανάλυση συναισθήματος

Παρά τη χρησιμότητά της, η ανάλυση συναισθήματος αντιμετωπίζει αρκετές προκλήσεις [79, 14, 55]:

Εντοπισμός σαρκασμού: Ο σαρκασμός συχνά εκφράζει ένα συναίσθημα αντίθετο στο κυριολεκτικό νόημα, καθιστώντας δύσκολη τη σωστή ερμηνεία του από τους αλγορίθμους.

Κατανόηση του πλαισίου-context: Οι λέξεις μπορεί να έχουν διαφορετική σημασία ανάλογα με το context στο οποίο βρίσκονται. Το γεγονός αυτό επηρεάζει την ερμηνεία του συναισθήματος. Για παράδειγμα, η λέξη "great" μπορεί να είναι θετική στην περίπτωση "great job" αλλά αρνητική στην περίπτωση "great, just what I needed" (σαρκαστικά).

Υποστήριξη πολλαπλών γλωσσών: Η ανάπτυξη μοντέλων που λειτουργούν σε διαφορετικές γλώσσες και διαλέκτους είναι πολύπλοκη λόγω των μοναδικών γλωσσικών χαρακτηριστικών και των πολιτισμικών πλαισίων.

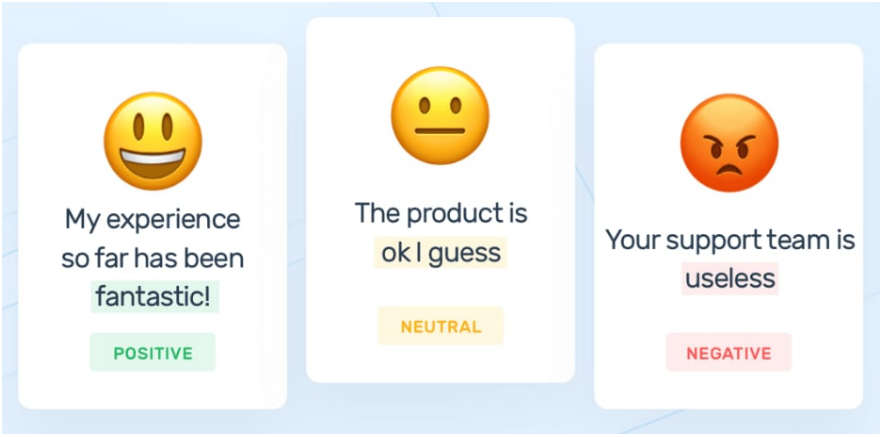


Figure 1.1.6: Παράδειγμα ανάλυσης συναισθήματος [64]

1.1.8 VADER: Valence Aware Dictionary for Sentiment Reasoning

Το VADER είναι ένα εργαλείο ανάλυσης συναισθήματος βασισμένο σε κανόνες και προσαρμοσμένο στα κείμενα των μέσων κοινωνικής δικτύωσης. Αναπτύχθηκε από τους C.J. Hutto και Eric Gilbert [44] και διαχειρίζεται την ανεπίσημη γλώσσα, τις συντομογραφίες και τα emoticons.

Ανάπτυξη λεξικού

Το λεξικό συναισθήματος VADER αναπτύχθηκε εξετάζοντας υπάρχουσες "τράπεζες λέξεων συναισθήματος" όπως LIWC, ANEW και GI [72, 12, 82]. Έπειτα επεκτάθηκε για να συμπεριλάβει χαρακτηριστικά που είναι κοινά στα microblogs, όπως τα emoticons, τα ακρωνύμια και η αργκό, με αποτέλεσμα να προκύψουν πάνω από 7.500 επικυρωμένα λεξιλογικά χαρακτηριστικά που αξιολογήθηκαν από ανθρώπινους βαθμολογητές μέσω του Amazon Mechanical Turk.

Πώς λειτουργεί

Το VADER συνδυάζει ένα λεξικό συναισθήματος με γραμματικούς και συντακτικούς κανόνες για να προσδιορίζει το συναίσθημα ενός κειμένου. Εφαρμόζει ευρετικούς κανόνες για τη στίξη, την κεφαλαιοποίηση, τους τροποποιητές βαθμού, τους συνδέσμους και την άρνηση. Το VADER παράγει τέσσερις βαθμολογίες-scores (pos, neu, neg, compound), που αντιπροσωπεύουν διαφορετικές πτυχές του συναισθήματος.

Scores:

- **Pos (Positive):** Ποσοστό του κειμένου που θεωρείται θετικό.
- **Neu (Neutral):** Ποσοστό του κειμένου που θεωρείται ουδέτερο.
- **Neg (Negative):** Ποσοστό του κειμένου που θεωρείται αρνητικό.
- **Compound:** Κανονικοποιημένη, σταθμισμένη σύνθετη βαθμολογία που κυμαίνεται από -1 (πιο αρνητική) έως +1 (πιο θετική).

Ερμηνεία του Compound Score:

- *Θετικό συνολικό συναίσθημα:* Compound score μεγαλύτερο του 0.05.
- *Ουδέτερο συνολικό συναίσθημα:* Compound score μεταξύ -0.05 και 0.05.
- *Αρνητικό συνολικό συναίσθημα:* Compound score μικρότερο του -0.05.

Απόδοση και Πλεονεκτήματα

Το VADER είναι αποτελεσματικό για την ανάλυση σύντομων, ανεπίσημων κειμένων, όπως tweets και αναρτήσεις στο Facebook, με επιδόσεις εξίσου καλές ή και καλύτερες από τους ανθρώπινους αξιολογητές και ακρίβεια ταξινόμησης F1 ίση με 0,96 [44]. Η προσέγγιση του, που βασίζεται σε κανόνες, είναι υπολογιστικά αποδοτική, καθιστώντας την κατάλληλη για ανάλυση συναισθήματος σε πραγματικό χρόνο.

1.1.9 Image captioning

Image captioning ονομάζεται η παραγωγή κειμένου που περιγράφει ό,τι απεικονίζεται σε μία εικόνα. Αξιοποιεί τεχνικές όρασης υπολογιστών και επεξεργασίας φυσικής γλώσσας. Είναι ζωτικής σημασίας για εφαρμογές όπως η παροχή βοήθειας σε άτομα με προβλήματα όρασης, η ενίσχυση της αναζήτησης εικόνων στο Ίντερνετ, η αυτοματοποίηση του content creation και η βελτίωση της αλληλεπίδρασης ανθρώπου-υπολογιστή [91].

Figure 1.1.7: Παράδειγμα αποτελεσμάτων image captioner ομαδοποιημένα με βάση την ποιότητα [91]

Τεχνικές και μοντέλα

Τα image captioning μοντελα έχουν εξελιχθεί από template-based μεθόδους σε προσεγγίσεις βαθιάς μάθησης που περιλαμβάνουν CNNs και RNNs. Το μοντέλο "Show and Tell" των Vinyals et al. (2015) χρησιμοποιεί ένα CNN για την εξαγωγή χαρακτηριστικών και ένα LSTM για τη δημιουργία λεζάντας. Αυτό το μοντέλο λειτουργεί σε τρία βήματα [91]:

Εξαγωγή χαρακτηριστικών: Ένα προ-εκπαιδευμένο CNN όπως το InceptionV3 εξάγει χαρακτηριστικά εικόνας.

Παραγωγή λεζάντας: Ένα LSTM παράγει τη λεζάντα λέξη προς λέξη με βάση αυτά τα χαρακτηριστικά.

Εκπαίδευση: Το μοντέλο εκπαιδεύεται σε σύνολα δεδομένων όπως το MS COCO [80].

Μετρικές αξιολόγησης

Τα image captioning μοντέλα αξιολογούνται συγκρίνοντας τις παραγόμενες λεζάντες με τις λεζάντες αναφοράς χρησιμοποιώντας μετρικές όπως BLEU, METEOR, ROUGE και CIDEr. Αυτές οι μετρικές μετρούν το n-gram overlap, την ακρίβεια, την ανάκληση και τη σημασιολογική ομοιότητα [91].

Προκλήσεις

Οι βασικές προκλήσεις που αντιμετωπίζει το image captioning περιλαμβάνουν:

Ποικιλομορφία στο οπτικό περιεχόμενο: Χειρισμός μεγάλης ποικιλίας αντικειμένων και σκηνών.

Κατανόηση πλαισίου-context: Κατανόηση των σχέσεων μεταξύ των αντικειμένων που εμφανίζονται και της συνολικής σκηνής.

Παραγωγή κειμένου: Παραγωγή εύγλωττων και γραμματικά ορθών προτάσεων [91].

1.1.10 GIT: A Generative Image-to-text Transformer

GIT (Generative Image-to-text Transformer) ενοποιεί τις εργασίες όρασης-γλώσσας, εστιάζοντας στο image και video captioning και στην απάντηση ερωτήσεων. Το GIT αναπτύχθηκε από τη Microsoft και χρησιμοποιεί έναν μόνο κωδικοποιητή εικόνας και έναν μόνο αποκωδικοποιητή κειμένου, απλοποιώντας την αρχιτεκτονική [94].

Αρχιτεκτονική μοντέλου

Απλοποιημένη δομή:

Κωδικοποιητής εικόνας: Ένας μετασχηματιστής όρασης τύπου Swin προ-εκπαιδευμένος σε ζεύγη εικόνας-κειμένου.

Αποκωδικοποιητής κειμένου: Δίκτυο transformer για την παραγωγή κειμένου από χαρακτηριστικά εικόνας.

Προ-εκπαίδευση

Το GIT προ-εκπαιδεύτηκε σε 0,8 δισεκατομμύρια ζεύγη εικόνας-κειμένου από σύνολα δεδομένων όπως τα COCO, CC3M, SBU Captions, Visual Genome, CC12M και ALT200M.

Στόχοι εκπαίδευσης:

Απώλεια γλωσσικής μοντελοποίησης (LM): Αντιστοίχιση εικόνων εισόδου σε περιγραφές κειμένου.

Αντιθετική προ-εκπαίδευση: Εκμάθηση ισχυρών αναπαραστάσεων εικόνων.

Fine-tuning για συγκεκριμένες εργασίες

Το GIT είναι δεχτεί fine-tuning για εργασίες όπως image captioning και visual question answering (VQA). Για εργασίες που αφορούν βίντεο, γίνεται δειγματοληψία και κωδικοποίηση πολλαπλών καρέ.

Scaling μοντέλου και βελτιστοποίηση επιδόσεων

Scaling Up:

Μεγαλύτερα μοντέλα και δεδομένα: Το GIT κλιμακώνει τα δεδομένα προ-εκπαίδευσης και το μέγεθος του μοντέλου για να βελτιώσει τις επιδόσεις, επιτυγχάνοντας κορυφαία αποτελέσματα σε διάφορα benchmarks.

Αξιολόγηση και Benchmarking

Το GIT ξεπέρασε τις επιδόσεις προηγούμενων μοντέλων σε benchmarks για image captioning, VQA και εργασίες που αφορούν βίντεο, ξεπερνώντας τις ανθρώπινες επιδόσεις στο TextCaps [94].

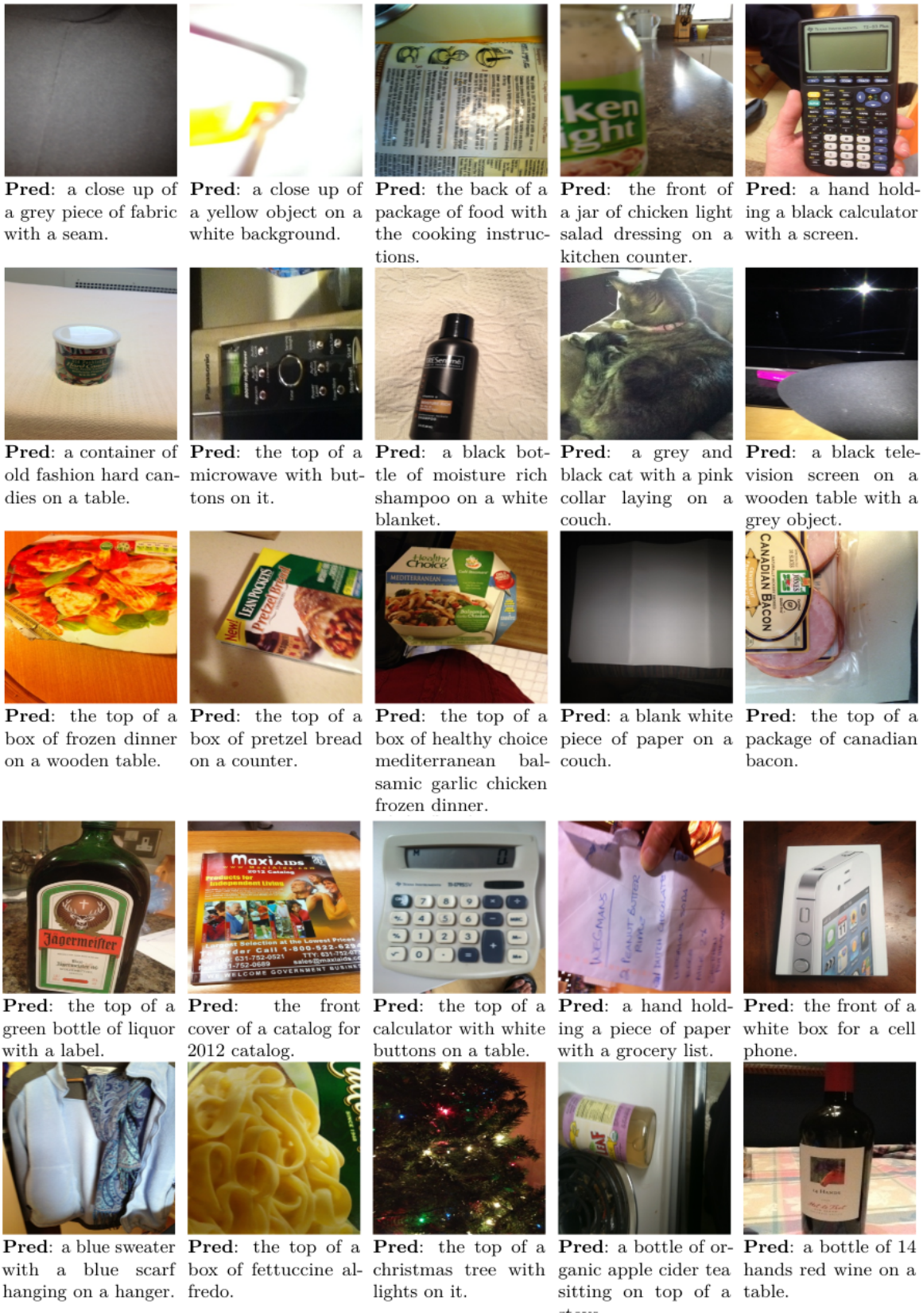


Figure 1.1.8: Αποτελέσματα του μοντέλου GIT model [94]

Βασικές συνεισφορές και καινοτομίες

Ενοποιημένη Αρχιτεκτονική: Απλοποίηση σε έναν κωδικοποιητή εικόνας και έναν αποκωδικοποιητή κειμένου.

Εκπαίδευση από άκρη σε άκρη: Ολόκληρο το δίκτυο εκπαιδεύτηκε σε μεγάλα σύνολα δεδομένων, επιτυγχάνοντας υψηλές επιδόσεις.

Ταξινόμηση με βάση τη γενιά: Παράγει ετικέτες κλάσεων αυτο-παλινδρομικά (auto-regressively) αντί να χρησιμοποιεί προκαθορισμένα λεξιλόγια.

Προκλήσεις και μελλοντική εργασία

Προκλήσεις: Έλεγχος των παραγόμενων λεζάντων και in-context μάθηση χωρίς ενημέρωση παραμέτρων.

Κοινωνικός αντίκτυπος: Βελτίωση της προσβασιμότητας με ταυτόχρονη διαχείριση πιθανής τοξικής γλώσσας σε σύνολα δεδομένων προ-εκπαίδευσης.

1.1.11 Σημασιολογική ομοιότητα

Η σημασιολογική ομοιότητα ή σημασιολογική απόσταση μετρά πόσο στενά σχετίζονται δύο λέξεις ή έννοιες. Είναι θεμελιώδης εργασία στην NLP με εφαρμογές στην ανάκτηση πληροφοριών, στην ταξινόμηση κειμένων και όχι μόνο.

Παραδοσιακές προσεγγίσεις: Οι πρώτες μέθοδοι χρησιμοποιούσαν δομημένες λεξιλογικές βάσεις δεδομένων όπως το WordNet [60], οργανώνοντας τις λέξεις σε σύνολα σε μια ιεραρχική δομή. Μέτρα όπως αυτό του Resnik [76] βασίζονταν στο πληροφοριακό περιεχόμενο των κοινών προγόνων.

Προκλήσεις: Αυτές οι μέθοδοι εξαρτώνται από την πληρότητα και την ακρίβεια της λεξικής βάσης δεδομένων και είναι στατικές, αποτυγχάνοντας να συλλάβουν τις σημασίες που εξαρτώνται από τα συμφραζόμενα.

Εφαρμογές:

- Ανάκτηση πληροφοριών: Βελτίωση των μηχανών αναζήτησης για την ανάκτηση σημασιολογικά συσχετιζόμενων εγγράφων.
- Ταξινόμηση κειμένου: Ομαδοποίηση εγγράφων με βάση το περιεχόμενο.
- Απάντηση ερωτήσεων: Εύρεση σχετικών απαντήσεων κατανοώντας την σημασιολογική ομοιότητα.

1.1.12 Embeddings: Σύλληψη σημασιολογικής απόστασης

Τα embeddings παρέχουν μια πυκνή, συνεχή αναπαράσταση των λέξεων που αποτυπώνει τις σημασιολογικές τους έννοιες με βάση τα συμφραζόμενα.

Τι είναι τα embeddings;

Τα embeddings είναι διανυσματικές αναπαραστάσεις των λέξεων σε έναν συνεχή χώρο, που αποτυπώνουν τις σημασιολογικές σχέσεις μεταξύ τους και επιτρέπουν την εις βάθος ανάλυση τους. Με βάση την distributional hypothesis, οι ενσωματώσεις τοποθετούν λέξεις με παρόμοια συμφραζόμενα κοντά η μία στην άλλη στο διανυσματικό χώρο.

Πώς δημιουργούνται τα embeddings;

Μέθοδοι για την δημιουργία embeddings περιλαμβάνουν:

- Word2Vec: Χρησιμοποιεί νευρωνικά δίκτυα με αρχιτεκτονικές όπως CBOW και Skip-Gram [59].
- GloVe: Αναλύει παγκόσμια στατιστικά στοιχεία συν-εμφάνισης λέξεων για τη δημιουργία ενσωματώσεων [73].
- Contextual Embeddings: Μοντέλα όπως το BERT δημιουργούν ενσωματώσεις που αλλάζουν ανάλογα με τα συμφραζόμενα, χρησιμοποιώντας μια αρχιτεκτονική transformer για αμφίδρομη κατανόηση [22].

Γιατί χρησιμοποιούμε embeddings;

Τα embeddings προσφέρουν:

- Μείωση διαστάσεων: πυκνά διανύσματα καθιστούν τους υπολογισμούς αποδοτικούς.
- Σημασιολογικός Πλούτος: Καταγραφή σύνθετων σχέσεων όπως συνώνυμα και αντώνυμα.
- Επίγνωση των συμφραζόμενων: Προσαρμογή σε διαφορετικές έννοιες λέξεων ανάλογα με τη χρήση.

Εφαρμογές των embeddings

- Ανάκτηση πληροφοριών: Ενισχύουν τις μηχανές αναζήτησης για σημασιολογική ομοιότητα.
- Ταξινόμηση κειμένου: Βελτιώνουν την απόδοση ταξινομητών με πλούσιες σημασιολογικές αναπαραστάσεις.
- Μετάφραση: Διευκολύνουν τις ακριβείς μεταφράσεις με τη σύλληψη των σημασιών που σχετίζονται με τα συμφραζόμενα.

1.1.13 Το πρόβλημα των Vanishing Gradients λόγω των ζωνών κορεσμού της συνάρτησης συνημιτόνου

Τα vanishing gradients αποτελούν μια σημαντική πρόκληση στη βαθιά μάθηση, ιδιαίτερα στη βελτιστοποίηση των embeddings κειμένου με τη συνάρτηση συνημιτόνου [40, 54].

Η συνάρτηση συνημιτόνου μετρά την ομοιότητα μεταξύ των embeddings κειμένου, η οποία ορίζεται ως εξής:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \cdot ||\mathbf{B}||}$$

όπου $\mathbf{A}$ και $\mathbf{B}$ είναι διανύσματα embedding. Οι ζώνες κορεσμού εμφανίζονται όταν $\cos(\theta) \approx \pm 1$, οδηγώντας σε μικρές κλίσεις και αργές ενημερώσεις παραμέτρων.

1.1.14 Angle-Optimized Embeddings Κειμένου

Τα υψηλής ποιότητας embeddings κειμένου βελτιώνουν τις εργασίες σημασιολογικής ομοιότητας κειμένου (STS). Το AnglE, που προτάθηκε από τους Li και Li [54], αντιμετωπίζει τα vanishing gradients βελτιστοποιώντας τις γωνίες στον μιγαδικό χώρο. Πιο συγκεκριμένα, το AnglE εισάγει τη βελτιστοποίηση γωνιών για να βελτιώσει τη ροή και τη βελτιστοποίηση των gradients, μετριάζοντας τις ζώνες κορεσμού της συνάρτησης συνημιτόνου.

Μεθοδολογία

Στρώμα εισόδου: Χρησιμοποιείται padding για τις προτάσεις, οι οποίες έπειτα αντιστοιχίζονται σε έναν d-διάστατο χώρο και περνούν μέσα από έναν κωδικοποιητή (BERT, RoBERTa, LLaMA) για τις αναπαραστάσεις συμφραζομένων.

Cosine Objective: Μετρά τη σημασιολογική ομοιότητα ανά ζεύγη, με στόχο τη μεγιστοποίηση της ομοιότητας για ζεύγη υψηλής ομοιότητας και την ελαχιστοποίησή της για ζεύγη χαμηλής ομοιότητας.

In-Batch Negative Objective: Βελτιώνει την απόδοση με τον εντοπισμό θετικών δειγμάτων εντός ενός batch, μειώνοντας το θόρυβο από λανθασμένα επισημασμένα αρνητικά.

Angle Objective: Βελτιστοποιεί τις διαφορές γωνίας στον μιγαδικό χώρο, βελτιώνοντας τη ροή των gradient και τη βελτιστοποίηση.

Αξιολόγηση

Εκτεταμένα πειράματα δείχνουν ότι το AnglE υπερτερεί των σύγχρονων μοντέλων σε εργασίες STS, συμπεριλαμβανομένων των εργασιών σύντομου κειμένου (short-text STS), μεγάλου κειμένου (long-text STS) και domain-specific εργασιών.

Απόδοση σε STS Tasks

Short-Text STS: Το AnglE-BERT πέτυχε μέση συσχέτιση Spearman 73,55%, ξεπερνώντας το 68,03% του SBERT.

Long-Text STS: Το AnglE-RAN είχε καλύτερες επιδόσεις σε σύνολα δεδομένων μεγάλου μήκους κειμένου από το GitHub Issues, αναδεικνύοντας την αποτελεσματικότητά του στον χειρισμό σύνθετων κειμένων.

Domain-Specific και LLM-Supervised Learning Η AnglE αποδίδει καλά σε domain-specific σενάρια με περιορισμένα επισημασμένα δεδομένα, και βελτιώνεται περαιτέρω με μάθηση με επίβλεψη από LLM.

Ablation Study Βρέθηκε ότι η βελτιστοποίηση της γωνίας είναι κρίσιμη, καθώς χωρίς αυτήν οι επιδόσεις μειώνονται σημαντικά.

Transfer και Non-Transfer Tasks Το AnglE ξεπερνά μοντέλα όπως το SimCSE τόσο σε transfer όσο και σε non-transfer tasks, αποδεικνύοντας ισχυρή γενίκευση.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICR-R	Avg.
Unsupervised Models								
GloVe (avg.) †	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT-flow ‡	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening ‡	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT ‡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT ‡	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT-BERT	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
DiffCSE-BERT	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
SimCSE-BERT	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
LLaMA2-7B ⋆	50.66	73.32	62.76	67.00	70.98	63.28	67.40	65.06
Supervised Models								
InferSent-GloVe †	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE †	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
ConSERT-BERT	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
CoSENT-BERT ⋆	71.35	77.52	75.05	79.68	76.05	78.99	71.19	75.69
SBERT †	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE-BERT	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
SimCSE-LLaMA2-7B ⋆	78.39	89.95	84.80	88.50	86.04	87.86	81.11	85.24
AnglE-BERT	75.09	85.56	80.66	86.44	82.47	85.16	81.23	82.37
AnglE-LLaMA2-7B	**79.00**	**90.56**	**85.79**	**89.43**	**87.00**	**88.97**	**80.94**	**85.96**

Figure 1.1.9: Evaluation results for transfer STS tasks [54]

Model	MRPC	STS-B	QQP	QNLI	GitHub Issues.	Avg.
	test	test	validation	validation	test	
SimCSE-BERT	48.13	76.27	65.84	33.00	60.38	56.72
SBERT	46.19	84.67	73.80	65.98	69.50	68.03
AnglE-RAN	58.70	80.23	74.87	63.04	**71.25**	69.62
AnglE-BERT	**62.20**	**86.26**	**76.54**	**72.19**	70.55	**73.55**

Figure 1.1.10: Evaluation results for non-transfer STS tasks [54]

1.1.15 Ερμηνευσιμότητα (interpretability) και εξηγησιμότητα (explainability) στην Τεχνητή Νοημοσύνη

Η ερμηνευσιμότητα της ΤΝ και η επεξηγησιμότητα χρησιμοποιούνται συχνά εναλλάξιμα ως ισοδύναμες έννοιες, αλλά έχουν διαφορετικές έννοιες.

Ερμηνευσιμότητα

Η ερμηνευσιμότητα της τεχνητής νοημοσύνης περιλαμβάνει την κατανόηση της συμπεριφοράς των μοντέλων μηχανικής μάθησης, δίνοντας έμφαση στα μοντέλα που είναι από τη φύση τους απλά και ερμηνεύσιμα, όπως η γραμμική παλινδρόμηση, τα δέντρα αποφάσεων ή τα μοντέλα που βασίζονται σε κανόνες. Είναι ζωτικής σημασίας για σενάρια όπου οι χρήστες πρέπει να εμπιστεύονται και να κατανοούν το μοντέλο σε βάθος [34]. Η πρόκληση έγκειται στην εξισορρόπηση της πολυπλοκότητας του μοντέλου και της ερμηνευσιμότητας [48].

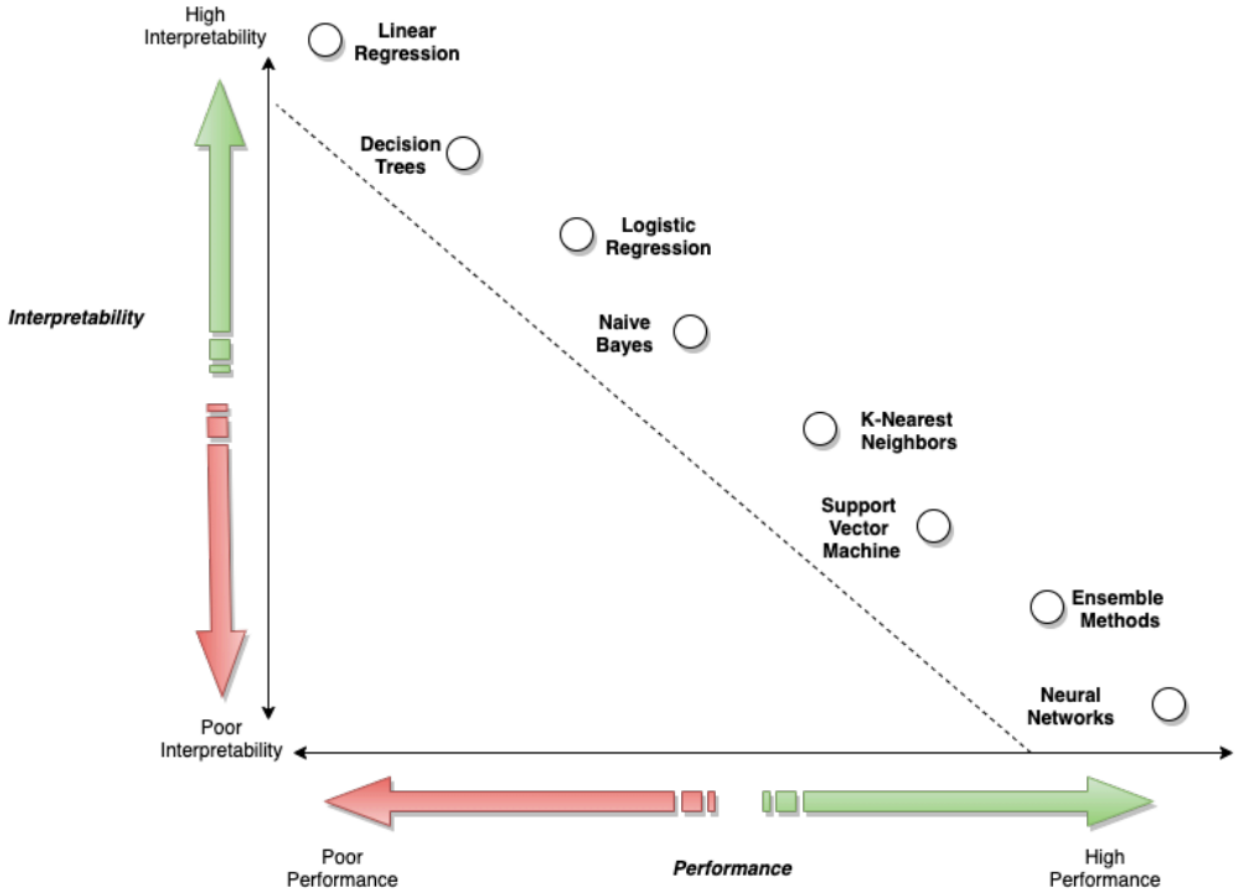


Figure 1.1.11: Interpretability versus performance trade-off given common ML algorithms [48]

Εξηγησιμότητα

Η επεξηγηματικότητα στην ΤΝ επικεντρώνεται στο να καταστήσει τις διαδικασίες λήψης αποφάσεων των μοντέλων ΤΝ διαφανείς, παρέχοντας σαφείς εξηγήσεις για τον τρόπο με τον οποίο προκύπτουν συγκεκριμένα αποτελέσματα. Είναι ζωτικής σημασίας για τη διασφάλιση της λογοδοσίας, της εμπιστοσύνης και της κανονιστικής συμμόρφωσης, ιδίως με πολύπλοκα μοντέλα όπως τα βαθιά νευρωνικά δίκτυα. Χρησιμοποιούνται post-hoc μέθοδοι, όπως οι βαθμολογίες σημασίας των χαρακτηριστικών και οι αντιπαραδειγματικές εξηγήσεις, για να διευκρινιστεί η συμπεριφορά του μοντέλου [48].

1.1.16 Εξηγήσεις με αντιπαράδειγμα

Οι εξηγήσεις με αντιπαράδειγμα παρέχουν πληροφορίες παρουσιάζοντας υποθετικά σενάρια που δείχνουν πώς μικρές αλλαγές στα δεδομένα εισόδου θα μπορούσαν να οδηγήσουν σε διαφορετικά αποτελέσματα, ενισχύοντας τη διαφάνεια και την εμπιστοσύνη στα συστήματα τεχνητής νοημοσύνης [62]. Για παράδειγμα, δείχνοντας ότι ένα δάνειο θα εγκρινόταν αν το εισόδημα του αιτούντος ήταν ελαφρώς υψηλότερο καθιστά τη διαδικασία λήψης αποφάσεων διαισθητική.

Προκλήσεις και εξελίξεις

Οι προκλήσεις περιλαμβάνουν τη διασφάλιση ότι τα αντιπαραδείγματα είναι ρεαλιστικά, εφαρμόσιμα και υπολογιστικά εφικτά. Οι πρόσφατες εξελίξεις χρησιμοποιούν τεχνικές όπως τα GANs για τη δημιουργία ρεαλιστικών εικόνων αντιπαραδειγμάτων και την ανάπτυξη μεθόδων που εφαρμόζονται σε διάφορες εργασίες μηχανικής μάθησης [90].

Εφαρμογές στον πραγματικό κόσμο

Οι εξηγήσεις με αντιπαράδειγμα χρησιμοποιούνται στα χρηματοοικονομικά για την εξήγηση των πιστωτικών αποφάσεων, στην υγειονομική περίθαλψη για την κατανόηση των διαγνωστικών μοντέλων και γίνονται πρότυπο για post-hoc εξηγήσεις μοντέλων [36].

1.1.17 Conceptual Edits ως εξηγήσεις με αντιπαράδειγμα

Οι Filandrianos et al. [30] προτείνουν τη δημιουργία εξηγήσεων με αντιπαράδειγμα με τη χρήση conceptual edits. Τα concepts (έννοιες) αντιπροσωπεύουν γενικευμένες μορφές αντικειμένων στα δεδομένα εισόδου, που συνδέονται με εξωτερική γνώση δομημένη ως ιεραρχίες concepts.

Το framework προσδιορίζει τις ελάχιστες επεξεργασίες εννοιών (ελάχιστα concept edits) για να αλλάξει η πρόβλεψη ενός ταξινομητή μαύρου κουτιού. Πολλαπλές εξηγήσεις με αντιπαράδειγμ μπορούν να εκτιμήσουν μια "συνολική" εξήγηση για μια περιοχή συνόλου δεδομένων και μια τάξη-στόχο.

Explanation Dataset: Το σύνολο δεδομένων αποτελείται από πλειάδες (x_i, C_i), όπου x_i είναι ένα δείγμα και C_i είναι ένα σύνολο εννοιών που περιγράφουν το δείγμα.

Conceptual Distance: Η εννοιολογική απόσταση (d_T) είναι η συντομότερη διαδρομή μεταξύ δύο εννοιών σε έναν γράφο TBox.

Concept Set Edit Distance: Η Concept set edit distance (απόσταση επεξεργασίας συνόλου εννοιών) μετρά το ελάχιστο κόστος μετατροπής ενός συνόλου εννοιών σε ένα άλλο, το οποίο είναι ζωτικής σημασίας για τη δημιουργία ουσιαστικών αντιπαραδειγμάτων.

Σημασία μιας μετατροπής: Η σημασία της μετατροπής ενός δείγματος σε άλλο ποσοτικοποιείται ως εξής:

$$\sigma(a, b) = \frac{|F(x_a) - F(x_b)|}{D_T(C_a, C_b)}$$

όπου F είναι ο ταξινομητής και D_T είναι η concept set edit distance.

Κατασκευή γράφου: Κατασκευάζεται ένας κατευθυνόμενος γράφος όπου οι κόμβοι αντιπροσωπεύουν δείγματα και οι ακμές αντιπροσωπεύουν μετασχηματισμούς, σταθμισμένους με το αντίστροφο της σημαντικότητάς τους.

Τοπικές και γενικευμένες εξηγήσεις με αντιπαράδειγμα

Τοπικές εξηγήσεις με αντιπαράδειγμα: Συγκεκριμένες αλλαγές που απαιτούνται για ένα μόνο δείγμα ώστε να επιτευχθεί η επιθυμητή ταξινόμηση του.

Γενικευμένες εξηγήσεις με αντιπαράδειγμα: Συνάθροιση των τοπικών εξηγήσεων για να συλλεχθούν πληροφορίες σχετικά με τις κοινές αλλαγές που απαιτούνται σε ένα υποσύνολο δεδομένων.

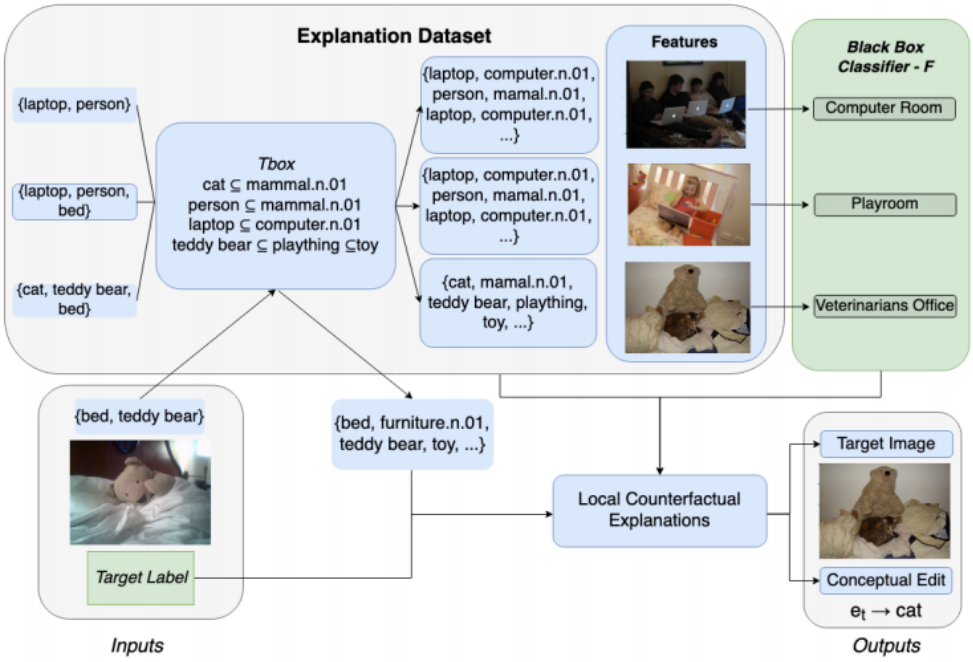


Figure 1.1.12: Conceptual Edits as Counterfactual Explanations framework [30].

Παραγωγή εξηγήσεων με αντιπαράδειγμα

Οι εξηγήσεις με αντιπαράδειγμα παράγονται ως εξής:

Υπολογισμός αποστάσεων εννοιών: Μέτρηση των αποστάσεων μεταξύ εννοιών χρησιμοποιώντας τη συντομότερη διαδρομή στο TBox.

Concept Set Edit Distance: Υπολογισμός του ελάχιστου κόστους που απαιτείται για τη μετατροπή ενός συνόλου εννοιών σε ένα άλλο.

Κατασκευή γράφου: Κατασκευή ενός γραφήματος επεξήγησης με κόμβους που αντιπροσωπεύουν στοιχεία συνόλου δεδομένων και ακμές που αντιπροσωπεύουν μετασχηματισμούς.

Υπολογισμός τοπικών εξηγήσεων με αντιπαράδειγμα: Εύρεση της συντομότερης διαδρομής στο γράφημα εξήγησης για να αλλάξει η ταξινόμηση ενός δείγματος.

Υπολογισμός γενικευμένων εξηγήσεων με αντιπαράδειγμα: Συνάθροιση πολλαπλών τοπικών εξηγήσεων για απόκτηση ευρύτερων γνώσεων.

Λεπτομερής ανάλυση αποτελεσμάτων

Πειράματα με χρήση του CLEVR-Hans3: Η μέθοδος εντόπισε προκαταλήψεις σε έναν ταξινομητή που εκπαιδεύτηκε στο σύνολο δεδομένων CLEVR-Hans3, εντοπίζοντας έτσι με επιτυχία προκαταλήψεις στα δεδομένα εκπαίδευσης.

Πειράματα με χρήση του COCO: Η μέθοδος παρήγαγε εξηγήσεις για μεταβάσεις μεταξύ κλάσεων στο σύνολο δεδομένων COCO. Επισημάνθηκαν οι βασικές έννοιες και οι πιθανές προκαταλήψεις των ταξινομητών.

1.1.18 Ομοιότητα γράφων

Η ομοιότητα γράφων μετρά πόσο όμοια είναι δύο γραφήματα όσον αφορά τη δομή και τις ιδιότητές τους. Είναι ζωτικής σημασίας σε τομείς όπως η βιοπληροφορική, η ανάλυση κοινωνικών δικτύων και η χημειοπληροφορική, όπου είναι συνήθης η σύγκριση πολύπλοκων δικτύων.

Βασικές έννοιες και μέθοδοι

Ισομορφισμός γράφων Δύο γραφήματα G_1 και G_2 είναι ισομορφικά εάν υπάρχει μια αμφιμονοσήμαντη αντιστοίχηση μεταξύ των συνόλων κορυφών τους που διατηρεί τη γειτνίαση. Ο έλεγχος του ισομορφισμού είναι μια υπολογιστική πρόκληση, ειδικά για μεγάλα ή γραφήματα με διαφορές, έστω και μικρές [3].

Απόσταση επεξεργασίας γραφημάτων (Graph Edit Distance-GED) Το GED μετρά τον ελάχιστο αριθμό τροποποιήσεων-edits (εισαγωγές, διαγραφές, αντικαταστάσεις) που απαιτούνται για τη μετατροπή ενός γραφήματος σε ένα άλλο, επιτρέποντας διαφορετικούς βαθμούς ανομοιότητας [32]. Χρησιμοποιείται στην αναγνώριση προτύπων και στην όραση υπολογιστών.

Ισομορφισμός υπογράφων Προσδιορίζει αν ένας γράφος περιέχεται σε έναν άλλο ως ακριβής αντιστοιχία, και είναι NP-πλήρης εργασία [88]. Συνήθης στη χημειοπληροφορική για τον εντοπισμό κοινών υποδομών σε μόρια.

Φασματικές μέθοδοι Σύγκριση των φασμάτων (ιδιοτιμών) πινάκων που σχετίζονται με γραφήματα, όπως ο πίνακας γειτνίασης ή ο πίνακας Laplacian. Παρόμοια φάσματα υποδηλώνουν παρόμοιες δομικές ιδιότητες [16]. Αυτές οι μέθοδοι χειρίζονται αποτελεσματικά μεγάλους γράφους.

Πυρήνες γράφων Χαρτογράφηση γραφημάτων σε ένα χώρο υψηλών διαστάσεων για τον υπολογισμό της ομοιότητας με χρήση συναρτήσεων πυρήνα, όπως ο πυρήνας Weisfeiler-Lehman [92]. Χρησιμοποιείται ευρέως στη μηχανική μάθηση για δεδομένα που βασίζονται σε γράφους.

Μέθοδοι που χρησιμοποιούν embeddings Αναπαράσταση γράφων ως διανύσματα με τη χρήση τεχνικών όπως το node2vec, το DeepWalk και τα συνελικτικά δίκτυα γράφων [38, 49]. Επιτρέπει τον αποδοτικό υπολογισμό της ομοιότητας και κλιμακωσιμότητα.

Εφαρμογές και σημασία

Η ομοιότητα γραφημάτων είναι καίριας σημασίας σε:

- **Βιοπληροφορική**: Σύγκριση δικτύων αλληλεπίδρασης πρωτεϊνών ή γενετικών ρυθμιστικών δικτύων [78].
- **Ανάλυση κοινωνικών δικτύων**: Κατανόηση της δομής και της εξέλιξης μιας κοινότητας [5].
- **Χημειοπληροφορική (Cheminformatics)**: Προσδιορισμός παρόμοιων χημικών ενώσεων με σύγκριση μοριακών γραφημάτων [75].
- **Ανάκτηση πληροφοριών**: Βελτίωση των μηχανών αναζήτησης μέσω της σύγκρισης δομών εγγράφων ως γραφημάτων [7].

1.1.19 Graph Edit Distance

Όπως αναφέρθηκε προηγουμένως, το GED ποσοτικοποιεί την ανομοιότητα μεταξύ γράφων υπολογίζοντας τον ελάχιστο αριθμό πράξεων επεξεργασίας που απαιτούνται για τη μετατροπή ενός γραφήματος σε ένα άλλο.

Πράξεις επεξεργασίας

- *Εισαγωγή κορυφής*: Προσθήκη μίας νέας κορυφής στο γράφημα.
- *Διαγραφή κορυφής*: Διαγραφή μίας υπάρχουσας κορυφής από το γράφημα.
- *Αντικατάσταση κορυφής*: Αντικατάσταση μίας κορυφής από μία άλλη.
- *Εισαγωγή ακμής*: Προσθήκη μίας νέα ακμής στο γράφημα.
- *Διαγραφή ακμής*: Διαγραφή μίας υπάρχουσας ακμής από το γράφημα.
- *Αντικατάσταση ακμής*: Αντικατάσταση μιας υπάρχουσας ακμής από μία άλλη.

Κάθε λειτουργία έχει ένα συγκεκριμένο κόστος και το GED ισούται με το άθροισμα αυτών των κόστων [77]. Το GED μεταξύ των γραφημάτων G_1 και G_2 είναι:

$$GED(G_1, G_2) = \min_{(e_1,\ldots,e_k)\in P(G_1,G_2)} \sum_{i=1}^{k} c(e_i)$$

Συνάρτηση κόστους

Η συνάρτηση κόστους c μπορεί να προσαρμοστεί ανάλογα με τις απαιτήσεις της εφαρμογής. Για παράδειγμα, η αντικατάσταση μιας κορυφής μπορεί να είναι λιγότερο δαπανηρή από τη διαγραφή και την εισαγωγή μιας νέας.

Υπολογισμός GED

Ο υπολογισμός του GED περιλαμβάνει την εύρεση της ακολουθίας των πράξεων επεξεργασίας με το ελάχιστο συνολικό κόστος, ένα NP-πλήρες πρόβλημα. Για μεγαλύτερους γράφους χρησιμοποιούνται προσεγγιστικές μέθοδοι, όπως οι αλγόριθμοι Hungarian, Hausdorff και BP-Beam [52, 31, 66].

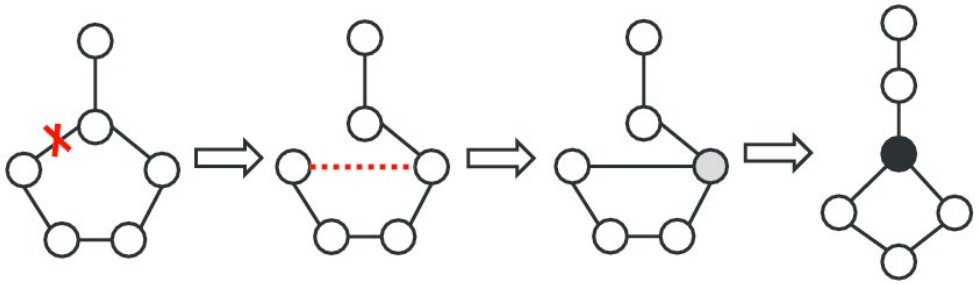


Figure 1.1.13: Graph Edit Distance μεταξύ δύο γράφων. [6]
Το ελάχιστο GED απαιτεί 3 πράξεις επεξεργασίας, οι οποίες εκτιμώνται να έχουν κόστος 3 στην περίπτωση που είναι ισοβαρείς.

Μια προσέγγιση που βασίζεται στη μετατροπή των γραφημάτων σε σύνολα και σε matching διμερών γραφημάτων απλοποιεί τον υπολογισμό του GED [20]. Τα συνδεδεμένα στοιχεία μετατρέπονται σε σύνολα με την αναδίπλωση των ρόλων σε concepts, απλοποιώντας το πρόβλημα στον υπολογισμό της απόστασης επεξεργασίας συνόλων. Αυτό επιλύεται χρησιμοποιώντας το framework για τη παραγωγή εξηγήσεων με αντιπαράδειγμα χρησιμοποιώντας conceptual edits.

1.1.20 Πρόβλεψη "virality" για βίντεο στο YouTube

Ο όρος "virality" αφορά την ταχεία και ευρεία διάδοση περιεχομένου, όπως βίντεο ή άρθρα, σε ψηφιακές πλατφόρμες, ιδίως μέσω των μέσων κοινωνικής δικτύωσης. Πιο συγκεκριμένα, το φαινόμενο αυτό αξιοποιεί τη δύναμη των κοινωνικών δικτύων για να δημιουργήσει σημαντική αλληλεπίδραση και προβολή, συχνά χωρίς την χρήση παραδοσιακής διαφήμισης, καθιστώντας το μια οικονομικά αποδοτική στρατηγική marketing, απόκτησης φήμης και άσκησης επιρροής [85].

Το YouTube, που ιδρύθηκε τον Φεβρουάριο του 2005 από τους Chad Hurley, Steve Chen και Jawed Karim, είναι μια κορυφαία πλατφόρμα διαμοιρασμού βίντεο. Αρχικά προοριζόταν ως ιστοσελίδα γνωριμιών με την χρήση βίντεο, αλλά τελικά εξελίχθηκε σε μια γενική πλατφόρμα κοινοποίησης βίντεο. Η Google εξαγόρασε το YouTube τον Νοέμβριο του 2006 έναντι 1,65 δισεκατομμυρίων δολαρίων σε μετοχές, ενισχύοντας σημαντικά τους πόρους της [101, 17]. Το YouTube Partner Program, που εισήχθη το 2007, επιτρέπει στους δημιουργούς περιεχομένου να εκμεταλλεύονται τα βίντεό τους, μετατρέποντας την πλατφόρμα σε κερδοφόρο χώρο για τους χρήστες. Το YouTube συνέχισε να καινοτομεί με χαρακτηριστικά όπως η ζωντανή ροή και συνδρομητικές υπηρεσίες όπως το YouTube Premium και το YouTube Music [42]. Σήμερα, εξυπηρετεί δισεκατομμύρια χρήστες παγκοσμίως, παραμένοντας κεντρικός κόμβος ψυχαγωγίας, εκπαίδευσης και ανταλλαγής πληροφοριών [42].

Ο όρος "virality" στο YouTube αναφέρεται στην ταχεία και ευρεία κοινοποίηση βίντεο, η οποία χαρακτηρίζεται από σημαντική και ξαφνική αύξηση των προβολών, των κοινοποιήσεων, των likes και των σχολίων μέσα σε

σύντομο χρονικό διάστημα. Ο αλγόριθμος του YouTube προωθεί το "viral" περιεχόμενο ενισχύοντας τα βίντεο που δέχονται εξαρχής μεγάλη προσοχή, δημιουργώντας έναν βρόχο ανατροφοδότησης που ενισχύει περαιτέρω την εμβέλεια τους [98, 1].

Σχετικές έρευνες

Η ανάλυση της δημοτικότητας του περιεχομένου έχει συγκεντρώσει σημαντικό ερευνητικό ενδιαφέρον. Αρχικές μελέτες, όπως η [96], είχαν περιορισμένα μεγέθη δείγματος. Οι Deza et al. [23] εντόπισαν οπτικά χαρακτηριστικά που επηρεάζουν τις πιθανότητες για virality μιας εικόνας. Οι Jiang et al. [45] εξέτασαν τα viral βίντεο και πρότειναν ένα μοντέλο για την πρόβλεψη της ημέρας όπου το βίντεο θα δεχτεί τον μέγιστο αριθμό προβολών. Οι Broxton et al. [13] ανακάλυψαν ότι τα δημοφιλή βίντεο παρουσιάζουν απότομες αιχμές και πτώσεις στον αριθμό προβολών. Οι Vallet et al. [89] ανέλυσαν τους παράγοντες που καθιστούν δημοφιλή τα tweets και τα βίντεο στο YouTube, αναπτύσσοντας ένα πλαίσιο πρόβλεψης χρησιμοποιώντας χαρακτηριστικά από πολλαπλές πλατφόρμες. Οι Pinto et al. [74] έδειξαν ότι οι μελλοντικές προβολές ενός βίντεο στο YouTube μπορούν να προβλεφθούν αξιοποιώντας τα αρχικά μοτίβα προβολών του. Οι Dubey et al. [25] πρότειναν ένα χωρικό transformer μοντέλο για την πρόβλεψη της δημοτικότητας μιας εικόνας χρησιμοποιώντας οπτικές ενδείξεις. Οι Stokowiec et al. [81] προέβλεψαν τη δημοτικότητα διαδικτυακού περιεχομένου χρησιμοποιώντας μόνο τον τίτλο, ενώ οι Chen et al. [15] εισήγαγαν μια πολυτροπική μέθοδο πρόβλεψης για micro-videos. Οι Zhang et al. [103] συνδύασαν οπτικές, κειμενικές και πληροφορίες χρηστών σε ένα attention model για την πρόβλεψη της δημοτικότητας εικόνων του Flickr. Οι Trzcinski et al. [86] πρότειναν ένα support vector regression μοντέλο για την πρόβλεψη της δημοτικότητας των διαδικτυακών βίντεο. Οι Bielski et al. [10] εισήγαγαν ένα πολυτροπικό self-attention μοντέλο για την πρόβλεψη της δημοτικότητας βίντεο. Οι Kong et al. [51] ανέπτυξαν το HIPie, ένα διαδραστικό σύστημα οπτικοποίησης που χρησιμοποιεί τη Hawkes Intensity Process (HIP) για την πρόβλεψη της δημοτικότητας βίντεο στο YouTube.

Πιο πρόσφατες μελέτες σχετικά με την πρόβλεψη της δημοτικότητας βίντεο έχουν διερευνήσει καινοτόμες προσεγγίσεις. Οι Bacic και Gilstrap [4] χρησιμοποίησαν βιομετρικά δεδομένα και μηχανική μάθηση για να προβλέψουν τη δέσμευση των θεατών βίντεο και την δημοτικότητα, επιτυγχάνοντας ακρίβεια άνω του 80%. Μια άλλη μελέτη χρησιμοποίησε μια εφαρμογή PyTorch του ViViT για την πρόβλεψη του virality σε βίντεο του TikTok [39]. Οι Wang κ.ά. [93] εισήγαγαν ένα framework που αξιοποίησε τη δυναμική μεταξύ YouTube και Twitter για την πρόβλεψη της δημοτικότητας βίντεο με υψηλή ακρίβεια χρησιμοποιώντας μόνο μία ημέρα δεδομένων εκπαίδευσης.

1.2 Προτεινόμενο Framework

1.2.1 Συνεισφορά

Οι συνεισφορές αυτής της διπλωματικής εργασίας είναι πολλαπλές και μπορούν να συνοψιστούν ως εξής:

- Η πιο αξιοσημείωτη συνεισφορά αυτής της διατριβής είναι η ανάπτυξη ενός ολοκληρωμένου πλαισίου που ενσωματώνει πολλαπλές προηγμένες αναλυτικές τεχνικές για την κατανόηση και την πρόβλεψη του virality των βίντεο στο YouTube. Το πλαίσιο συνδυάζει τη θεωρία γραφημάτων, την ανάλυση συναισθήματος, το image captioning, embeddings και τις εξηγήσεις με αντιπαράδειγμα, με όλα να διαδραματίζουν καθοριστικό ρόλο στην ολιστική ανάλυση του περιεχομένου των βίντεο.
- Graphs for Relationship Modeling: Η χρήση της θεωρίας γράφων για τη μοντελοποίηση των σχέσεων μεταξύ διαφόρων στοιχείων μεταδεδομένων βίντεο (όπως tags, thumbnails) εφαρμόστηκε εδώ πρώτη φορά. Τα γραφήματα παρέχουν ένα οπτικό και μαθηματικό μέσο για την κατανόηση των πολύπλοκων αλληλεπιδράσεων και εξαρτήσεων που συμβάλλουν στην δημοτικότητα ενός βίντεο.
- Αξιοποιήσιμες γνώσεις: Η συμπερίληψη εξηγήσεων με αντιπαράδειγμα παρέχει αξιοποιήσιμες γνώσεις, δείχνοντας πώς μικρές τροποποιήσεις σε στοιχεία βίντεο (π.χ. τίτλος, tags) μπορούν να επηρεάσουν σημαντικά την τηλεθέαση και τη συμμετοχή. Αυτή η πτυχή ενισχύει την πρακτική χρησιμότητα του framework για τους δημιουργούς περιεχομένου.

1.2.2 Προτεινόμενη μέθοδος

Για την ανακάλυψη των παραγόντων που ωθούν συνηθισμένα βίντεο να γίνονται viral sensations, θα χρειαστούμε ένα προσαρμοσμένο σύνολο δεδομένων από viral βίντεο που θα περιέχει τις εικόνες thumbnails τους και τα

μεταδεδομένα κειμένου τους. Από αυτά τα δεδομένα δημιουργούμε αναπαραστάσεις γράφων, όπου οι κόμβοι και οι ακμές κωδικοποιούν διάφορες ιδιότητες των βίντεο. Χρησιμοποιούνται αλγόριθμοι βασισμένοι σε γράφους και σημασιολογικές αντιπαραθετικές μέθοδοι για τον χειρισμό και την ανάλυση αυτών των δομών γράφων.

YouTube Trending Video Dataset

Η παρούσα διατριβή χρησιμοποιεί το σύνολο δεδομένων YouTube Trending Video Dataset της Kaggle[1], το οποίο ενημερώνεται καθημερινά για να αντανακλά τις τελευταίες τάσεις στο YouTube. Το σύνολο δεδομένων παρέχει πληροφορίες σχετικά με τα trending βίντεο σε διάφορες περιοχές ανά τον κόσμο, με έως και 200 καταχωρημένα trending βίντεο ανά ημέρα. Παρέχει λεπτομερή μεταδεδομένα για κάθε βίντεο και τις μετρήσεις επιδόσεών του, τα οποία συλλέγονται χρησιμοποιώντας το API του YouTube.

Οι στήλες του αποτελούνται από τα εξής χαρακτηριστικά:

- **Video ID**: Μοναδικό αναγνωριστικό για κάθε βίντεο.
- **Title**: Ο τίτλος του βίντεο.
- **Published Date**: Ημερομηνία και ώρα δημοσίευσης του βίντεο.
- **Channel ID**: Μοναδικό αναγνωριστικό του καναλιού δημοσίευσης.
- **Channel Title**: Όνομα του καναλιού δημοσίευσης.
- **Category ID**: Αναγνωριστικό για την κατηγορία του βίντεο.
- **Trending Date**: Ημερομηνία κατά την οποία το βίντεο ήταν "τάση".
- **Tags**: Ετικέτες που σχετίζονται με το βίντεο.
- **View Count**: Αριθμός προβολών.
- **Likes**: Αριθμός των likes.
- **Dislikes**: Αριθμός των dislikes.
- **Comment Count**: Αριθμός σχολίων.
- **Thumbnail Link**: Σύνδεσμος προς την εικόνα thumbnail του βίντεο.
- **Description**: Σύντομη περιγραφή του βίντεο.
- **Comments Disabled**: Δείχνει αν τα σχόλια είναι απενεργοποιημένα.
- **Ratings Disabled**: Δείχνει αν οι αξιολογήσεις είναι απενεργοποιημένες.

Το προσαρμοσμένο σύνολο δεδομένων που δημιουργήσαμε

Ελήφθησαν αποφάσεις σχετικά με τα χαρακτηριστικά και τα μεταδεδομένα που έπρεπε να διατηρηθούν ή να αγνοηθούν, με βάση τον αντίκτυπό τους στον στόχο μας:

- **Video ID**: Διατηρείται για τη μοναδική ταυτοποίηση κάθε βίντεο.
- **Title**: Διατηρείται καθώς είναι από τους βασικούς παράγοντες που καθορίζουν το αν ένας χρήστης θα επιλέξει να δει ένα βίντεο.
- **Published Date**: Σημαντική για την αξιολόγηση του virality με την πάροδο του χρόνου.
- **Channel ID & Title**: Δεν διατηρούνται καθώς δεν είναι χρήσιμα για την ενίσχυση του virality, όταν το κανάλι δεν είναι γνωστό.
- **Category ID**: Διατηρείται για πειράματα που αφορούν συγκεκριμένες κατηγορίες.
- **Trending Date**: Διατηρείται για τη μέτρηση των προβολών σε συγκεκριμένες χρονικές στιγμές.
- **Tags**: Διατηρούνται καθώς επηρεάζουν την αναγνωρισιμότητα και το virality.

[1] *https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset?select=US_youtube_trending_data.csv*

- **View Count**: Κύριος καθοριστικός παράγοντας του virality, συνεπώς διατηρείται για να μπορούμε να συγκρίνουμε βίντεο μεταξύ τους.
- **Likes & Dislikes**: Δεν διατηρούνται, καθώς είναι post-viral μετρήσεις.
- **Comment Count**: Δεν διατηρείται καθώς δεν είναι ορατός πριν από το κλικ και είναι επίσης post-viral μέτρηση.
- **Thumbnail Link**: Διατηρείται καθώς αποτελεί κρίσιμο οπτικό στοιχείο.
- **Comments Disabled & Ratings Disabled**: Διατηρούνται καθώς επηρεάζουν τις δυνατότητες αλληλεπίδρασης του χρήστη με το βίντεο.
- **Description**: Δεν διατηρείται καθώς δεν αποτελεί πρωταρχικό παράγοντα απόφασης για τους θεατές.

Το σύνολο δεδομένων μας έχει πάνω από 240.000 καταχωρίσεις, με περίπου 43.000 μοναδικές καταχωρίσεις. Τα αριθμητικά δεδομένα παραμένουν αναλλοίωτα, ενώ τα κατηγορικά δεδομένα (τίτλος και tags) και τα thumbnails απαιτούν περαιτέρω ανάλυση και επεξεργασία.

Η ανάλυση του τίτλου περιλαμβάνει:

- **Εξαγωγή λέξεων-κλειδιών**: Απομάκρυνση stopwords και σημείων στίξης, μετατροπή σε πεζά γράμματα.
- **Ανάλυση συναισθήματος**: Χρήση του μοντέλου VADER για την εξαγωγή σύνθετης βαθμολογίας (compound score).
- **Σημεία στίξης**: Καταμέτρηση συγκεκριμένων σημείων στίξης.
- **Μήκος**: Μέτρηση του αριθμού των λέξεων.
- **Κεφαλαίο πρώτο γράμμα προτάσεων**: Ελέγξτε αν το πρώτο γράμμα κάθε πρότασης είναι κεφαλαίο.
- **Αναλογία κεφαλαίων λέξεων**: Υπολογίζει την αναλογία των λέξεων γραμμένων με κεφαλαία γράμματα προς τις λέξεις γραμμένες με πεζά.

Τα thumbnails αναλύονται με τη χρήση του μοντέλου GIT για τη δημιουργία λεζάντων. Έπειτα εξάγουμε και διατηρούμε λέξεις-κλειδιά από αυτές τις λεζάντες.

Για τα tags χρησιμοποιούμε την τεχνική Term Frequency - Inverse Document Frequency για να εξάγουμε τα βασικά θέματα που κυριαρχούν στην πλειοψηφία τους. Διατηρούμε επίσης τον αριθμό των διαφορετικών tags που περιέχει το βίντεο.

Μετατροπή σε γράφους γνώσης

Οι μοναδικές σειρές του συνόλου δεδομένων μετατρέπονται σε γράφους γνώσης χρησιμοποιώντας το Resource Description Framework (RDF). Οι γράφοι αναπαριστούν τα βασικά στοιχεία κάθε βίντεο ως ξεχωριστές οντότητες με τις ιδιότητες τους κωδικοποιημένες ως σχέσεις. Δεν συμπεριλάβαμε τα view count, published date και trending date στα γραφήματα, καθώς αυτά θα χρησιμοποιηθούν για τη σύγκριση των βίντεο και δεν υπάρχει λόγος να συγκρίνουμε τις ημερομηνίες που τα βίντεο δημοσιεύτηκαν και έγιναν επίσημα τάσεις ή τον αριθμό προβολών. Αυτά τα χαρακτηριστικά διατηρήθηκαν μόνο για να χωρίσουμε αργότερα το σύνολο δεδομένων σε μικρότερα σύνολα. Ένα παράδειγμα γραφήματος παρουσιάζεται στην εικόνα 1.2.1.

Σύγριση

Για να παράγουμε λογικές προτάσεις για τη βελτιστοποίηση του περιεχομένου ενός δοθέντος γράφου YouTube βίντεο, το συγκρίνουμε με κάθε γράφημα στο σύνολο δεδομένων μας. Επιλέγεται ο μετασχηματισμός με το χαμηλότερο κόστος. Η βιβλιοθήκη cece [20] χρησιμοποιείται για τον υπολογισμό των διαφορών και την παροχή προτάσεων αλλαγών για την ενίσχυση της δημοτικότητας του δοθέντος βίντεο.

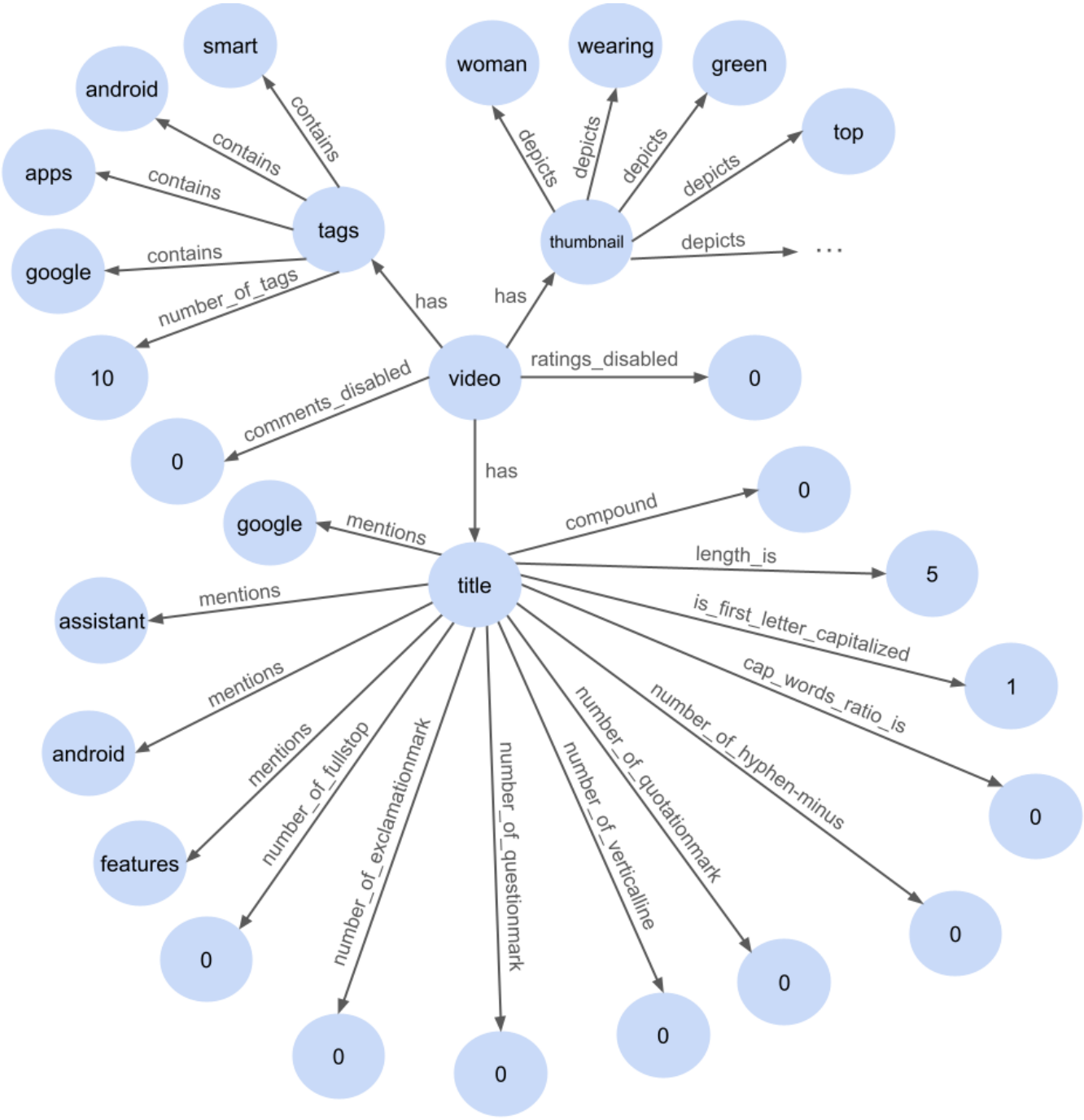


Figure 1.2.1: Παράδειγμα γράφου γνώσης ενός βίντεο του YouTube

Καθώς η σύγκριση ολόκληρων γραφημάτων ήταν υπερβολικά υπολογιστικά κοστοβόρα, χωρίζουμε τους γράφους σε τρία μικρότερα γραφήματα: τίτλος, thumbnail και tags. Το καθένα από αυτά περιέχει μόνο τις λέξεις-κλειδιά που εξήχθηκαν για αυτό, προκειμένου να συγκρίνουμε τους γράφους αποκλειστικά με βάση το θέμα. Αυτοί οι γράφοι παρουσιάζονται στις εικόνες 1.2.2, 1.2.3, 1.2.4. Για να βρούμε το πιο θεματικά όμοιο βίντεο στο σύνολο δεδομένων μας με το δοθέν, θα συγκρίνουμε τίτλους με τίτλους, thumbnails με thumbnails και tags με tags. Για την σύγκριση, οι γράφοι μετατρέπονται σε queries. Ως εκ τούτου, κάθε βίντεο αντιπροσωπεύεται από τρία queries και το κόστος μετασχηματισμού αυτών των queries υπολογίζεται για να βρεθεί το πιο παρόμοιο βίντεο στο σύνολο δεδομένων. Τα τρία queries για το παράδειγμά μας:

τίτλος: ['google', 'assistant', 'features', 'android']

thumbnail: ['woman', 'wearing', 'green', 'striped', 'top', 'showing', 'computer', ...]

tags: ['android', 'smart', 'google', 'apps']

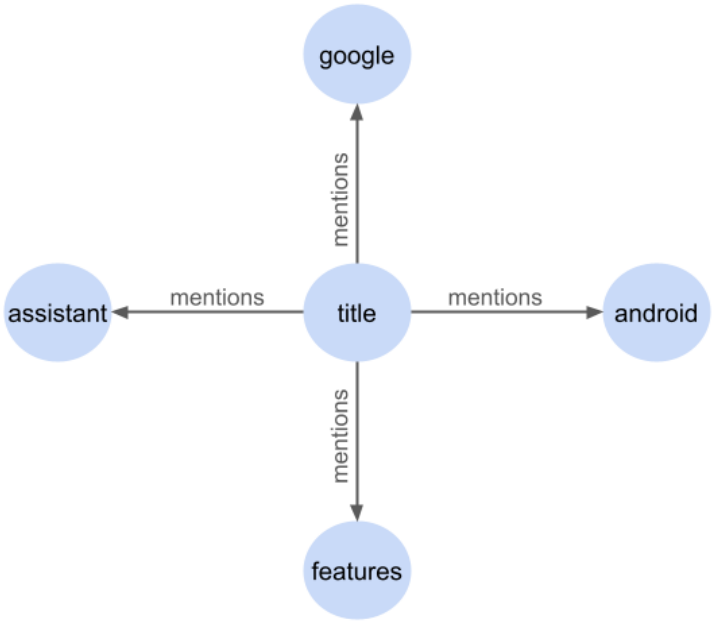


Figure 1.2.2: Παράδειγμα του γράφου του τίτλου

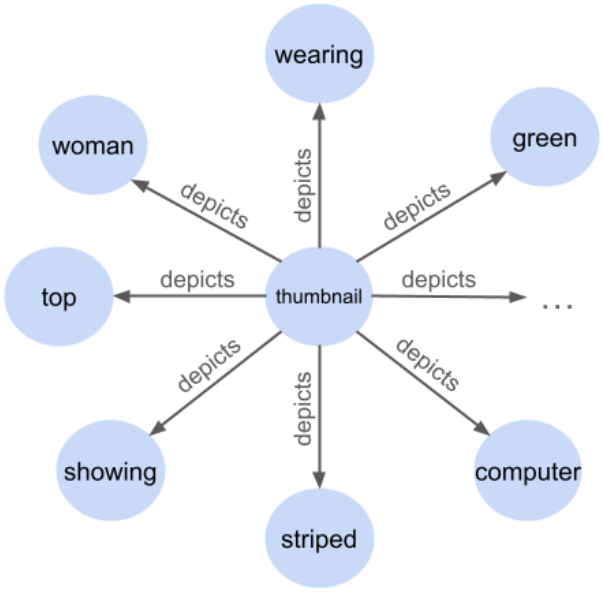


Figure 1.2.3: Παράδειγμα του γράφου του thumbnail

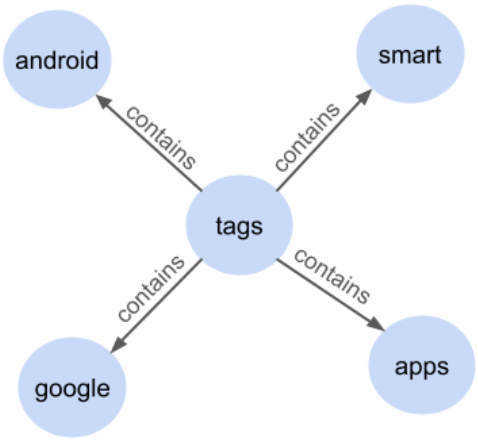


Figure 1.2.4: Παράδειγμα του γράφου των tags

Αφού υπολογίσουμε τα πιο παρόμοια γραφήματα με βάση το θέμα τους, χρησιμοποιούμε ξανά τη βιβλιοθήκη cece μαζί με δικές μας συναρτήσεις για να υπολογίσουμε τις διαφορές μεταξύ των δύο βίντεο και να προτείνουμε αλλαγές για το δοθέν βίντεο, ώστε να αυξήσουμε τις πιθανότητες του να γίνει viral. Οι αλλαγές περιλαμβάνουν αλλαγές στα tags, στο τι δείχνει η εικόνα του thumbnail, στις λέξεις του τίτλου, στη στίξη, στην κεφαλαία γραφή κ.λπ. καθώς κάνουμε χρήση όλων των πληροφοριών που έχουμε εξάγει μέχρι στιγμής.

1.3 Πειράματικό Μέρος

Εδώ θα περιγράψουμε τα πειράματα που πραγματοποιήσαμε χρησιμοποιώντας το framework που κατασκευάσαμε ώστε να προσδιορίσουμε πώς διαφέρουν τα viral βίντεο του YouTube από εκείνα με πολύ μικρότερο αριθμό προβολών. Στόχος μας είναι να εντοπίσουμε κοινά μοτίβα και χαρακτηριστικά που ενισχύουν το virality των βίντεο σε διάφορες κατηγορίες αλλά και συλλογικά.

1.3.1 Γενικά πειράματα

Ξεκινήσαμε τα πειράματά μας με μια γενική προσέγγιση, χρησιμοποιώντας το μικτό σύνολο δεδομένων μας, χωρίς να το χωρίσουμε σε διαφορετικές κατηγορίες. Αυτό έγινε σε μια προσπάθεια να εντοπίσουμε καθολικές τάσεις και χαρακτηριστικά των viral βίντεο. Δεδομένου ότι ο αριθμός των προβολών μεταξύ των δειγμάτων μας έχει αρκετά μεγάλες διαφορές, ήταν δυνατό να χρησιμοποιήσουμε το σύνολο δεδομένων μας και για τη δημιουργία των δεδομένων δοκιμής. Πιο συγκεκριμένα, εξήγαμε όλα τα βίντεο με τουλάχιστον 5.000.000 προβολές μέσα σε μια εβδομάδα για να τα χρησιμοποιήσουμε ως σύνολο viral δεδομένων μας. Πρόκειται για περίπου 4500 βίντεο. Εξήγαμε επίσης τα 4500 βίντεο με το μικρότερο αριθμό προβολών εντός μιας εβδομάδας για να τα χρησιμοποιήσουμε ως δεδομένα δοκιμών. Για το πρώτο σύνολο δεδομένων ο ελάχιστος αριθμός προβολών είναι 5.002.539, ενώ για το δεύτερο ο μέγιστος αριθμός προβολών είναι 325.458. Η διαφορά είναι αρκετά μεγάλη, οπότε είναι λογικό να αξιολογήσουμε τις ανομοιότητές τους.

Τα στατιστικά αποτελέσματα από το πείραμά μας παρουσιάζονται στους πίνακες 11.1, 11.2, 11.3.

Γενικά, πρέπει να τονιστεί ότι υπάρχουν καλές και κακές αντιστοιχίες μεταξύ βίντεο που έγιναν με τη μέθοδό μας. Οι καλές αφορούν βίντεο του συνόλου δοκιμής που σχετίζονται με κοινά, γενικά ζητήματα ή δημοφιλή θέματα. Για αυτά τα βίντεο, είναι ευκολότερο να βρεθεί μια αντιστοιχία στο σύνολο δεδομένων μας, καθώς είναι πιο πιθανό τέτοια θέματα να έχουν συζητηθεί και από viral βίντεο. Για βίντεο, ωστόσο, που έχουν εξαιρετικά συγκεκριμένα ή αφανή θέματα είναι πολύ πιο δύσκολο. Ως εκ τούτου, στις περισσότερες περιπτώσεις, αυτά τα βίντεο αντιστοιχίζονται με ένα δείγμα στο σύνολο δεδομένων με το οποίο δεν μοιράζονται πολλά κοινά χαρακτηριστικά.

Feature	Increased	Decreased	Remains the same
Title length	32.5%	59.9%	7.6%
Title's capital words ratio	30.2%	48.2%	21.6%
Title's compound score	36.2%	34.6%	29.2%
Number of exclamation marks in title	11.9%	20.8%	67.4%
Number of question marks in title	2%	4.7%	93.3%
Number of full stops in title	9.5%	15%	75.5%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	17.5%	21.7%	60.8%
Number of hyphens in title	27.9%	14.9%	57.2%
Number of different tags	37.3%	60.9%	1.8%

Table 1.1: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Feature	True->False	False->True	True->True	False->False
Ratings disabled	1%	0.8%	0%	98.2%
Comments disabled	2.7%	1.4%	0.1%	95.8%
First letter capitalization in title	12.5%	18%	66.3%	3.1%

Table 1.2: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Είναι προφανές από τα παραπάνω στατιστικά στοιχεία ότι, σε σύγκριση με τα βίντεο με λιγότερες προβολές, τα viral βίντεο έχουν μικρότερους τίτλους, λιγότερες λέξεις με κεφαλαία γράμματα και λιγότερα και πιο εστιασμένα tags. Όσον αφορά το αν οι τίτλοι τους έχουν θετική ή αρνητική χροιά, τα αποτελέσματα είναι εξίσου μοιρασμένα. Στις περισσότερες περιπτώσεις, τα σημεία στίξης δεν χρησιμοποιούνται πραγματικά και η συμβουλή είναι να μειωθεί η χρήση τους.

Από τον πίνακα 11.2 μπορούμε να καταλάβουμε ότι τα περισσότερα βίντεο διατηρούν ενεργοποιημένα τόσο τα σχόλια όσο και τις αξιολογήσεις, ανεξάρτητα από τον αριθμό των προβολών τους. Το ίδιο ισχύει και για την κεφαλαιοποίηση του πρώτου γράμματος κάθε πρότασης. Πραγματοποιείται στην πλειονότητα των περιπτώσεων, ειδικά στο σύνολο δεδομένων μας για τα viral βίντεο. Όταν αυτό δεν συμβαίνει, το πιθανότερο είναι ότι ο αλγόριθμός μας θα συμβουλεύσει να εφαρμοστεί.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
man	video
person	game
game	trailer
movie	music
poster	diy

Table 1.3: Στατιστικά αποτελέσματα για thumbnails και tags

Αναλύοντας τις συχνότερες προτεινόμενες αλλαγές στα thumbnails διαπιστώθηκε ότι οι πιο συνηθισμένες

προσθήκες ήταν οι λέξεις "man" και "person", και επίσης η πλειοψηφία των μετασχηματισμών επικεντρώθηκε σε λέξεις όπως αυτές. Αυτό δείχνει ότι η εμφάνιση ανθρώπων στο thumbnail αυξάνει τη δημοτικότητα ενός βίντεο και τις πιθανότητες να προβληθεί. Οι υπόλοιπες πιο δημοφιλείς επεξεργασίες προέρχονταν από τις μεγαλύτερες κατηγορίες βίντεο στα δεδομένα, τις οποίες θα αναλύσουμε στη συνέχεια.

Όσον αφορά τα tags, δεν υπήρχαν αλλαγές αρκετά συχνές και σημαντικές ώστε να αξίζει να αναφερθούν. Οι δημοφιλείς τροποποιήσεις προέρχονταν και πάλι από την αφθονία δειγμάτων ορισμένων κατηγοριών.

1.3.2 Πειράματα βάσει την κατηγορία

Συνολικά, υπάρχουν 15 κατηγορίες βίντεο στο σύνολο δεδομένων μας. Αυτές είναι:

- **Category ID = 1:** Ταινίες & Κινουμένα σχέδια (1703 δείγματα)
- **Category ID = 2:** Αυτοκίνητα & Οχήματα (876 δείγματα)
- **Category ID = 10:** Μουσική (6899 δείγματα)
- **Category ID = 15:** Κατοικίδια & Ζώα (190 δείγματα)
- **Category ID = 17:** Αθλητισμός (5614 δείγματα)
- **Category ID = 19:** Ταξίδια & Εκδηλώσεις (255 δείγματα)
- **Category ID = 20:** Gaming (8787 δείγματα)
- **Category ID = 22:** Άνθρωποι & Blogs (3769 δείγματα)
- **Category ID = 23:** Κωμωδία (2121 δείγματα)
- **Category ID = 24:** Ψυχαγωγία (8539 δείγματα)
- **Category ID = 25:** Ειδήσεις & Πολιτική (1574 δείγματα)
- **Category ID = 26:** Howto & Style (1122 δείγματα)
- **Category ID = 27:** Εκπαίδευση (1032 δείγματα)
- **Category ID = 28:** Επιστήμη & Τεχνολογία (1319 δείγματα)
- **Category ID = 29:** Μη κερδοσκοπικοί οργανισμοί & Ακτιβισμός (19 δείγματα)

Θα πειραματιστούμε μόνο με 5 από αυτές: Μουσική, Αθλητισμός, Gaming, Άνθρωποι και Blogs και Ψυχαγωγία, καθώς τα υπόλοιπα έχουν πολύ λίγα δείγματα για να μπορέσουμε να τα χρησιμοποιήσουμε τόσο για το σύνολο δεδομένων όσο και για τις δοκιμές και να έχουν σημαντική διαφορά στον αριθμό των προβολών, ώστε τα πειράματα να έχουν νόημα.

Μουσική

Αυτή η κατηγορία αποτελείται από μουσικά βίντεο, επίσημα κλιπ μουσικών βίντεο και βίντεο με στίχους. Δεδομένου ότι έχει 6899 δείγματα, δημιουργούμε ένα σύνολο δεδομένων που περιέχει τα 2000 δείγματα με τις περισσότερες προβολές. Το σύνολο δεδομένων δοκιμής μας αποτελείται από τα 2000 με τις λιγότερες προβολές. Τα στατιστικά αποτελέσματα αυτού του πειράματος παρουσιάζονται στους πίνακες.

Feature	Increased	Decreased	Remains the same
Title length	44.4%	46%	9.6%
Title's capital words ratio	45.3%	24.1%	30.5%
Title's compound score	27.6%	30.1%	42.3%
Number of exclamation marks in title	1.2%	2.9%	95.9%
Number of question marks in title	0.9%	1.1%	98%
Number of full stops in title	13.3%	16.9%	69.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	5.5%	4.5%	90%
Number of hyphens in title	19.4%	19.9%	60.7%
Number of different tags	60.4%	35.8%	3.8%

Table 1.4: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Feature	True->False	False->True	True->True	False->False
Ratings disabled	0.6%	0.7%	0%	98.7%
Comments disabled	0.4%	0.3%	0%	99.3%
First letter capitalization in title	8.5%	8.6%	81.5%	1.4%

Table 1.5: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Στους πίνακες 1.4, 1.5 μπορούμε να δούμε ότι στους τίτλους των viral βίντεο μας υπάρχει μεγαλύτερη αναλογία λέξεων με κεφαλαία προς πεζά γράμματα σε σύγκριση με το σύνολο δεδομένων δοκιμής. Και πάλι, όσον αφορά τα σημεία στίξης τα δύο σύνολα δεν φαίνεται να διαφέρουν πολύ. Στην πλειονότητα των αντιστοιχιών συνιστάται επίσης να αυξηθεί ο αριθμός των διαφορετικών tags. Η τάση να γράφεται με κεφαλαίο το πρώτο γράμμα κάθε πρότασης στον τίτλο είναι πολύ υψηλή και στα δύο σύνολα. Επιπλέον, όπως και στο σενάριο των μικτών κατηγοριών, τα σχόλια και οι αξιολογήσεις είναι ενεργοποιημένα και στα δύο σύνολα.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
person	music
man	video
group	new
standing	bts
woman	entertainment

Table 1.6: Στατιστικά αποτελέσματα για thumbnails και tags

Όσον αφορά το thumbnail, φαίνεται από τις πιο δημοφιλείς επεξεργασίες στον πίνακα 11.6 ότι υπάρχει μια προτίμηση να απεικονίζει ανθρώπους, πιθανότατα τον καλλιτέχνη ή τους καλλιτέχνες που δημιούργησαν το τραγούδι. Στη στήλη tags του πίνακα, βλέπουμε ότι τα πιο δημοφιλή και γενικά tags περιέχουν τις λέξεις "music", "video", "new" και "entertainment". Ωστόσο, συνήθως προτείνονται επίσης ετικέτες που περιέχουν το όνομα ενός δημοφιλούς συγκροτήματος αγοριών "BTS". Αυτό οφείλεται στο γεγονός ότι δεν υπήρχε τρόπος

να συγκεντρωθούν τα ονόματα όλων των καλλιτεχνών και των συγκροτημάτων, ώστε να αποτραπεί ο αλγόριθμος από το να μετατρέπει τα ονόματα των καλλιτεχνών εισόδου σε αυτά. Αυτό είναι ένα γενικό πρόβλημα σε αυτή την κατηγορία, καθώς δεν έχει καμία ουσία να προτείνεται η αλλαγή του τίτλου ενός βίντεο που είναι τραγούδι ή η αλλαγή του ονόματος του καλλιτέχνη. Επομένως, η εύρεση ουσιαστικών αντιστοιχιών μεταξύ μουσικών βίντεο είναι δύσκολη, με την έννοια ότι ο τίτλος, το thumbnail ή τα tags στις περισσότερες περιπτώσεις δεν περιγράφουν την ουσία του τραγουδιού, το είδος του και τη συνολική ενέργεια και "ατμόσφαιρα" του.

Αθλητισμός

Αυτή η κατηγορία αποτελείται από βίντεο που αφορούν τον αθλητισμό. Δεδομένου ότι έχει 5614 δείγματα, δημιουργούμε ένα σύνολο δεδομένων που περιέχει τα 2000 με τις περισσότερες προβολές. Το σύνολο δεδομένων δοκιμής μας αποτελείται από τα 2000 με τις λιγότερες προβολές. Τα στατιστικά αποτελέσματα αυτού του πειράματος παρουσιάζονται στους πίνακες 1.7, 1.8, 1.9.

Feature	**Increased**	**Decreased**	**Remains the same**
Title length	40.6%	51.5%	7.9%
Title's capital words ratio	49.8%	31.8%	18.4%
Title's compound score	30.9%	30.2%	38.9%
Number of exclamation marks in title	7.7%	14.9%	77.3%
Number of question marks in title	1.2%	4.5%	94.2%
Number of full stops in title	20.2%	19.0%	60.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	48.5%	21.9%	29.6%
Number of hyphens in title	16.1%	16.6%	67.2%
Number of different tags	42.7%	53.9%	3.4%

Table 1.7: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Feature	**True->False**	**False->True**	**True->True**	**False->False**
Ratings disabled	0.3%	0.8%	0%	98.9%
Comments disabled	0.4%	0%	0%	99.6%
First letter capitalization in title	18.8%	12.8%	64.4%	4.1%

Table 1.8: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Στους πίνακες 1.7, 1.8 μπορούμε να δούμε ότι στους τίτλους των viral βίντεο μας υπάρχει μεγαλύτερη αναλογία λέξεων με κεφαλαία προς πεζά γράμματα σε σύγκριση με το σύνολο δεδομένων δοκιμής. Και πάλι, όσον αφορά τα σημεία στίξης τα δύο σύνολα δεν φαίνεται να διαφέρουν πολύ, εκτός από τις κάθετες γραμμές, οι οποίες σύμφωνα με τις απαντήσεις θα έπρεπε να είναι αυξημένες στα μισά βίντεο της δοκιμής. Σε ένα ελαφρώς μεγαλύτερο ποσοστό αντιστοιχιών προτείνεται επίσης να μειωθεί ο αριθμός των διαφορετικών tags σε σύγκριση με το να αυξηθεί. Η τάση να γράφεται με κεφαλαίο το πρώτο γράμμα κάθε πρότασης στον τίτλο είναι και πάλι υψηλή και στα δύο σύνολα, αν και με μεγαλύτερο αριθμό εξαιρέσεων σε αυτή την κατηγορία. Επιπλέον, όπως και στο σενάριο των μικτών κατηγοριών, τα σχόλια και οι αξιολογήσεις είναι ενεργοποιημένα και στα δύο σύνολα.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
player	game
players	league
game	basketball
football	highlights
basketball	nba

Table 1.9: Στατιστικά αποτελέσματα για thumbnails και tags

Όσον αφορά το thumbnail, και πάλι προτιμάται να απεικονίζει ανθρώπους, και συγκεκριμένα σε αυτή την κατηγορία, παίκτες. Συνιστώνται επίσης εικόνες του εν λόγω αθλήματος, με μερικούς από τους πιο δημοφιλείς μετασχηματισμούς να είναι προς το "ποδόσφαιρο" και το "μπάσκετ", τα οποία είναι αναμφισβήτητα τα πιο δημοφιλή αθλήματα στις ΗΠΑ, όπου συλλέχθηκαν τα δεδομένα. Αυτό είναι αναμενόμενο, καθώς η πλειονότητα τόσο του συνόλου δεδομένων των viral βίντεο όσο και του συνόλου δεδομένων δοκιμής θα αφορά αυτά τα δύο αθλήματα. Για τα βίντεο που συζητούν πιο άγνωστα αθλήματα υπάρχει η πιθανότητα να μην βρουν ένα καλό ταίρι για να συγκριθούν. Οι πιο δημοφιλείς μετασχηματισμοί θεμάτων tags είναι προς γενικούς αθλητικούς όρους, αλλά και τύπους αθλημάτων.

Gaming

Αυτή η κατηγορία αποτελείται από περιεχόμενο που σχετίζεται με βιντεοπαιχνίδια, συμπεριλαμβανομένων, ενδεικτικά, του gameplay, κριτικών, tutorials και ειδήσεων της βιομηχανίας. Είναι η κατηγορία με τον μεγαλύτερο αριθμό δειγμάτων (8787 δείγματα). Δημιουργούμε ένα σύνολο δεδομένων που περιέχει τα 2500 με τις περισσότερες προβολές. Το δοκιμαστικό μας σύνολο δεδομένων αποτελείται από τα 2500 με τις λιγότερες προβολές. Τα στατιστικά αποτελέσματα αυτού του πειράματος παρουσιάζονται στους πίνακες 1.10, 1.11, 1.12.

Feature	Increased	Decreased	Remains the same
Title length	34.8%	55.0%	10.2%
Title's capital words ratio	30.6%	44.7%	24.7%
Title's compound score	38.4%	33.9%	27.7%
Number of exclamation marks in title	15.3%	22.6%	62.1%
Number of question marks in title	2.0%	5.4%	92.6%
Number of full stops in title	14.3%	13.9%	71.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	8.8%	13.0%	78.3%
Number of hyphens in title	14.0%	18.0%	67.9%
Number of different tags	38.3%	56.9%	4.8%

Table 1.10: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Για άλλη μια φορά, προτιμώνται μικρότεροι τίτλοι και λιγότερες λέξεις με κεφαλαία. Τα σημεία στίξης στον τίτλο υπάρχουν σε μικρό ποσοστό των δειγμάτων και συνιστάται ο αριθμός τους να μειώνεται στις περισσότερες περιπτώσεις. Ο αριθμός των tags είναι επίσης γενικά μικρότερος σε σύγκριση με τον αριθμό των tags στα βίντεο του συνόλου δεδομένων δοκιμής. Σε κάθε βίντεο του συνόλου δεδομένων μας φαίνεται ότι τόσο οι αξιολογήσεις όσο και τα σχόλια είναι ενεργοποιημένα και συνιστάται πάντα να είναι ενεργοποιημένα και στο σύνολο δεδομένων

δοκιμής. Η κεφαλαιοποίηση του πρώτου γράμματος σε κάθε πρόταση του τίτλου είναι επίσης κοινή πρακτική και για τα δύο σύνολα δεδομένων.

Feature	**True->False**	**False->True**	**True->True**	**False->False**
Ratings disabled	1.8%	0%	0%	98.2%
Comments disabled	2.6%	0%	0%	97.4%
First letter capitalization in title	15.9%	19.7%	60.3%	4%

Table 1.11: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Top 5 thumbnail edits/additions	**Top 5 tags edits/additions**
game	minecraft
video	game
series	battle
man	clash
screenshot	fortnite

Table 1.12: Στατιστικά αποτελέσματα για thumbnails και tags

Όσον αφορά τα thumbnails, οι πιο δημοφιλείς αλλαγές είναι προς εικόνες στιγμιότυπων οθόνης από βιντεοπαιχνίδια και επίσης απεικονίσεις ανδρών, ενδεχομένως των δημιουργών του περιεχομένου, ή των άβαταρ που χρησιμοποιούνε. Τα tags που περιέχουν τις λέξεις "παιχνίδι" και "μάχη" είναι επίσης δημοφιλείς, μαζί με διάφορα εμπορικά βιντεοπαιχνίδια, δηλαδή Minecraft, Fortnite. Αυτό σημαίνει ότι πολλά δείγματα σε αυτή την κατηγορία, ειδικά αυτά με τις περισσότερες προβολές, αφορούν τα πιο δημοφιλή βιντεοπαιχνίδια, πράγμα αναμενόμενο.

Άνθρωποι & Blogs

Αυτή η κατηγορία περιλαμβάνει ένα ευρύ φάσμα περιεχομένου που περιστρέφεται γύρω από τις προσωπικές εμπειρίες, την αφήγηση ιστοριών, τον τρόπο ζωής και την αλληλεπίδραση με την κοινότητα. Δεδομένου ότι αυτή η κατηγορία έχει 3769 μοναδικά δείγματα, δημιουργούμε ένα σύνολο δεδομένων που περιέχει τα 1000 δείγματα με τις περισσότερες προβολές. Το σύνολο δεδομένων δοκιμής μας αποτελείται από τα 1000 με τις λιγότερες προβολές. Τα στατιστικά αποτελέσματα αυτού του πειράματος παρουσιάζονται στους πίνακες 1.13, 1.14, 1.15.

Feature	Increased	Decreased	Remains the same
Title length	39.0%	52.3%	8.7%
Title's capital words ratio	26.8%	45.6%	27.5%
Title's compound score	40.4%	26.5%	33.0%
Number of exclamation marks in title	13.7%	31.3%	55.0%
Number of question marks in title	3.2%	5.4%	91.4%
Number of full stops in title	11.5%	16.6%	72.0%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	10.7%	11.4%	77.9%
Number of hyphens in title	9.9%	9.1%	81.0%
Number of different tags	30.5%	50.9%	18.6%

Table 1.13: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Feature	True->False	False->True	True->True	False->False
Ratings disabled	0.3%	2.3%	0.1%	97.3%
Comments disabled	2.3%	1.2%	0%	96.5%
First letter capitalization in title	14.7%	22%	58%	5.4%

Table 1.14: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Για άλλη μια φορά προτιμώνται μικρότεροι τίτλοι από εκείνους των δοκιμαστικών δεδομένων, καθώς και λιγότερες λέξεις γραμμένες με κεφαλαία. Επιπλέον, τα δεδομένα δοκιμής φαίνεται να περιέχουν περισσότερα θαυμαστικά από τα δείγματα του συνόλου δεδομένων μας. Ο αριθμός των tags τείνει επίσης να είναι μικρότερος για το σύνολο δεδομένων μας. Όπως και σε όλα τα άλλα σύνολα δεδομένων, οι αξιολογήσεις και τα σχόλια είναι ενεργοποιημένα. Η κεφαλαιοποίηση του πρώτου γράμματος στον τίτλο είναι επίσης εμφανής και στα δύο σύνολα δεδομένων, αλλά υπάρχει ένα μικρό τμήμα (22%) του συνόλου δεδομένων δοκιμής όπου αυτό παραβιάζεται.

Τα thumbnails του viral συνόλου δεδομένων μας, σε αντίθεση με αυτά του συνόλου δεδομένων δοκιμής, απεικονίζουν άνδρες και γυναίκες, που κάθονται ή κρατούν κάτι, ή μια εικόνα ίσως μιας οικογένειας ή ενός ατόμου σε διακοπές. Με άλλα λόγια, τα thumbnails τείνουν να δείχνουν ανθρώπους σε καθημερινές καταστάσεις. Ο πιο δημοφιλής μετασχηματισμός tags είναι προς θέματα οικογένειας και ζωής, vlogs και tags που περιέχουν τη λέξη funny (αστείο).

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
man	family
woman	real
holding	life
picture	funny
sitting	vlogs

Table 1.15: Στατιστικά αποτελέσματα για thumbnails και tags

Ψυχαγωγία

Αυτή η κατηγορία περιλαμβάνει μια ποικιλία περιεχομένου που αποσκοπεί στη διασκέδαση, την ψυχαγωγία και την καθήλωση των θεατών. Αυτό μπορεί να καλύπτει πολλά υποείδη, από κωμωδία και δράμα μέχρι ειδήσεις διασημοτήτων. Περιέχει 8539 δείγματα και είναι η δεύτερη μεγαλύτερη κατηγορία. Δημιουργούμε ένα σύνολο δεδομένων χρησιμοποιώντας τα 2500 δείγματα με τον υψηλότερο αριθμό προβολών. Δημιουργούμε επίσης ένα σύνολο δοκιμαστικών δεδομένων που αποτελείται από τα 2500 δείγματα με τον χαμηλότερο αριθμό προβολής. Τα στατιστικά αποτελέσματα από τα πειράματα παρουσιάζονται στους πίνακες 1.16, 1.17, 1.18.

Εκτός από το μήκος του τίτλου, το οποίο τα αποτελέσματά μας συνιστούν να είναι γενικά μικρότερο από ό,τι είναι στα δείγματα δοκιμής, δεν υπάρχουν άλλες ισχυρές γενικές κατευθύνσεις στους δύο πρώτους πίνακες. Αυτό μπορεί να εξηγηθεί από τον τύπο της κατηγορίας που εξετάζουμε. Σε αυτή την κατηγορία, δεδομένου ότι ασχολείται με ειδήσεις διασημοτήτων, κωμωδία και δράμα, τα βίντεο που λαμβάνουν τη μεγαλύτερη προβολή είναι είτε αυτά που συζητούν τρέχοντα γεγονότα και κουτσομπολιά, είτε είναι πολύ μοναδικά και διασκεδαστικά με τον τρόπο τους. Αυτό σημαίνει ότι δεν υπάρχει πραγματικά μια μυστική φόρμουλα που τα κάνει να πετύχουν, αλλά στην πραγματικότητα έχει να κάνει περισσότερο με τον συγχρονισμό και την ατομικότητα. Ως εκ τούτου, είναι πολύ δύσκολο να ταιριάξουν τα μη viral βίντεο με τα viral βίντεο με τρόπο που θα βοηθήσει στην αύξηση της δημοτικότητάς τους.

Για αυτή την κατηγορία, τα thumbnails τείνουν και πάλι να απεικονίζουν ανθρώπους και πρόσωπα. Τα tags περιέχουν γενικά τις λέξεις "αστείο", "diy" (κάνε το μόνος σου), "συμβουλές" κ.λπ. και προτείνονται συχνά αλλαγές προς τέτοιου είδους tags από τις εξηγήσεις με αντιπαράδειγμα.

Feature	**Increased**	**Decreased**	**Remains the same**
Title length	36.0%	55.3%	8.6%
Title's capital words ratio	36.2%	37.4%	26.4%
Title's compound score	36.4%	36.8%	26.8%
Number of exclamation marks in title	13.0%	19.4%	67.6%
Number of question marks in title	3.3%	5.8%	90.9%%
Number of full stops in title	9.5%	12.4%	78.1%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	22.1%	19.0%	58.9%%
Number of hyphens in title	14.9%	14.1%	71.0%
Number of different tags	43.2%	50.3%	6.5%

Table 1.16: Στατιστικά αποτελέσματα για αριθμητικά δεδομένα

Feature	**True->False**	**False->True**	**True->True**	**False->False**
Ratings disabled	0.4%	0.4%	0%	99.2%
Comments disabled	1%	1%	0%	97.9%
First letter capitalization in title	14.1%	16%	66%	3.9%

Table 1.17: Στατιστικά αποτελέσματα για κατηγορικά δεδομένα

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
man	funny
person	diy
woman	tips
movie	activities
face	challenge

Table 1.18: Στατιστικά αποτελέσματα για thumbnails και tags

1.3.3 Συγκεντρωτικά στατιστικά αποτλέσματα

Η ενδελεχής εξέταση των viral βίντεο στο YouTube σε σχέση με τα βίντεο με χαμηλό αριθμό προβολών σε διάφορες κατηγορίες αποκάλυψε μια σειρά από κοινά μοτίβα και χαρακτηριστικά που διαφοροποιούν το επιτυχημένο υλικό. Οι συντομότεροι τίτλοι, συχνά με λιγότερες λέξεις με κεφαλαία και λίγα σημεία στίξης, αποτελούν κοινό χαρακτηριστικό των viral βίντεο, αντανακλώντας μια προτίμηση για σαφήνεια και συντομία που πιθανώς απευθύνεται σε ένα ευρύτερο κοινό. Τα σημεία στίξης, όπως θαυμαστικά, ερωτηματικά και τελεία, χρησιμοποιούνται με φειδώ και ο αριθμός των tags τείνει να είναι μικρότερος αλλά πιο εστιασμένος. Τόσο οι βαθμολογίες όσο και τα σχόλια επιτρέπονται σχεδόν πάντα στα viral βίντεο, γεγονός που υποδηλώνει ότι το viewer engagement και η αλληλεπίδραση είναι σημαντικοί παράγοντες για την επιτυχία των βίντεο. Επιπλέον, είναι κοινή πρακτική να γράφεται με κεφαλαίο το πρώτο γράμμα των προτάσεων των τίτλων, καθώς προσδίδει επαγγελματισμό και ελκυστικότητα.

Επιπλέον, η δημοτικότητα των viral βίντεο εξαρτάται σε μεγάλο βαθμό από τα thumbnails τους, και ένα επαναλαμβανόμενο θέμα είναι η απεικόνιση των ανθρώπων, ιδίως σε καθημερινές ή γνώριμες καταστάσεις. Είναι πιθανό ότι αυτή η οπτική προσέγγιση προσελκύει τους θεατές δημιουργώντας μια αίσθηση περιέργειας ή οικειότητας. Επιπροσθέτως, ευρεία, γενικώς αγαπητά θέματα όπως "μουσική", "παιχνίδι", "αστείο", "οικογένεια" και "ζωή" χρησιμοποιούνται συχνά στα tags των επιτυχημένων βίντεο, καθώς είναι πιθανό να κεντρίσουν το ενδιαφέρον ενός μεγάλου κοινού.

Τα μοτίβα αυτά βρέθηκαν να ισχύουν για διάφορες κατηγορίες βίντεο, όπως η ψυχαγωγία, οι άνθρωποι & τα blogs, ο αθλητισμός, η μουσική και το gaming. Ο δυναμικός και ποικίλος χαρακτήρας των μουσικών βίντεο αντικατοπτρίζεται στη μεγαλύτερη αναλογία λέξεων με κεφαλαία και στα περισσότερα tags στην κατηγορία Μουσική. Τα αθλητικά βίντεο έδιναν προτεραιότητα σε τίτλους που τόνιζαν σημαντικούς όρους όπως "παίκτης" και "παιχνίδι", με μεγαλύτερη αναλογία κεφαλαίων λέξεων και λιγότερα tags. Παρόμοιες τάσεις μπορούσαν να παρατηρηθούν στα gaming βίντεο, τα οποία παρουσίαζαν κυρίως γνωστούς τίτλους όπως το Fortnite και το Minecraft. τα thumbnails και τα tags που αφορούσαν τα βίντεο της κατηγορίας People & Blogs έτειναν να ευνοούν θέματα καθημερινής ζωής και οικογένειας. Η μεγάλη ποικιλία περιεχομένου που παρατηρείται στα βίντεο ψυχαγωγίας -από κωμωδίες έως ειδήσεις διασημοτήτων- αναδεικνύει την αξία του συγχρονισμού και της μοναδικότητας, που καθιστούν πιο δύσκολο τον εντοπισμό μιας μοναδικής συνταγής επιτυχίας.

Συνολικά, η έρευνά μας δείχνει ότι, ανεξάρτητα από την κατηγορία περιεχομένου ενός βίντεο, η προσεκτικά μελετημένη επιλογή tags, η έξυπνη διάρθρωση τίτλων και τα συναρπαστικά thumbnails είναι βασικά στοιχεία που ενισχύουν το virality των βίντεο. Η απλότητα και η σχετικότητα αποτελούν βασικά συστατικά αυτής της διορατικής μεθοδολογίας που βοηθά τους δημιουργούς περιεχομένου να διευρύνουν το κοινό τους και τον αντίκτυπο των βίντεο τους.

1.4 Συμπεράσματα

1.4.1 Συζήτηση

Η παρούσα διπλωματική εργασία διερευνά τους περίπλοκους μηχανισμούς που επηρεάζουν την δημοτικότητα των βίντεο στην πλατφόρμα του YouTube. Αξιοποιήσαμε προηγμένα αναλυτικά εργαλεία, όπως η θεωρία γραφημάτων, η ανάλυση συναισθήματος και οι εξηγήσεις με αντιπαράδειγμα και εντοπίσαμε βασικά χαρακτηριστικά που

αυξάνουν τις πιθανότητες επιτυχίας των βίντεο. Πιο συγκεκριμένα, η έρευνά μας αναδεικνύει τη σημασία των συνοπτικών και ενδιαφερόντων τίτλων, των εικόνων thumbnail που απεικονίζουν ανθρώπους και των εστιασμένων tags. Η παροχή της δυνατότητας στους χρήστες να σχολιάζουν και να βαθμολογούν βίντεο έχει επίσης αποδειχθεί ότι ενισχύει τη δημοτικότητα των βίντεο.

Πραγματοποιήσαμε μια σειρά από γενικά πειράματα και πειράματα σε συγκεκριμένες κατηγορίες για να ανακαλύψουμε τα μοτίβα που ξεχωρίζουν τα βίντεο με υψηλό αριθμό προβολών από εκείνα με λιγότερες προβολές. Συγκεκριμένα, στην κατηγορία με τίτλο "Μουσική" τα βίντεο με περισσότερα tags και λέξεις με κεφαλαία στον τίτλο τείνουν να έχουν καλύτερες επιδόσεις. Στην κατηγορία με τίτλο "Αθλητισμός" τα tags τείνουν να είναι πιο ακριβή και λιγότερα. Στην κατηγορία με τίτλο "Gaming" δημοφιλή ονόματα βιντεοπαιχνιδιών κατέλαβαν τα tags και τα thumbnails τους συχνά απεικονίζουν στιγμιότυπα οθόνης από τα εν λόγω παιχνίδια. Τα βίντεο που βρέθηκαν στην κατηγορία "People & Blogs" τείνουν να παρουσιάζουν καθημερινές καταστάσεις στα thumbnails τους και έχουν tags που σχετίζονται με θέματα οικογένειας και καθημερινής ζωής. Τα βίντεο στην κατηγορία "Ψυχαγωγία", έχουν ποικίλο περιεχόμενο και τονίζουν την αξία του συγχρονισμού με τις εξελίξεις στην πραγματικότητα και της μοναδικότητας για την επίτευξη virality.

Η έρευνά μας υπογραμμίζει πόσο κρίσιμη είναι η χρήση thumbnails και tags για την προσέλκυση θεατών. Πιο συγκεκριμένα, διαπιστώθηκε ότι τα thumbnails που απεικονίζουν ανθρώπους, ιδίως σε σχετικές καταστάσεις, τείνουν να προσελκύουν μεγαλύτερη προσοχή. Επιπλέον, τα δημοφιλή και γενικά tags απευθύνονται σε ευρύ κοινό, αυξάνοντας έτσι τις πιθανότητες να κοινοποιηθεί και να συστηθεί ένα βίντεο.

Αν και η μελέτη αυτή προσφέρει πολύτιμες πληροφορίες, αντιμετωπίζει ορισμένους περιορισμούς. Το γεγονός ότι αναπτύξαμε και εκτελέσαμε του κώδικά μας στο Google Collab συνεπάγεται ότι δεν είχαμε τη δυνατότητα να εκτελέσουμε πειράματα με σύνολα δεδομένων μεγαλύτερα από 5000 δείγματα, καθώς τα αποτελέσματα θα χρειάζονταν πολύ χρόνο για να παραχθούν. Επιπλέον, το σύνολο δεδομένων μας περιείχε δεδομένα μόνο για τα ίδια τα βίντεο και όχι για γεγονότα που λάμβαναν χώρα στην πραγματικότητα την ίδια περίοδο, τα οποία θα μπορούσαν να βοηθήσουν στον εντοπισμό των αιτιών για τις οποίες ένα συγκεκριμένο βίντεο έγινε δημοφιλές σε μια συγκεκριμένη χρονική στιγμή. Τέλος, η μελέτη μας περιορίστηκε σε δεδομένα από τις ΗΠΑ.

Συμπερασματικά, η παρούσα διπλωματική εργασία συμβάλλει στη βαθύτερη κατανόηση των παραγόντων που ωθούν ένα βίντεο να γίνει viral. Οι γνώσεις που αποκομίσαμε από τη μελέτη μας ενδιαφέρουν όχι μόνο τους ακαδημαϊκούς ερευνητές, αλλά και οποιονδήποτε ασχολείται με τη δημιουργία και το μάρκετινγκ ψηφιακού περιεχομένου, καθώς θα μπορεί να βελτιώσει την ικανότητά του να παράγει βίντεο με υψηλό αριθμό προβολών, λαμβάνοντας υπόψη τα ευρήματα και τις συστάσεις μας. Καθώς τα ψηφιακά μέσα συνεχίζουν να εξελίσσονται, οι αρχές που περιγράφονται στην παρούσα έρευνα θα παραμείνουν σχετικές για την επίτευξη επιτυχίας στον ανταγωνιστικό κόσμο του YouTube.

1.4.2 Μελλοντικές Κατευθύνσεις

Κλείνοντας την παρούσα διατριβή, θα θέλαμε να δώσουμε μερικές προτάσεις για μελλοντικές βελτιώσεις της εργασίας μας ή για διαφορετικές ερευνητικές κατευθύνσεις. Αρχικά, η ανάλυση των thumbnails όλων των βίντεο του YouTube μπορεί να βελτιωθεί σημαντικά. Πιο συγκεκριμένα, αντί για την απλή δημιουργία μιας λεζάντας που περιγράφει την εικόνα, μπορούν να γίνουν πολλά όσον αφορά την ανάλυση χρώματος, την ανίχνευση αντικειμένων ή ακόμη και την ανάλυση συναισθήματος των προσώπων των ανθρώπων που υπάρχουν στην εικόνα. Θα είχε μεγάλο ενδιαφέρον να εξεταστεί κατά πόσον αυτές οι λεπτομέρειες έχουν πραγματικό αντίκτυπο στην δημοτικότητα, καθώς μπορεί κανείς διαισθητικά να σκεφτεί ότι τα φωτεινότερα χρώματα θα μπορούσαν να τραβήξουν την προσοχή των θεατών πιο εύκολα και γρήγορα ή ότι οι συγκλονιστικές εκφράσεις των ανθρώπων στην εικόνα ιντριγκάρουν τους θεατές, αναγκάζοντάς τους να εξερευνήσουν ορισμένα βίντεο. Ένας άλλος τρόπος με τον οποίο μπορούν να ενσωματωθούν εικόνες στην ανάλυση του virality θα ήταν να μην εξετάζεται μόνο το thumbnail, αλλά και στιγμιότυπα οθόνης από το σύνολο του βίντεο, ιδίως από τα πρώτα δευτερόλεπτα. Η ανάλυση των εν λόγω στιγμιότυπων οθόνης θα μπορούσε να δώσει πολλές ενδιαφέρουσες λεπτομέρειες. Για παράδειγμα, εάν το βίντεο αποτελείται από ένα άτομο που μιλάει για ένα συγκεκριμένο θέμα, η ανάλυση της έκφρασής του καθ' όλη τη διάρκεια της ομιλίας του θα μπορούσε να μας δώσει ενδείξεις για το γιατί το συγκεκριμένο βίντεο πέτυχε τόσο μεγάλη δημοτικότητα. Με άλλα λόγια, το αν ένα άτομο είναι ενθουσιώδες και ζωηρό καθ' όλη τη διάρκεια του βίντεο ή ήρεμο και συγκρατημένο θα μπορούσε να δείξει γιατί συγκέντρωσε μεγάλο αριθμό προβολών ή όχι, καθώς η ενέργεια του ατόμου που μιλάει σίγουρα επηρεάζει τον τρόπο με τον οποίο γίνεται αντιληπτή η ομιλία. Ένα άλλο μονοπάτι προς το οποίο μπορεί να επεκταθεί η έρευνά μας είναι

η συνεκτίμηση γεγονότων του πραγματικού κόσμου που λαμβάνουν χώρα περίπου την ίδια χρονική περίοδο με την ημερομηνία ανάρτησης των βίντεο. Αυτό έχει μεγάλη σημασία, καθώς, για παράδειγμα, η προβολή ενός δημοφιλούς επεισοδίου τηλεοπτικής σειράς με ένα συγκεκριμένο τραγούδι ως soundtrack μπορεί να προκαλέσει αύξηση του αριθμού των προβολών του επίσημου βίντεο του εν λόγω τραγουδιού στο YouTube. Άλλα παραδείγματα της επιρροής των γεγονότων του πραγματικού κόσμου στη δημοτικότητα των βίντεο στο YouTube είναι οι πολλές περιπτώσεις βίντεο με ειδήσεις διασημοτήτων. Η δημοτικότητα τέτοιων βίντεο δεν πηγάζει απλώς από το ίδιο το βίντεο, αλλά και από τη δημοτικότητα και τη συνάφεια των διασημοτήτων που συζητούν. Επιπλέον, ο συγχρονισμός παίζει μεγάλο ρόλο σε τέτοια σενάρια, καθώς, στις περισσότερες περιπτώσεις, τα βίντεο που ανεβαίνουν πρώτα παίρνουν τη μερίδα του λέοντος των προβολών. Επομένως, δεν πρέπει να λαμβάνονται υπόψη μόνο τα μεγάλα κοινωνικά γεγονότα, αλλά και η εγγύτητα των βίντεο σε αυτά είναι αρκετά σημαντική. Τέλος, η έρευνά μας περιορίστηκε στην εξέταση δεδομένων μόνο από τις ΗΠΑ. Στο μέλλον θα πρέπει να εξεταστούν και δεδομένα από άλλες χώρες.

Chapter 2

Introduction

In today's digital age, virality is a state chased by all content creators and influencers. In Youtube, viral videos specifically constitute a cultural phenomenon, capturing the attention of viewers and influencing or even sparking trends across social media platforms. Videos with the ability to become viral sensations are of great interest and value not only for content creators and influencers but also for marketers and businesses all over the world. Understanding the mechanics behind a video's virality potential can help improve successful content creation, so that it connects better with viewers, encourages interaction and reaches a large audience.

YouTube is objectively one of the largest video-sharing platforms and therefore receives more than two billion monthly visits from logged-in users. As a result, it constitutes the perfect space for a video to gain exposure and plays a significant role in the viral video phenomenon. YouTube videos that go viral are often turned into cultural icons, shape public opinion, establish trends and sometimes even have an actual impact on world events. In YouTube, common, everyday people can be elevated to celebrity status and new or already established brands can gain unheard-of visibility.

The importance of viral videos extends beyond the satisfaction of accumulating large view counts. For content creators and influencers, achieving virality often leads to financial rewards through ad revenue, as well as offers for sponsorship deals and increased channel popularity and subscribers, which will in turn augment their chances of achieving virality again in the future. For businesses, viral marketing campaigns drive sizeable sales growth, as well as brand recognition and establishment, which in turn leads to customer loyalty. For social movements or non-profit organizations, going viral equates to amplified support and advertisement of their cause which takes them closer to accomplishing actual social change. The rapid and extensive sharing of viral videos often leads to a cascade of effects, where their impact and reach are significantly boosted in a way that traditional marketing methods miss.

Despite the widespread fascination with viral videos, the underlying factors that drive their success are still quite obscure. Though their content can greatly vary, from humorous sketches and emotional stories to educational tutorials and breaking news, common elements should exist among these categories. Such elements could include compelling narratives, high production quality, relatability, and emotional resonance. Additionally, the timing of the video's release, the use of strategic keywords and tags, and the engagement metrics (likes, comments, shares) could play crucial roles in increasing a video's popularity. Recognizing and understanding these factors is critical for anyone aiming to produce videos with the same or even higher chances to captivate viewers.

This thesis seeks to develop an understanding of the factors that contribute to the virality of YouTube videos. Analyzing video metadata and their impact on viewership, our goal is to discover patterns that set highly viewed videos apart from those that receive less attention. We make use of sentiment analyzers, image captioners, embedding representations and counterfactual explanations to construct a detailed framework and become cognizant of the dynamics of video popularity.

The structure of this thesis includes several core sections. Initial chapters introduce essential concepts which we will be utilizing to create our framework. More specifically, these chapters include graph theory, sentiment

analysis, image captioners, semantic similarity, graph similarity and counterfactual explanations. These chapters are indispensable in order for the reader to be able to fully comprehend the methodologies applied in our research.

Graph theory allows us to represent and analyze the network of connections between the various metadata of our videos. More specifically, knowledge graphs will be used to visualize and understand how certain videos gain traction and spread across the platform. Sentiment analysis is used to help identify positive or negative connotations present in metadata, providing insights into how positive or negative sentiments influence viewer engagement. Image captioners aid in exploring thumbnail images and turning them into textual representations, which are much more easily compared and understood. Semantic similarity is the tool which we use to compare and contrast the textual information in our metadata in order to identify differences between videos in our dataset. Counterfactual explanations help us calculate the minimum graph edit distance through a framework developed by Filandrianos et al in [30]. They also grant us the advantage of being able to directly view the changes made in turning one graph into another, providing actionable recommendations for content creators to optimize their videos for maximum virality.

By combining the above information we construct a framework of video comparison in order to gather information specific to viral videos. In the section of the experiments, we discuss the trials we executed to test the framework and gather the results we desired. We explain and interpret the statistical findings in detail, as well as their implications. More specifically, we first examine a dataset of mixed video categories so as to identify universal trends. Moving forward, we divide our analysis into category-specific datasets and look for distinct traits that influence the virality potential of each category.

The factors taken into consideration are multiple. To begin with, the title is analyzed in many different ways. We extract keywords, apply sentiment analysis, track the use and multitude of punctuation, check for the percentage of words in capital form and whether or not the first letter of each sentence is capitalized and also take the title's length into consideration. The thumbnail is analyzed as well, using image captioners to produce a description of what is depicted. The tags are counted and keywords are extracted as well. These detailed analyses provide a comprehensive view of what drives a video to become viral. By comparing the aforementioned characteristics for each video we aim to identify trends present within them. This analysis will help us understand common patterns and variations in the content, such as the frequency of certain keywords, images, tags, and the overall effectiveness of different video formats.

All in all, this thesis not only improves our understanding of viral videos academically, but it also delivers practical strategies in order to improve online video content. More specifically, our findings underscore the importance of title optimization, engaging thumbnails and effective tag usage to boost viewer engagement. The insights gained from this study can help shape the future of digital content creation, offering a roadmap for achieving greater visibility and impact in the crowded online landscape.

Apart from contributing to the field of digital content analysis, this thesis also lays the groundwork for future studies in this ever-evolving area. More precisely, as new data analysis techniques continue to evolve, they can easily join the study of YouTube video virality, resulting into better insights. Artificial intelligence offers some very exciting possibilities for the future of digital content creation and distribution. Moreover, our research highlights the importance of understanding viewers' behavior and preferences. Therefore, the analysis of viewer interactions, such as likes, comments, and shares, could potentially produce some even deeper insights into the factors driving virality. Additionally, understanding the role of social media algorithms in promoting content is crucial to maximize a video's reach. Platforms like YouTube use complex algorithms to determine which videos are recommended to users, and by aligning content with these algorithms, creators can increase the likelihood of their videos being seen by a broader audience.

Besides, the implications of our research extend beyond the realm of online video content. In actuality, the methodologies used and the results found can also be applied to many different areas of digital media and marketing, becoming a blueprint for achieving success in various digital platforms. The understanding of the concept of virality through our work can aid content creators, businesses and influencers to better navigate the digital landscape. In other words, the principles and strategies uncovered here can be adapted with ease to a wide range of contexts, which ensures that they continue to stay relevant throughout the ever-evolving digital landscape. Through our work, we aspire to inspire content creators of all backgrounds and kinds with the tools and knowledge necessary to harness the full potential of virality.

In conclusion, this thesis not only sheds light on the intricacies of what makes a video go viral but also provides a comprehensive framework for leveraging these insights in practical applications. The detailed examination of various factors, combined with advanced analytical techniques, makes this research a valuable resource for anyone involved in the digital media space. By understanding and applying the principles outlined in this study, content creators and marketers can significantly enhance their ability to produce viral content, ultimately achieving greater success in the competitive world of online video.

Chapter 3

Graphs

A graph is a pivotal structure in mathematics and computer science, which provides a powerful framework for modeling relationships between objects. A graph G consists of vertices (nodes) and edges (links) that connect pairs of vertices. This chapter starts by exploring the fundamental concepts of graph theory, detailing how graphs are defined, represented, and utilized to solve various problems. We delve into different types of graphs, such as undirected and directed graphs, and their respective applications. Furthermore, we discuss bipartite graphs and their significance in solving matching problems, highlighting algorithms like the Hungarian method. Finally, the chapter introduces knowledge graphs, which extend traditional graphs by incorporating semantic information, making them indispensable in artificial intelligence and data integration applications. Through this comprehensive overview, we aim to highlight the versatility and importance of graphs in numerous fields.

Contents

3.1 Graph Theory Basics

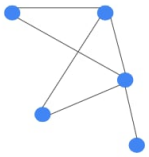
A graph, denoted G, is a mathematical structure used to model pairwise relations between objects. A graph is defined as a pair $G = (V, E)$, where:

- V is a non-empty set of elements called vertices or nodes.
- E is a set of unordered pairs of distinct vertices, called edges or links.

Each edge $e \in E$ is represented as $e = \{u, v\}$, where u and v are vertices in V. If the order of the vertices in each edge matters, the graph is called a directed graph or digraph, and each edge e is represented as an ordered pair $e = (u, v)$. In such graphs, an edge e is sometimes called a directed edge or arc. If $e = (u, v)$, then the vertices u and v are known as the tail and head of the edge e, respectively. The edge e can also be referred to as an outgoing edge from u or an incoming edge to v [99].

In academic literature, the term "graph" typically refers to a simple graph, which is defined as a graph that does not contain self-loops (edges that connect a vertex to itself) and has at most one edge between any two vertices.

An example of an undirected graph can be seen in Figure 3.1.1a. This graph can also be described using an adjacency list or an adjacency matrix. An adjacency list of a graph consists of each vertex listed separately, followed by a list of vertices to which it is connected. An **adjacency list** of a graph is a collection where each vertex is listed in a separate row, followed by a list of vertices to which it is adjacent. Conversely, an **adjacency matrix** of a graph G with vertex set $\{v_1, \ldots, v_n\}$ is an $n \times n$ matrix. In this matrix, the entry in row i and column j is 1 if there is an edge between vertices v_i and v_j in G, and 0 otherwise. Each method of representation has its own computational advantages and disadvantages. For graphs with relatively few edges, known as sparse graphs, an adjacency list is more space-efficient than an adjacency matrix. In contrast, for graphs with many edges, referred to as dense graphs, the space usage of both representations is similar. Representing a graph with a drawing is often advantageous, as one of the main appeals of graph theory is that graphs can be visually depicted and analyzed.



(a) Visual Representation

Vertex	Adjacent to
1	2, 3
2	1, 3, 4
3	1, 2, 4
4	2, 3, 5
5	4

(b) Adjacency List

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

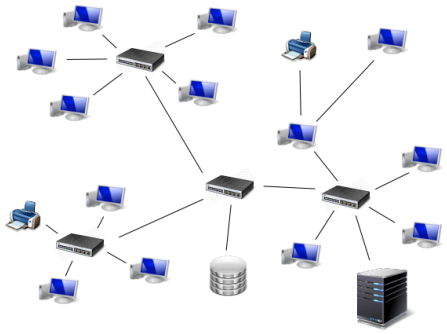
(c) Adjacency Matrix

Figure 3.1.1: Representation of undirected graph [9].

A graph whose vertices are named is called a labeled graph. The neighborhood $N(u)$ of an entity u in a graph is defined as the set of nodes adjacent to it. The number of edges that are incident to a vertex is called the degree of the vertex.

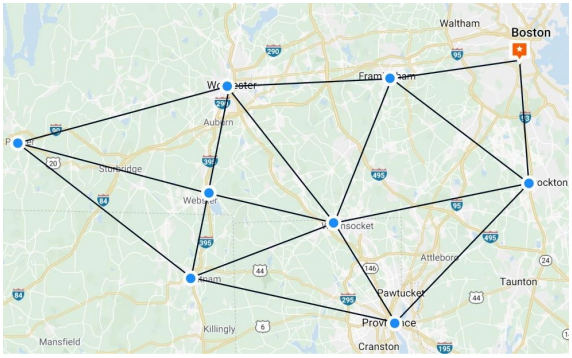
A **weighted graph** is a type of graph in which each edge has an associated numerical value, called a weight. These weights can represent various attributes such as distances, costs, capacities, or any other measure that applies to the connection between two vertices. In the adjacency matrix, instead of displaying the value 1 to signify the existence of an edge, weighted graphs display the weight of each edge.

Graphs are indispensable tools in various fields due to their ability to model complex relationships and structures. Their flexibility and the wealth of algorithms developed for them make them suitable for solving a wide range of problems in computer science, engineering, biology, social sciences, and more. Whether it's for analyzing social networks, optimizing transportation routes, or managing network traffic, graphs play a crucial role in many modern applications. Some examples of graph models are depicted in Figure 3.1.2.



(a) A computer network graph

(b) Social Network graph



(c) Transportation network graph

Figure 3.1.2: Graph reprentation examples

3.2 Bipartite Graphs and Minimum Weight Full Matching Problem

A graph $G = (V, E)$ is termed bipartite if its vertex set V can be split into two disjoint subsets, U and W, in such a way that no edges connect vertices within the same subset. This property makes bipartite graphs particularly useful in various applications across computer science, mathematics, and related fields. The division (U, W) is known as a bipartition of G. Figure 3.2.1 depicts an example of a bipartite graph [63].

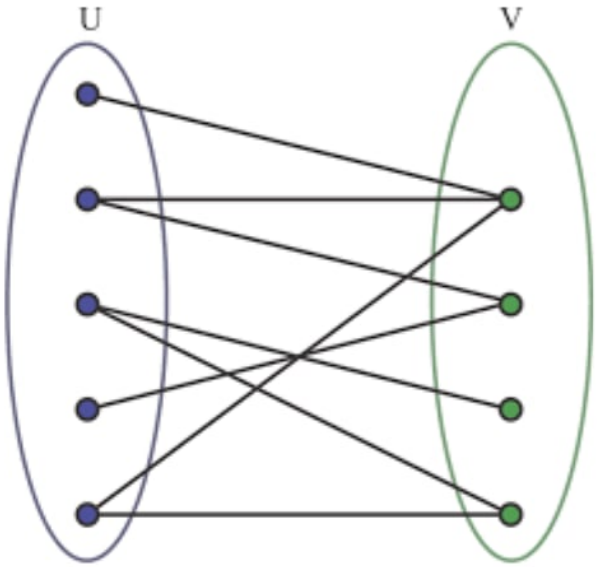


Figure 3.2.1: An example bipartite graph [63]

A graph is termed equally bipartite if it is bipartite and the bipartition results in subsets X and Y having the same number of vertices [99].

A bipartite graph $G = (V, E)$ with vertex sets partitioned into U and W is called a complete bipartite graph if each vertex in U is connected to every vertex in W [63].

Bipartite graphs are frequently used in matching problems where one needs to pair vertices in U with vertices in W. One such problem is the problem of determining a **minimum weight full matching**, also known as the rectangular linear assignment problem, a fundamental issue in combinatorial optimization. This problem involves finding the optimal way to match all elements of two sets such that the total cost is minimized.

Problem Definition: Given two sets U and V with potentially different sizes and a set of edges E connecting elements of U to elements of V, each edge $(u, v) \in E$ has an associated weight $w(u, v)$. The goal is to find a matching $M \subseteq E$ that pairs each element of U with exactly one element of V (and possibly vice versa) such that the sum of the weights of the edges in M is minimized.

Mathematical Formulation: The problem can be formulated as follows:

$$\min \sum_{(u,v)\in M} w(u, v)$$

subject to:

$$|M| = \min(|U|, |V|)$$

This ensures that the number of edges in the matching M is equal to the smaller of the sizes of U and V, guaranteeing that each element in the smaller set is matched to exactly one element in the larger set.

Several algorithms have been proposed to solve this problem, such as:

- Using the Hungarian algorithm (or else the Kuhn-Munkres algorithm): A method to find the optimal assignment in polynomial time. The time complexity of the Hungarian algorithm is $O(n^3)$ [52, 65].
- Using Linear Programming: Formulating the problem as a linear program and solving it using standard techniques. This displays exponential worst-case scenario time complexity[18].

- Using the Auction Algorithm: A decentralized algorithm that iteratively improves the assignment by mimicking an auction process. The worst-case time complexity of the auction algorithm is $O(n^3)$ [8].

However, the solution that we are interested in is Karp's algorithm [46]. In his research, Richard M. Karp introduces an efficient method for addressing the assignment problem. The proposed algorithm achieves an expected time complexity of $O(mn \log n)$ by utilizing priority queues and assuming the independence of edge costs as random variables. This approach offers a significant improvement in computational efficiency for large assignment problems.

3.3 Knowledge Graphs

3.3.1 General information

The term "knowledge graph" has a range of definitions, some of which conflict with one another, spanning from specific technical descriptions to broader, more general interpretations. Despite the term "knowledge graph" being cited in literature since at least 1972 [97], its modern significance emerged with Google's announcement of the Google Knowledge Graph in 2012 [28]. Their Knowledge Graph was developed using data from DBpedia [2], Freebase [11], and other sources. The Google Knowledge Graph effectively complemented Google's string-based search, and its success popularized the term "Knowledge Graph" in the online community. This announcement sparked a wave of similar initiatives from companies Facebook [87], LinkedIn, Airbnb, Microsoft, Amazon, Uber [97] and eBay [67]. Knowledge graphs have been implemented across multiple fields, from social sciences [70, 84, 47] to music [19] to medicine [102], demonstrating their versatility and utility. Additionally, the application of knowledge graphs in assisting machine learning models with various tasks has consistently yielded positive results [21, 19, 102].

A data graph is the foundation of every knowledge graph. For our purposes, we follow Hogan et al. [41] and define a knowledge graph as a data graph designed to collect and represent knowledge about the real world. In this graph, nodes symbolize entities of interest, and edges depict various possible relationships between these entities. This data graph adheres to a graph-based data model, which could be a directed edge-labeled graph, a heterogeneous graph, a property graph, or similar structures. Generally, the terms "knowledge graph" and "knowledge base" are considered synonymous and used interchangeably.

The graphs which we utilize for our endeavor are heterogeneous directed edge-labeled knowledge graphs.

Directed edge-labeled graph: A directed edge-labeled graph (or del graph) is defined as a set of nodes connected by directed labeled edges. In knowledge graphs, nodes represent entities and edges represent binary relations between these entities. This structure allows for flexibility in integrating new data compared to traditional relational models, as it doesn't require predefined schemas [41].

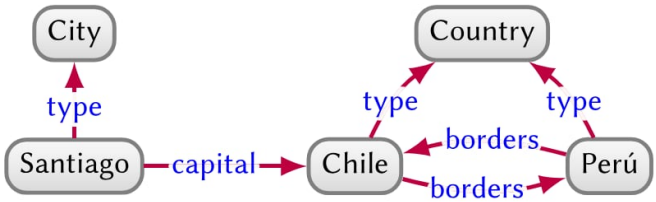
Heterogeneous graph: A heterogeneous graph ([100], [43], [95]) is a graph where each node and edge has a specific type. These graphs are similar to del graphs, as they use edge labels to denote types, but differ in that the node types are part of the graph model itself rather than being represented as a special relationship. For example, an edge is called homogeneous if it connects two nodes of the same type (e.g., borders); otherwise, it is called heterogeneous (e.g., capital). Heterogeneous graphs are useful for partitioning nodes by type, which is beneficial for machine learning tasks. Unlike del graphs, heterogeneous graphs typically have a one-to-one relationship between nodes and types, but there can be nodes with zero types and nodes with multiple types.

Knowledge graphs are essential for integrating and organizing complex data relationships in a machine-readable format, enhancing data integration, search accuracy, and AI applications. They encode semantic relationships, providing deeper contextual understanding crucial for chatbots and virtual assistants. Their flexibility and scalability allow easy adaptation to new information, making them ideal for handling complex queries in fraud detection, drug discovery, and supply chain management. By revealing hidden patterns and connections, knowledge graphs improve decision-making processes, benefiting fields like healthcare, finance, and education. As data volume and complexity grow, the role of knowledge graphs in unlocking data value and enhancing decision-making becomes increasingly vital.

Given the crucial role of knowledge graphs in processing diverse information in a machine-readable format, substantial research has been ongoing in this area over the past few years. Knowledge graphs have been

employed in various AI systems [50], including recommender systems, question-answering systems, and information retrieval. They are also widely used across numerous fields such as education and healthcare to improve human life and society [83], [71].

However, research on knowledge graphs still faces notable technical challenges. For instance, gathering knowledge from multiple sources and integrating it into a coherent knowledge graph is quite a difficult task using today's technologies. Hence, while knowledge graphs hold great promise for modern society, their development is still fraught with technical difficulties [71].



(a) Del graph

(b) Heterogeneous graph

Figure 3.3.1: Data about capitals and countries in a del graph and a heterogeneous graph [41].

3.3.2 Resource Description Framework

We will utilize the Resource Description Framework (RDF) as the data model for our knowledge graphs. RDF, a standard by the World Wide Web Consortium (W3C), was initially created to serve as a metadata data model. It has since evolved into a general-purpose method for describing and exchanging graph data. RDF utilizes a simple yet effective model based on triples (subject, predicate, object) to represent data, allowing it to describe complex relationships and integrate data from various sources.

RDF is especially significant in knowledge graph construction and management. It enables the merging of data even when underlying schemas differ, facilitating seamless data integration and interoperability. This capability is crucial for applications in artificial intelligence (AI), where knowledge graphs enhance systems like recommender engines, question-answering systems, and information retrieval tools. By providing structured and semantically rich data, RDF supports advanced data analytics and machine learning tasks, including node classification and link prediction.

Additionally, RDF offers various serialization formats, with Turtle being the most popular due to its human-readable syntax. RDF-star, an extension of RDF, introduces the concept of quoted triples, allowing for more complex and nested data representations, which further enhances its utility in sophisticated data mining and AI applications.

Research continues to address the challenges and opportunities in RDF's application, focusing on areas such as knowledge graph completion, knowledge fusion, and reasoning. These efforts aim to improve the quality and comprehensiveness of knowledge graphs, making RDF a cornerstone of modern data representation and analysis [71], [33], [27].

At its core, RDF uses a simple structure known as a triple to represent data. Each triple consists of three components:

- Subject: The resource being described.
- Predicate: The property or relationship of the subject.

- Object: The value of the property or another resource linked to the subject.

For example, to represent the fact that "John Smith created a webpage," the RDF triple would look like this:

- Subject: http://www.example.org/index.html
- Predicate: http://purl.org/dc/elements/1.1/creator
- Object: http://www.example.org/people/JohnSmith

Or to represent the knowledge that the book named "John Smith Autobiography was" written by John Smith and is about John Smith, the needed RDF triples could be the following:

First:

- Subject: http://www.example.com/book45
- Predicate: http://www.example.com/title
- Object: "John Smith Autobiography"

Second:

- Subject: http://www.example.com/book45
- Predicate: http://www.example.com/subject
- Object: blank node

Third:

- Subject: blank node
- Predicate: http://www.example.com/wrote
- Object: http://www.example.com/book45

Fourth:

- Subject: blank node
- Predicate: http://www.example.com/name
- Object: "John Smith"

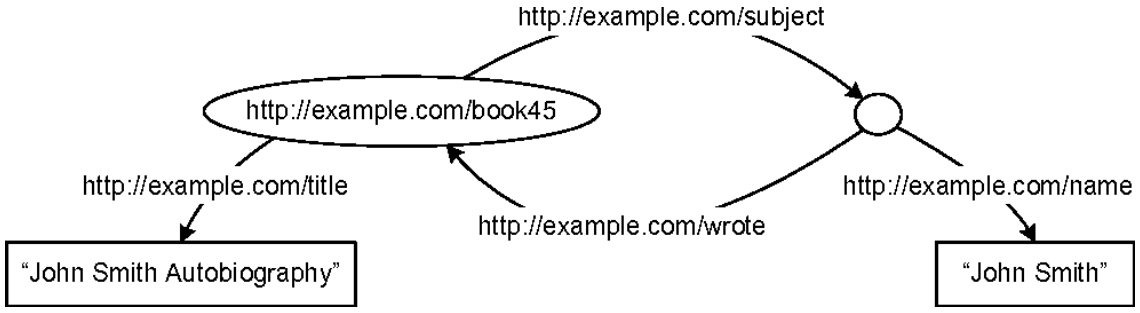


Figure 3.3.2: RDF example graph

URIs and Literals: RDF uses Uniform Resource Identifiers (URIs) to uniquely identify subjects and predicates. Objects can be either URIs or literals (data values such as strings, numbers, or dates). This system allows RDF to describe any resource, whether it is a web page, a person, or an abstract concept [57], [71], [27].

Blank nodes: Blank nodes in RDF (Resource Description Framework) are used to represent resources that do not have a global identifier (URI) or when the specific identity of the resource is not important. They are

useful for modeling complex or nested data structures, providing a way to represent intermediate nodes or anonymous resources without the need to generate unique URIs [57], [71], [27].

RDF Graph: A collection of RDF triples forms an RDF graph, a directed labeled graph where nodes represent subjects and objects, and edges represent predicates. This graph structure enables complex relationships and data integration across different datasets [57], [71], [27].

Chapter 4

Sentiment Analysis

Natural Language Processing (NLP) has made significant strides, particularly with the development of large language models (LLMs) that exhibit advanced comprehension and reasoning abilities [35, 68, 37]. Among various NLP tasks, sentiment analysis, stands out for its exceptional accuracy [29]. Sentiment analysis, often referred to as opinion mining, is a significant technique in the field of Natural Language Processing (NLP) that focuses on identifying and extracting subjective information from text. This technique involves categorizing the sentiment conveyed in a piece of text as positive, negative, or neutral, providing valuable insights into people's emotions and opinions.

Sentiment analysis is widely used in various sectors such as business, politics, healthcare, and social media monitoring to gauge public sentiment, monitor brand reputation, and improve customer experiences. This chapter explores the fundamental principles of sentiment analysis, its diverse applications, and the methodologies employed, including lexicon-based approaches, machine learning techniques, and hybrid methods.

Additionally, we discuss the tool VADER (Valence Aware Dictionary and sEntiment Reasoner) which is particularly effective for analyzing sentiments in social media texts. By understanding these concepts, we aim to highlight the importance of sentiment analysis in harnessing the power of human sentiment for strategic decision-making and enhanced interaction with digital content.

Contents

4.1 Introduction to Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is a Natural Language Processing (NLP) technique aimed at identifying and extracting subjective information from text. This process involves determining the sentiment expressed in a piece of text, classifying it as positive, negative, or neutral. By analyzing emotions, opinions, and attitudes, sentiment analysis provides valuable insights into public perception and emotional response [79].

4.2 Applications of Sentiment Analysis

Sentiment analysis has widespread applications across various domains [55]:

Business: In the business sector, sentiment analysis is crucial for understanding customer feedback and market trends. Companies use it to monitor brand reputation, assess customer satisfaction, and analyze product reviews. For instance, by examining social media posts and online reviews, businesses can gauge public sentiment towards their products and services, enabling them to make informed decisions and tailor their marketing strategies.

Politics: In politics, sentiment analysis is used to measure public opinion on policies, political figures, and events. It helps political analysts and campaign managers understand voter sentiment, track changes in public opinion, and devise strategies to address public concerns. By analyzing social media discussions and news articles, sentiment analysis can reveal insights into voter behavior and predict election outcomes.

Healthcare: In the healthcare industry, sentiment analysis plays a role in improving patient care and services. By analyzing patient feedback from surveys, social media, and online forums, healthcare providers can identify areas for improvement, understand patient experiences, and enhance the quality of care. Sentiment analysis can also monitor public sentiment regarding health policies and medical treatments.

Social Media Monitoring: Sentiment analysis is extensively used in social media monitoring to track and analyze user sentiments. Social media platforms generate vast amounts of unstructured data containing valuable opinions and emotions. Sentiment analysis tools can process this data to understand public reactions to events, brands, and topics. Companies, governments, and organizations leverage this information to respond to public sentiment, manage crises, and engage with their audience effectively.

4.3 Techniques in Sentiment Analysis

Several approaches are employed in sentiment analysis [79]:

Lexicon-based Methods: Lexicon-based methods rely on predefined dictionaries of sentiment-laden words. These dictionaries assign sentiment scores to words or phrases, which are then used to determine the overall sentiment of a text. While straightforward, lexicon-based methods can struggle with context and the nuances of language, such as sarcasm and idioms.

Machine Learning Methods: Machine learning methods involve training algorithms on labeled datasets to predict sentiment. These methods can capture more complex patterns and context within the text. Common machine learning approaches include:

Supervised Learning: Algorithms are trained on annotated datasets where the sentiment is already labeled. Examples include Naive Bayes, Support Vector Machines, and neural networks.

Unsupervised Learning: These methods identify sentiment without labeled data, often using clustering or topic modeling techniques to infer sentiment from patterns in the data.

Hybrid Methods: Hybrid methods combine lexicon-based and machine learning approaches to improve accuracy. By leveraging the strengths of both methods, hybrid approaches can achieve better performance in sentiment classification. They use lexicons to handle common sentiment words and machine learning models to capture contextual information and complex patterns.

4.4 Challenges in Sentiment Analysis

Despite its utility, sentiment analysis faces several challenges [79], [14], [55]:

Sarcasm Detection: Detecting sarcasm is a significant challenge in sentiment analysis. Sarcastic statements often express a sentiment opposite to the literal meaning of the words, making it difficult for algorithms to interpret correctly. Advanced NLP techniques and contextual understanding are required to address this issue effectively.

Context Understanding: Understanding the context in which words are used is crucial for accurate sentiment analysis. Words can have different meanings based on their context, and sentiment can change depending on the situation. For example, the word "great" can express positive sentiment in "great job," but negative sentiment in "great, just what I needed" (sarcastically). Capturing this nuance is challenging for sentiment analysis models.

Multilingual Support: Developing sentiment analysis models that work across different languages and dialects is complex. Each language has unique linguistic features, idioms, and cultural contexts that affect sentiment expression. Multilingual models must be trained on diverse datasets to handle these variations effectively.

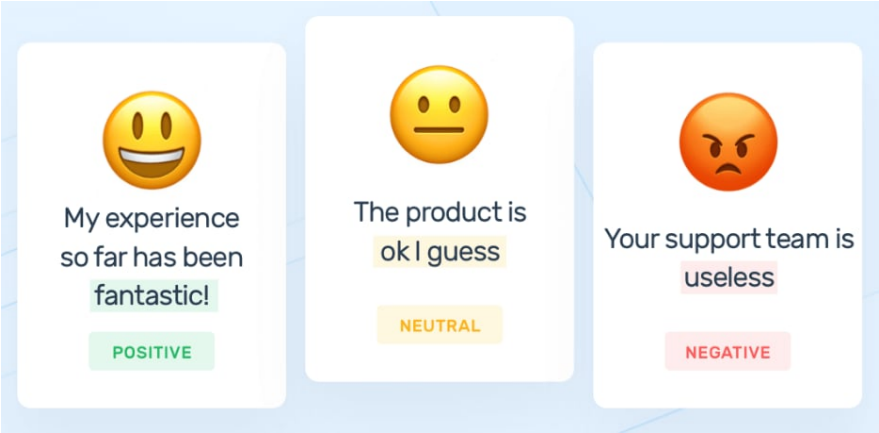


Figure 4.4.1: An example of sentiment analysis [64]

4.5 VADER: Valence Aware Dictionary for Sentiment Reasoning

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool specifically attuned to social media texts. Developed by C.J. Hutto and Eric Gilbert at Georgia Institute of Technology [44], VADER is designed to handle the informal language, abbreviations, and emoticons commonly used in social media contexts.

4.5.1 Lexicon Development

To develop the VADER sentiment lexicon, the researchers began by examining existing sentiment wordbanks such as LIWC [72], ANEW [12], and GI [82]. They expanded this list by incorporating lexical features common in microblogs, including Western-style emoticons (e.g., ":-)" for a smiley face), sentiment-related acronyms and initialisms (e.g., LOL and WTF), and commonly used slang (e.g., "nah," "meh," and "giggly"). This initial list comprised over 9,000 lexical features.

Next, they assessed the applicability of each feature using a wisdom-of-the-crowd (WotC) approach, gathering intensity ratings from ten independent human raters through Amazon Mechanical Turk (AMT)[1]. Each feature was rated on a scale from -4 (Extremely Negative) to +4 (Extremely Positive), resulting in over 90,000 ratings.

[1]Amazon Mechanical Turk. (n.d.). Retrieved June 9, 2024, from https://www.mturk.com

They retained features with non-zero mean ratings and standard deviations less than 2.5, refining the list to over 7,500 validated lexical features.

The validated lexicon includes examples such as "okay" (0.9), "good" (1.9), "great" (3.1), "horrible" (-2.5), the frowning emoticon ":(" (-2.2), and "sucks" (-1.5). This gold-standard list, with associated valence scores indicating sentiment polarity and intensity, forms the core of the VADER sentiment lexicon [44].

4.5.2 How it works

VADER combines a sentiment lexicon with a set of grammatical and syntactical rules to determine the sentiment of a text. The lexicon includes a list of words, acronyms, initialisms, and emoticons, each assigned a sentiment intensity score. VADER then applies five heuristics to account for the nuances of sentiment expression in text:

1. Punctuation: Exclamation points increase sentiment intensity.
2. Capitalization: Words in ALL-CAPS are given more weight.
3. Degree Modifiers: Words like "very" or "extremely" modify the intensity of the sentiment.
4. Conjunctions: The word "but" signals a shift in sentiment polarity.
5. Negation: The presence of negation words flips the sentiment polarity.

When a sentence is analysed using VADER, 4 scores (pos, neu, neg, compound) are produced and each one represents different aspects of the sentiment expressed in the text. Here's a breakdown of what each score signifies:

Pos (Positive): This score indicates the proportion of the text that is perceived to be positive. It is a float value between 0 and 1, where higher values indicate a stronger presence of positive sentiment in the text.

Neu (Neutral): This score represents the proportion of the text that is neutral. Like the positive score, it is a float value between 0 and 1. A higher value indicates that a larger portion of the text is neutral, meaning it does not strongly convey positive or negative sentiment.

Neg (Negative): This score indicates the proportion of the text that is perceived to be negative. It is also a float value between 0 and 1, with higher values indicating a stronger presence of negative sentiment.

Compound: The compound score is a normalized, weighted composite score that takes into account all the other scores (positive, neutral, and negative) to provide a single sentiment score. It ranges from -1 (most extreme negative) to +1 (most extreme positive). This score is calculated by summing the valence scores of each word in the text, adjusted according to the heuristics, and then normalized to be between -1 and 1.

Interpretation of Compound Score:

Positive Sentiment: If the compound score is greater than 0.05, the sentiment of the text is generally considered positive.

Neutral Sentiment: If the compound score is between -0.05 and 0.05, the sentiment is considered neutral.

Negative Sentiment: If the compound score is less than -0.05, the sentiment is generally considered negative.

4.5.3 Performance and Advantages

VADER is particularly effective in analyzing sentiments expressed in short, informal texts such as tweets and Facebook posts. It has been shown to perform as well as, and in some cases better than, human raters and other sentiment analysis tools, with an F1 classification accuracy of 0.96 compared to 0.84 for individual human raters [44]. VADER's rule-based approach allows it to be computationally efficient, making it suitable for real-time sentiment analysis on streaming data. With its lexicon-based and rule-based approach, VADER offers a robust solution for analyzing sentiments in social media texts, providing high accuracy and efficiency without the need for extensive training data.

4.5.4 Why it was chosen

We chose to use VADER for this thesis due to its special attunement to social media texts. More specifically, we needed a sentiment analysis tool that would be able to determine the sentiment of informal texts written by people on YouTube, which are expected to be very similar to texts written in other social media platforms. VADER also takes into account capitalization, punctuation and emoticons to calculate its scores, which is very useful again in the context of YouTube videos.

Chapter 5

Image Captioners

Image captioning is a sophisticated task that merges the fields of computer vision and natural language processing to generate descriptive text for images. This task necessitates a deep understanding of visual content and the ability to produce coherent, contextually relevant textual descriptions.

The significance of image captioning is evident in its diverse applications, ranging from aiding visually impaired individuals and enhancing image search engines to automating content creation and improving human-computer interactions. This chapter delves into the evolution of image captioning techniques, highlights the advancements brought by deep learning models, and discusses the evaluation metrics and challenges faced in this domain.

Special attention is given to innovative models like GIT (Generative Image-to-text Transformer), which have set new benchmarks in unifying vision and language tasks, showcasing the latest strides in image captioning technology and its potential future directions.

Contents

5.1 Overview

Image captioning is a complex task that involves generating a textual description of an image. This task lies at the intersection of computer vision and natural language processing, requiring models to understand and interpret visual content and subsequently generate coherent and contextually relevant descriptions. The importance of image captioning spans across various applications, including assisting visually impaired individuals, enhancing image search engines, automating content creation, and improving human-computer interaction [91].

5.2 Techniques and Models

The development of image captioning systems has evolved significantly over the years. Early methods relied on template-based approaches, where predefined templates were filled with detected objects and their attributes. However, these methods were limited by their rigidity and inability to generalize well to diverse images.

With the advent of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), more sophisticated approaches emerged. CNNs are used to extract visual features from images, while RNNs, particularly Long Short-Term Memory (LSTM) networks, are employed to generate sequences of words based on these features.

A landmark model in image captioning is the "Show and Tell" model developed by Vinyals et al. (2015). This model uses a CNN to encode the image into a fixed-dimensional feature vector, which is then fed into an LSTM to produce the caption. The process can be summarized in three steps [91]:

Feature Extraction: A pre-trained CNN, such as InceptionV3, extracts features from the image.

Caption Generation: An LSTM network generates the caption word-by-word, conditioned on the extracted features.

Training: The model is trained end-to-end using a dataset of images paired with corresponding captions, such as the MS COCO dataset [80].

5.3 Evaluation Metrics

Evaluating image captioning models involves comparing the generated captions to reference captions. Common metrics include BLEU, METEOR, ROUGE, and CIDEr. These metrics, borrowed from machine translation and text summarization tasks, measure the n-gram overlap between the generated and reference captions, assessing aspects like precision, recall, and semantic similarity [91].

5.4 Challenges

Despite significant progress, image captioning remains a challenging task due to:

Diversity in Visual Content: Images can contain a wide variety of objects, scenes, and activities.

Contextual Understanding: Generating accurate and contextually appropriate descriptions requires a deep understanding of the relationships between objects and the overall scene.

Language Generation: Producing fluent and grammatically correct sentences is crucial for usability [91].

Figure 5.4.1: A selection of image captioner results, grouped by human rating [91]

5.5 GIT: A Generative Image-to-text Transformer for Vision and Language

GIT (Generative Image-to-text Transformer) is a novel model developed to unify vision-language tasks, particularly focusing on image and video captioning as well as question answering. Developed by a team at Microsoft [94], GIT simplifies the architecture of vision-language models by using a single image encoder and a single text decoder. This streamlined design contrasts with the complex structures of previous models that typically involve multiple encoders and decoders along with external modules like object detectors and OCR systems [56, 69].

5.5.1 Model Architecture

Simplified Structure

GIT's architecture consists of:

Single Image Encoder: A Swin-like vision transformer is used as the image encoder, pre-trained on a large dataset of image-text pairs using a contrastive task to learn robust image representations.

Single Text Decoder: The text decoder is a transformer network responsible for generating text descriptions from image features. This unification of various vision-language tasks into a single language modeling task simplifies the design and training process.

5.5.2 Pre-training Approach

Dataset GIT was pre-trained on a massive dataset of 0.8 billion image-text pairs. This dataset includes:

- COCO (Common Objects in Context)

- Conceptual Captions (CC3M)
- SBU Captions
- Visual Genome
- Conceptual Captions (CC12M)
- ALT200M, plus an additional 0.6 billion web-crawled image-text pairs

Training Objectives

Language Modeling (LM) Loss: The primary task during pre-training is to map the input image to its associated text description using a language modeling objective.

Contrastive Pre-training for Image Encoder: The image encoder is pre-trained with a contrastive task to learn strong image representations.

5.5.3 Fine-tuning for Specific Tasks

Image Captioning

The same LM task used during pre-training is applied to fine-tune the model on image captioning datasets.

Visual Question Answering (VQA)

During fine-tuning, the question is treated as a prefix to the text description, and the model generates answers in an auto-regressive manner.

Video Captioning and Question Answering

For video tasks, multiple frames are sampled and encoded independently. These frame features are concatenated and passed through the text decoder.

5.5.4 Model Scaling and Performance Optimization

Scaling Up

Larger Models and Data: GIT scales up the pre-training data and the model size to improve performance significantly. Larger datasets and more extensive model architectures have shown to boost performance across various benchmarks.

Comparison with Prior Models: GIT achieves new state-of-the-art results, surpassing previous models by a large margin on several tasks.

5.5.5 Evaluation and Benchmarking

Benchmarks GIT was evaluated on a wide range of benchmarks for image captioning (COCO, nocaps, VizWiz-Caption, TextCaps), VQA (VQAv2, TextVQA, VizWiz-VQA), and video tasks (MSVD, MSRVTT, VATEX).

State-of-the-Art Performance

GIT outperformed previous models on many of these benchmarks, including surpassing human performance on TextCaps with a CIDEr score of 138.2 compared to the human score of 125.5.

5.5.6 Technical Approach

Image Encoder

The image encoder uses a Swin-like vision transformer pre-trained on image-text pairs using a contrastive task, which helps in learning robust image representations without relying on object detectors.

Text Decoder

The text decoder employs multiple transformer blocks with self-attention and feed-forward layers to generate text descriptions from image features.

Pre-training Task

The model is trained to map input images to their associated text descriptions using a language modeling objective.

5.5.7 Key Contributions and Innovations

Unified Architecture

GIT's architecture is simplified to just one image encoder and one text decoder, which contrasts with the complex structures of previous models that often involve multiple encoders and decoders.

End-to-End Training

The entire network is trained end-to-end on large-scale datasets, enabling it to achieve strong performance across various vision-language tasks.

Generation-Based Classification

GIT introduces a new scheme for image classification, where the model generates class labels in an auto-regressive way rather than relying on predefined vocabularies.

5.5.8 Challenges and Future Work

Generative Challenges

Generative models, like GIT, have inherent challenges, such as controlling the generated captions and performing in-context learning without parameter updates.

Societal Impact

GIT aims to improve accessibility for visually impaired individuals, but care must be taken to manage the potential for toxic language in pre-training datasets.

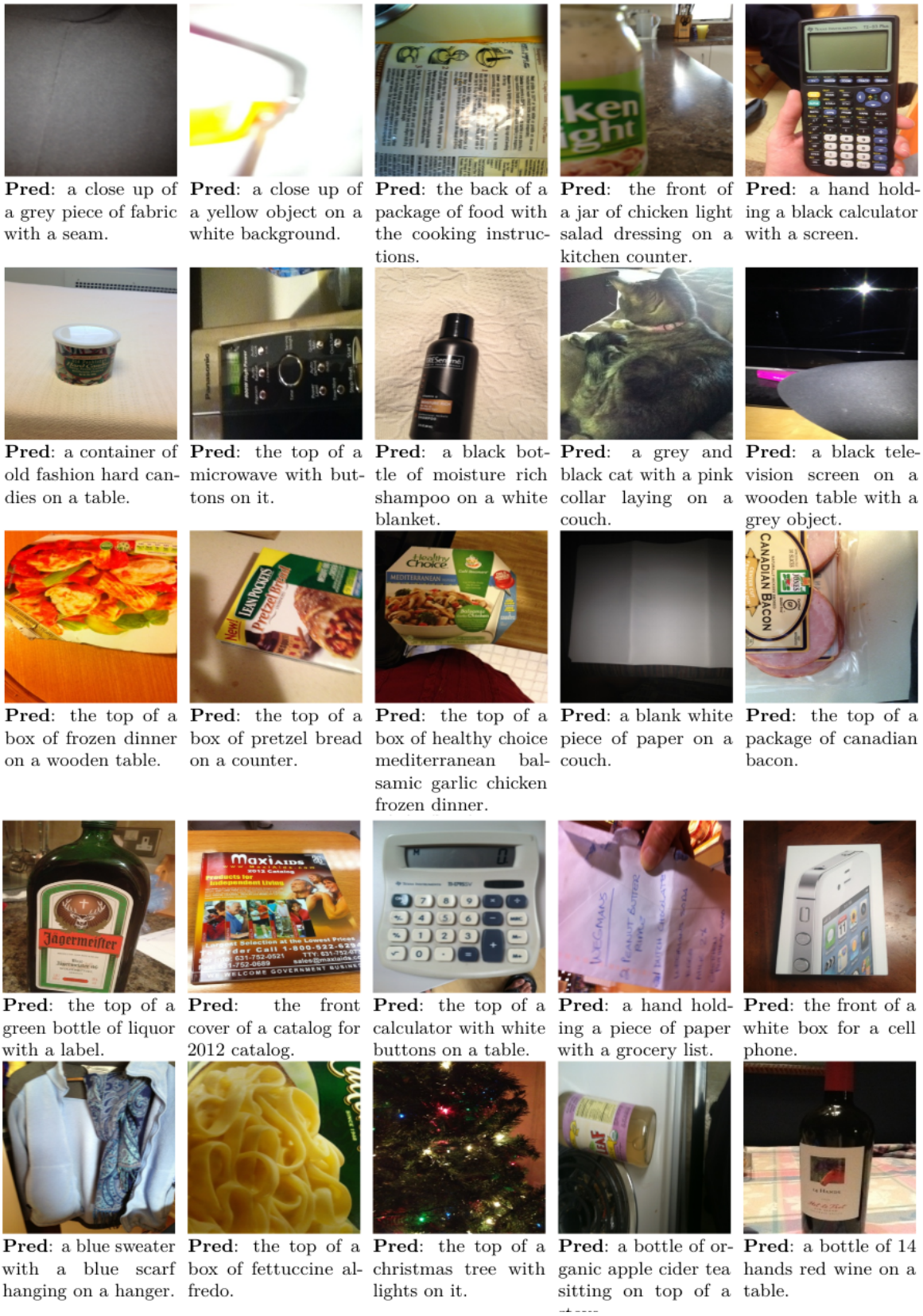


Figure 5.5.1: Visualization of the GIT model on random test images of VizWiz-Captions [94]

5.5.9 Why it was chosen

We chose to use the GIT model in our thesis as it is reasonably fast and its accuracy in describing details present in the images is of great value. The analysis of image **??** is an example of its dominance over other models we tested.

Figure 5.5.2: An example of a thumbnail image from our dataset

The GIT model produces the caption "a man with a cast on his leg playing basketball", which is very descriptive and accurate. Another model we experimented with, Blip [53], generates the caption "a man standing on a basketball court with a ball", which is again quite good, but fails to notice the cast on the man's leg. We also tested the vit-gpt2 model [61], which generated the caption "a man holding a tennis racket on a tennis court", which is not accurate. From this enlightening example, it is obvious that GIT tests better in general.

Chapter 6

Semantic Similarity

Semantic similarity, also referred to as semantic distance, measures the degree to which two words or concepts are related in meaning. This foundational concept in natural language processing (NLP) is essential for various tasks, including information retrieval, text classification, and question answering. By quantifying how closely related different pieces of text are, semantic similarity enables machines to interpret and process human language more effectively. This chapter explores the core principles of semantic similarity, delves into traditional and modern methods for measuring it, and highlights the challenges and advancements in this field. Through the use of embeddings, particularly, we examine how semantic relationships can be captured and utilized in computational models, enhancing the capabilities of NLP applications.

Contents

6.1 Introduction

Semantic similarity, also known as semantic distance, measures how closely related two words or concepts are in meaning. It is a fundamental concept in natural language processing (NLP) that has broad applications in information retrieval, text classification, and other NLP tasks. This chapter delves into the general concept of semantic distance and explores embeddings, a powerful method to quantify and utilize semantic distance in computational models.

6.2 Semantic Similarity: An Overview

Semantic similarity is crucial for understanding and processing human language computationally. It provides a way to quantify the similarity between words, phrases, or larger text units, enabling machines to mimic human-like understanding of language.

Traditional Approaches: Early methods for calculating semantic distance relied on structured lexical databases such as WordNet [60]. These databases organize words into sets of synonyms (synsets) and arrange these synsets into a hierarchical structure. Semantic similarity between two words can be computed based on the distance between their synsets in this hierarchy. For instance, Resnik [76] proposed a measure based on the information content of the common ancestor in the hierarchy.

Challenges: Despite their utility, traditional methods have limitations. They depend heavily on the comprehensiveness and accuracy of the lexical database, which may not cover all possible words or usages. Additionally, these methods are static and do not capture context-dependent meanings of words.

Applications: Semantic distance is used in various applications, including:

- Information Retrieval: Improving search engines by retrieving documents that are semantically related to the query.
- Text Classification: Grouping documents into categories based on their content.
- Question Answering: Finding the most relevant answers by understanding the semantic similarity between questions and potential answers.

6.3 Embeddings: Capturing Semantic Distance

Embeddings are a significant advancement in NLP, providing a dense, continuous representation of words that capture their semantic meanings based on context. They address many limitations of traditional methods by learning from large text corpora and adapting to context-specific nuances.

6.3.1 What Are Embeddings?

Embeddings are vector representations of words or phrases in a continuous vector space. Unlike traditional representations that might use one-hot encoding (where each word is represented as a binary vector with only one position marked as 1 and all others as 0), embeddings represent words as dense vectors of real numbers. These vectors capture semantic relationships between words, allowing for more nuanced analysis and manipulation.

The central idea behind embeddings is the distributional hypothesis, which states that words that appear in similar contexts tend to have similar meanings (Harris, 1954). By analyzing large corpora of text, embedding models learn to place words with similar meanings close to each other in the vector space.

6.3.2 How Are Embeddings Created?

Several methods have been developed to create embeddings, including:

- Word2Vec: This model, introduced by Mikolov et al. [59], uses neural networks to learn word vectors. It has two main architectures:

1. Continuous Bag of Words (CBOW): Predicts a target word from a surrounding context.
2. Skip-Gram: Predicts surrounding context words from a target word.

Both architectures aim to maximize the likelihood of context words given a target word, thereby capturing the semantic relationships between words.

- GloVe (Global Vectors for Word Representation): Developed by Pennington et al. [73], GloVe creates embeddings by analyzing global word-word co-occurrence statistics from a corpus. It constructs a large co-occurrence matrix and then applies matrix factorization techniques to generate word vectors that capture both local and global semantic information.
- Contextual Embeddings: Unlike static embeddings like Word2Vec and GloVe, contextual embeddings change based on the word's usage in different sentences. Models like BERT (Bidirectional Encoder Representations from Transformers) generate these embeddings by considering the entire context in which a word appears [22]. BERT uses a transformer architecture to understand the context bidirectionally, providing more accurate and context-aware word representations.

6.3.3 Why Use Embeddings?

Embeddings offer several advantages over traditional methods:

- Dimensionality Reduction: Dense vectors reduce the dimensionality of the word representation, making computations more efficient.
- Semantic Richness: They capture complex semantic relationships between words, such as synonyms, antonyms, and analogies.
- Context Awareness: Contextual embeddings adapt to different meanings of a word based on its usage, providing more accurate representations.

6.3.4 Applications of Embeddings

- Information Retrieval: Enhances search engines to understand and retrieve documents based on semantic similarity rather than mere keyword matching.
- Text Classification: Improves the performance of classifiers by providing rich semantic representations of texts.
- Machine Translation: Facilitates more accurate translations by capturing the contextual meanings of words, leading to coherent and contextually appropriate translations.

6.4 The challenge of vanishing gradients due to the saturation zones of the cosine function

6.4.1 Introduction to the Problem

In the realm of machine learning and neural networks, the issue of vanishing gradients is a significant challenge, particularly in deep learning models. This problem occurs when the gradients used for updating the network parameters during backpropagation become exceedingly small, effectively stalling the training process. One specific context where this issue arises is in the optimization of text embeddings using the cosine function [40], [54].

6.4.2 Cosine Function in Text Embedding

Text embeddings are numerical representations of text data that capture semantic information and are crucial for various natural language processing tasks, including Semantic Textual Similarity (STS). The cosine function is commonly used to measure the similarity between these text embeddings. The cosine similarity metric is defined as:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \cdot ||\mathbf{B}||} \tag{6.4.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the embedding vectors, and θ is the angle between them. This function effectively captures the semantic similarity by measuring the cosine of the angle between two vectors in a multi-dimensional space.

6.4.3 Saturation Zones in the Cosine Function

The saturation zones of the cosine function refer to regions where the derivative (or gradient) of the function approaches zero. These zones typically occur when the angle θ is close to 0 or π radians, corresponding to cosine values near 1 or -1. In these regions, changes in the input vectors produce minimal changes in the cosine similarity value, leading to very small gradients. Mathematically, this can be visualized as:

$$\text{Gradient of } \cos(\theta) \approx 0 \text{ when } \cos(\theta) \approx \pm 1 \tag{6.4.2}$$

6.4.4 Impact on Optimization

When the gradients are near zero, the update steps for the parameters during backpropagation become extremely small. This phenomenon significantly hinders the learning process, as the model's parameters get updated at an exceedingly slow rate, if at all. Consequently, the model may struggle to converge to an optimal solution, leading to subpar performance in tasks that depend on fine-tuned embeddings.

6.5 Angle-Optimized Text Embeddings

High-quality text embeddings are crucial for improving semantic textual similarity (STS) tasks, which are essential components in large language model (LLM) applications. Traditional text embedding models often face challenges with vanishing gradients due to their reliance on the cosine function in the optimization objective, which has saturation zones that impede gradient flow and hinder the optimization process. To address this issue, a novel angle-optimized text embedding model called AnglE has been proposed by Li and Li [54].

6.5.1 Core Idea

The core idea of AnglE is to introduce angle optimization in a complex space. This approach effectively mitigates the adverse effects of the saturation zone in the cosine function, enhancing the gradient flow and improving the optimization process. By optimizing the angle difference between embeddings, AnglE ensures more robust and effective training.

6.5.2 Methodology

Input Layer: The model first applies padding to ensure a consistent length for input sentences and maps each word to a continuous d-dimensional space to produce word embeddings. These embeddings are concatenated to form the model input, which is then passed through an encoder such as BERT, RoBERTa, or LLaMA to obtain contextual representations.

Cosine Objective: The cosine objective function is used to measure the pairwise semantic similarity between representations. It aims to maximize the cosine similarity for high similarity pairs and minimize it for low similarity pairs. However, due to the saturation zones of the cosine function, this approach can lead to vanishing gradients.

In-Batch Negative Objective: To further improve performance, the model integrates an in-batch negative objective. This involves identifying and assigning positive samples within a batch, reducing potential noise from incorrectly labeled negatives and enhancing the generalization of the model.

Angle Objective: To counteract the limitations of the cosine function, AnglE introduces an angle objective. This involves dividing the text embedding into real and imaginary parts in a complex space and computing the

angle difference between these parts. By normalizing and optimizing this angle difference, AnglE effectively mitigates the saturation zone issues of the cosine function, leading to better gradient flow and improved optimization.

6.5.3 Evaluation

Extensive experiments on various tasks, including short-text STS, long-text STS, and domain-specific STS tasks, demonstrate that AnglE outperforms state-of-the-art STS models that ignore the cosine saturation zone.

Performance on STS Tasks

Short-Text STS: AnglE was tested against several state-of-the-art (SOTA) models using short-text STS datasets. The results showed that AnglE outperformed other models significantly. For example, AnglE-BERT achieved an average Spearman correlation of 73.55

Long-Text STS: A new dataset, collected from GitHub Issues, was introduced to evaluate the model's performance on long-text STS tasks. AnglE demonstrated superior performance on this dataset as well, particularly with the AnglE-RAN variant, which performed better in long-text scenarios than AnglE-BERT. This highlights the model's ability to handle long and complex texts effectively, which is crucial for real-world applications like legal documents and technical reports.

Domain-Specific and LLM-Supervised Learning

The study also evaluated AnglE's performance in domain-specific scenarios with limited labeled data. The results were promising, showing that AnglE could still perform well even with less annotated data. The introduction of LLM-supervised learning, where large language models are used as annotators, further improved AnglE's performance. The ensemble of LLMs yielded the best results, indicating the robustness and effectiveness of this approach in overcoming data scarcity issues.

Ablation Study

An ablation study was conducted to understand the contributions of different components of the AnglE model. The results revealed that the angle optimization component was critical for the model's superior performance. When the angle objective was removed, there was a more significant drop in performance compared to the removal of the in-batch negative (ibn) objective. This emphasizes the importance of optimizing angle differences in a complex space to mitigate the saturation zone issues of the cosine function.

Transfer and Non-Transfer Tasks

AnglE's performance was evaluated in both transfer and non-transfer settings:

Transfer Tasks: In transfer tasks, AnglE was trained on NLI datasets (like MNLI and SNLI) and evaluated on various STS benchmarks. AnglE-BERT and AnglE-LLaMA consistently outperformed previous SOTA models such as SimCSE-BERT and SimCSE-LLaMA, with gains of 0.80

Non-Transfer Tasks: In non-transfer tasks, AnglE outperformed models like SimCSE and SBERT even when trained on smaller datasets. This indicates that AnglE is highly effective in learning high-quality embeddings within the same dataset context.

These results highlight the effectiveness of angle optimization in generating high-quality text embeddings and its practical utility in diverse NLP scenarios.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICR-R	Avg.
Unsupervised Models								
GloVe (avg.) †	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT-flow ‡	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening ‡	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-BERT ‡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT ‡	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT-BERT	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
DiffCSE-BERT	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
SimCSE-BERT	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
LLaMA2-7B ⋆	50.66	73.32	62.76	67.00	70.98	63.28	67.40	65.06
Supervised Models								
InferSent-GloVe †	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE †	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
ConSERT-BERT	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
CoSENT-BERT ⋆	71.35	77.52	75.05	79.68	76.05	78.99	71.19	75.69
SBERT †	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE-BERT	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
SimCSE-LLaMA2-7B ⋆	78.39	89.95	84.80	88.50	86.04	87.86	81.11	85.24
AnglE-BERT	75.09	85.56	80.66	86.44	82.47	85.16	81.23	82.37
AnglE-LLaMA2-7B	**79.00**	**90.56**	**85.79**	**89.43**	**87.00**	**88.97**	**80.94**	**85.96**

Figure 6.5.1: Evaluation results for transfer STS tasks [54]

Model	MRPC	STS-B	QQP	QNLI	GitHub Issues.	Avg.
	test	test	validation	validation	test	
SimCSE-BERT	48.13	76.27	65.84	33.00	60.38	56.72
SBERT	46.19	84.67	73.80	65.98	69.50	68.03
AnglE-RAN	58.70	80.23	74.87	63.04	**71.25**	69.62
AnglE-BERT	**62.20**	**86.26**	**76.54**	**72.19**	70.55	**73.55**

Figure 6.5.2: Evaluation results for non-transfer STS tasks [54]

Chapter 7

Counterfactual Explanations

Counterfactual explanations are an essential technique within interpretable AI, providing insights by illustrating hypothetical scenarios that highlight how slight modifications to the input data could alter a model's decision. This chapter delves into the differences between AI interpretability and explainability, emphasizing the need for transparent and comprehensible AI systems. Counterfactual explanations offer a user-friendly approach to understanding AI decisions by answering "what if" questions, which helps users grasp the minimal changes required to achieve different outcomes. The chapter further explores the challenges and advancements in creating effective counterfactuals, including ensuring realism and actionability. Additionally, it introduces the concept of using conceptual edits for generating counterfactual explanations, which leverages external knowledge graphs to provide more meaningful and intuitive explanations. Through practical applications in finance and healthcare, the chapter demonstrates the growing importance and utility of counterfactual explanations in enhancing the transparency and trustworthiness of AI models.

Contents

7.1 Artificial Intelligence Interpretability and Explainability

Often the terms AI interpretability and AI explainability are used interchangeably. However, there are significant differences between the two.

7.1.1 AI Interpretability

AI interpretability, involves understanding and making sense of the behavior of machine learning models. It refers to the extent to which a human can comprehend the cause-and-effect relationships within a model. Interpretability is about building models that are inherently understandable, either through their simplicity or through methods that allow humans to gain insights into how the model works. Interpretability emphasizes model design and selection, preferring models that are straightforward and interpretable by nature, such as linear regression, decision trees, or rule-based models. This is important for scenarios where users need to trust and understand the model thoroughly [34].

In AI interpretability the challenge lies in balancing model complexity and interpretability. While simpler models are easier to interpret, they may not capture complex patterns in the data as effectively as more complex models [48].

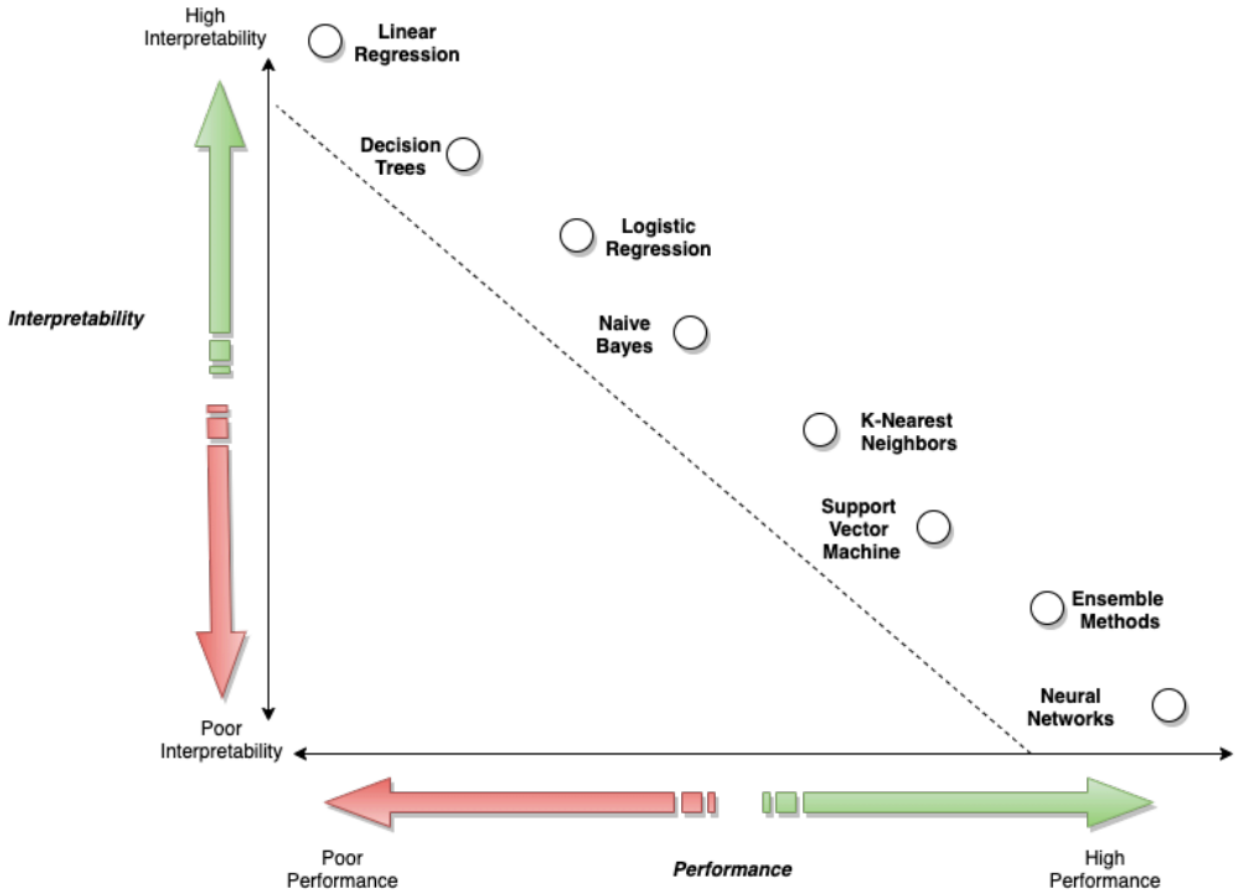


Figure 7.1.1: Interpretability versus performance trade-off given common ML algorithms [48]

7.1.2 AI Explainability

AI explainability refers to the ability to describe the internal mechanisms of a machine learning model in a way that is understandable to humans. It focuses on making the decision-making processes of AI models transparent and comprehensible by providing clear, logical explanations of how specific outputs are derived from given inputs. Explainability is often associated with complex, "black-box" models like deep neural

networks. The goal is to explain how these complex models arrive at specific decisions. It is crucial for ensuring accountability, trust, and regulatory compliance in AI systems, which are demanded for high-stakes decisions in fields like healthcare, finance, and legal systems. To achieve explainability post-hoc methods such as feature importance scores, partial dependence plots, and counterfactual explanations are used to elucidate the model's behavior after it has been trained.

One of the main challenges of AI explainability is ensuring that the explanations are accurate and meaningful without oversimplifying the model's behavior. Explainability is crucial for gaining user trust and meeting regulatory requirements [48].

7.2 Counterfactual Explanations

Counterfactual explanations are a key technique within the field of interpretable AI. They provide insights by presenting hypothetical scenarios that illustrate how slight changes to the input data could lead to different outcomes. Essentially, they answer "what if" questions, helping users understand the minimal changes required to alter a model's decision [62].

For example, a counterfactual explanation might show that if a loan applicant's income were slightly higher, their application would be approved instead of rejected. This type of explanation is intuitive and aligns with human reasoning, making it easier for users to comprehend and trust AI systems [62].

Counterfactual explanations enhance transparency by demystifying the decision-making process of AI models, turning them from opaque "black boxes" into more transparent systems that users can understand and interact with more confidently. They are particularly valuable in scenarios where users need to know not just the decision but the factors influencing it and how changes in those factors could alter the outcome.

7.2.1 Challenges and Advancements

Creating effective counterfactual explanations involves several challenges. These include ensuring that the counterfactuals are realistic and actionable, maintaining minimal changes to avoid confusion, and addressing computational complexities associated with generating these explanations [90].

Recent advancements have focused on improving the feasibility and relevance of counterfactuals. For instance, techniques such as using Generative Adversarial Networks (GANs) to create realistic counterfactual images have been explored. Moreover, the field is working towards unifying terminology and developing methods applicable across different types of machine learning tasks, including regression and classification.

7.2.2 Real-World Applications

Counterfactual explanations are being applied across various sectors to enhance the interpretability of AI systems. In finance, they help explain credit decisions, providing transparency and facilitating regulatory compliance [36]. In healthcare, they assist in understanding diagnostic models, thereby improving patient trust and outcomes. These explanations are becoming a de-facto standard for post-hoc model explanations, bridging the gap between complex AI systems and human understanding.

7.3 Conceptual Edits as Counterfactual Explanations

Filandrianos et al. [30] introduce a theoretical framework for generating counterfactual explanations using conceptual edits. Within this framework, concepts represent the generalized forms of objects found in the input data and are associated with external knowledge structured as concept hierarchies [**dimitriou2024grapheditscounterfactualexplanationslymperaiou2023counterfactual**, 20, 24]. An overview of the proposed framework is illustrated in Figure 7.3.1.

This paper addresses the "black-box" problem by providing counterfactual explanations, which clarify how a model's decision could be altered. The proposed method leverages external knowledge graphs to provide more meaningful explanations.

7.3.1 Framework overview

The framework identifies minimal concept edits that change the prediction of a black-box classifier to a desired class. By accumulating multiple counterfactual explanations, it can also estimate a "global" explanation for a dataset region and a target class.

Here's an in-depth look at the framework's core components and their functionality:

Explanation Dataset

The framework requires an explanation dataset, which is a set of tuples (x_i, C_i). Each tuple consists of a sample x_i that the classifier can process and a semantic description C_i, which is a set of concepts describing the sample. For example, in an image classification task, x_i could be an image, and C_i could be a set of labels describing objects within the image.

Conceptual Distance

The framework introduces the concept of a TBox, a terminology box that organizes concepts into a hierarchy. The TBox graph represents these concepts and their relationships. Conceptual distance (d_T) is defined as the shortest path between two concepts in the TBox graph. This distance helps determine how closely related two concepts are.

Concept Set Edits and Edit Distance

A concept set edit involves operations like replacing, deleting, or inserting concepts within a set C_i. Concept set edit distance measures the minimal cost of transforming one set of concepts into another using these edits. This distance is crucial for generating meaningful counterfactuals that reflect small yet significant changes.

Significance of Transformation

The significance of transforming one sample into another is quantified as

$$\sigma(a, b) = \frac{|F(x_a) - F(x_b)|}{D_T(C_a, C_b)} \tag{7.3.1}$$

where F is the classifier, $a = (x_a, C_a), b = (x_b, C_b)$ and D_T is the concept set edit distance. High significance indicates that small conceptual changes lead to significant changes in the classifier's output, making it a useful metric for generating impactful counterfactual explanations.

Graph Construction

The framework constructs a directed graph where each node represents a sample from the explanation dataset, and edges represent possible transformations between samples, weighted by the inverse of their significance. Dijkstra's algorithm is used to find the shortest path in this graph, providing the optimal counterfactual explanation by identifying the sequence of minimal conceptual edits needed to change the classifier's prediction.

Local and Generalized Counterfactual Explanations

Local Counterfactual Explanations: These provide specific changes needed for a single sample to achieve a desired classification. It focuses on the minimal conceptual edits necessary for this change.

Generalized Counterfactual Explanations: These extend the local explanations by aggregating them to offer insights into common changes required across a subset of the dataset. This helps understand broader patterns and common biases in the classifier.

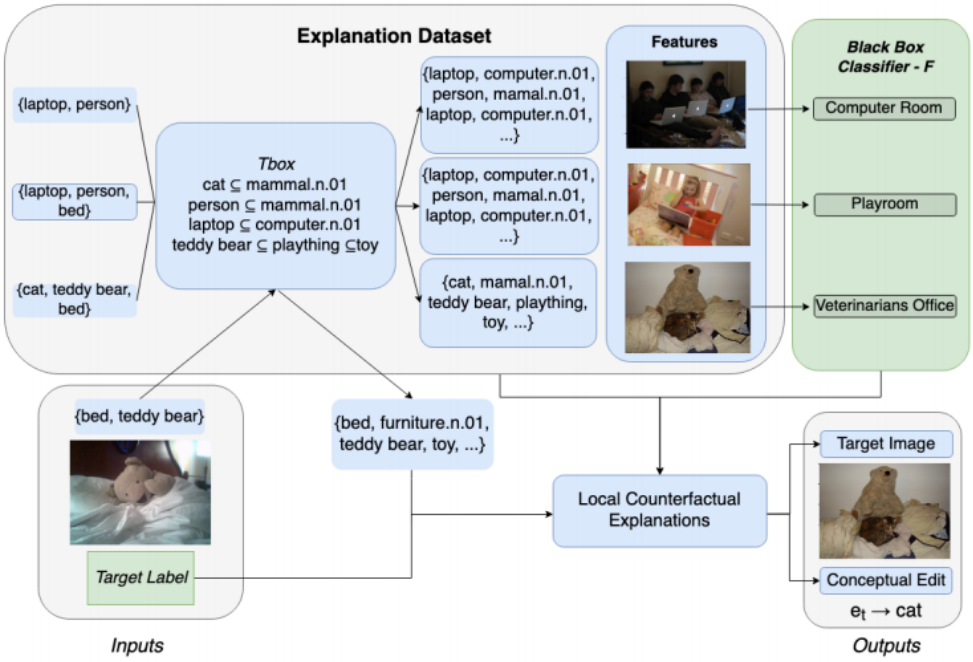


Figure 7.3.1: Conceptual Edits as Counterfactual Explanations framework [30].

7.3.2 How the counterfactual explanations are generated

In the "Conceptual Edits as Counterfactual Explanations" framework, counterfactuals are generated through a series of methodical steps that integrate conceptual understanding with machine learning. Here is an in-depth explanation of the process:

Concept Distance Calculation

Concept Distance is a measure of how far apart two concepts are within a given taxonomy (TBox). The distance is determined using the shortest path in an undirected graph representation of the TBox. For instance, in a TBox where "Cat" is a subclass of "Mammal" and "Mammal" is a subclass of "Animal," the distance between "Cat" and "Dog" (another subclass of "Mammal") is 2, traversing through "Mammal."

Concept Set Edit Distance

*Concept Set Edits:*Transformations involve replacing, deleting, or inserting concepts in a set. Each operation has a cost based on the concept distance. For example, replacing "Cat" with "Dog" has a cost equal to the distance between these two concepts.

Edit Distance Calculation: This involves computing the minimum cost required to transform one set of concepts into another. It uses a bipartite graph to match concepts from two sets and computes the minimum weight full matching using Karp's algorithm [46] to determine the cost.

Graph Construction

Explanation Graph: Each element in the explanation dataset is represented as a node. Edges between nodes represent the possible transformations with weights based on the inverse of the significance of the transformation, defined as the change in classifier output divided by the concept set edit distance.

Algorithm Implementation: The explanation graph is constructed by iterating through pairs of elements in the dataset, calculating their concept set edit distances and associated transformation significances.

Finding Counterfactual Explanations

Local Counterfactual Explanations: The goal is to find the shortest path in the explanation graph from a given sample to any sample in the desired class. This path represents the sequence of edits needed to change the classification of the sample.

Generalized Counterfactual Explanations: These are statistical summaries of multiple local counterfactual explanations within a specified region of the dataset. They highlight the most frequent edits that lead to a change in classification, providing broader insights into the classifier's behavior.

7.3.3 Detailed Analysis of Results

Experiments with CLEVR-Hans3

Objective: To detect biases in a classifier trained on the CLEVR-Hans3 dataset, where the training set contains biases (e.g., all Large Cubes are Grey).

Method: The authors created two explanation datasets, one from the training set to compare with the FACE algorithm, and another from the test set to detect training biases.

Findings: The proposed counterfactual algorithm effectively identified the biases, consistently suggesting the addition of a "grey cube" to change classifications to the biased class.

Experiments with COCO

Objective: To apply the method to a more intuitive task using the COCO dataset, which contains real-world images annotated with objects linked to external knowledge (WordNet).

Method: They used a pre-trained scene classifier and generated explanations for transitions between classes like "Bedroom" to "Kitchen" or "Veterinarian's Office."

Findings: The generalized counterfactual explanations successfully identified key concepts (e.g., adding a "cat" to classify a "Bedroom" image as a "Veterinarian's Office"). This highlighted potential biases and provided insights into the classifier's decision-making process.

Chapter 8

Graph Similarity

Graph similarity measures the structural and property-based resemblance between two graphs, playing a crucial role in various fields such as bioinformatics, social network analysis, and cheminformatics. Understanding how similar two graphs are can reveal important insights about the underlying systems they represent, whether it be protein interaction networks, social structures, or molecular compounds. Various methods are employed to measure graph similarity, including graph isomorphism, graph edit distance, subgraph isomorphism, spectral methods, graph kernels, and embedding-based techniques. Each method offers unique advantages and is suitable for different applications, ranging from exact structural matches to more flexible and scalable approximate comparisons. This chapter explores these key concepts and methods, their applications, and the specific challenges and solutions associated with computing graph similarity.

Contents

8.1 Introduction

Graph similarity is a measure of how alike two graphs are in terms of their structure and properties. This concept is crucial in various domains, including bioinformatics, social network analysis, and cheminformatics, where comparing complex networks is a common task.

8.2 Key Concepts and Methods

8.2.1 Graph Isomorphism

Two graphs G_1 and G_2 are isomorphic if there exists a bijection between their vertex sets that preserves adjacency. Essentially, G_1 can be transformed into G_2 simply by renaming its vertices [26]. Checking graph isomorphism is computationally challenging and is not suitable for large or slightly differing graphs [3].

8.2.2 Graph Edit Distance (GED)

GED measures the minimum number of edit operations (insertions, deletions, substitutions of vertices or edges) required to transform one graph into another. It provides a flexible way to measure similarity, allowing for varying degrees of difference [32]. Used in pattern recognition and computer vision to compare shapes and structures.

8.2.3 Subgraph Isomorphism

Determines if one graph (subgraph) is contained within another as an exact match. This is more stringent than graph isomorphism and is computationally NP-complete [88]. Common in cheminformatics for identifying common substructures in molecules.

8.2.4 Spectral Methods

These methods compare the spectra (eigenvalues) of matrices associated with graphs, such as the adjacency matrix or Laplacian matrix. Similar spectra indicate similar structural properties [16]. Spectral methods can handle large graphs efficiently and are useful for approximate similarity measures.

8.2.5 Graph Kernels

Graph kernels map graphs into a high-dimensional space where their similarity can be computed using kernel functions. Popular graph kernels include the Weisfeiler-Lehman kernel and the shortest-path kernel [92]. Widely used in machine learning for graph-based data, such as molecular activity prediction.

8.2.6 Embedding-Based Methods

These methods represent graphs as vectors in a continuous vector space using graph embedding techniques like node2vec, DeepWalk, and graph convolutional networks [38, 49]. They enable efficient similarity computation and are scalable to large datasets.

8.3 Applications and Importance

Graph similarity is pivotal in numerous applications:

- **Bioinformatics**: Comparing protein-protein interaction networks or genetic regulatory networks to identify functional similarities [78].
- **Social Network Analysis**: Understanding community structure and evolution by comparing social graphs [5].
- **Cheminformatics**: Identifying similar chemical compounds by comparing molecular graphs [75].

- **Information Retrieval**: Enhancing search engines by comparing document structures represented as graphs [7].

8.4 Graph Edit Distance

Graph Edit Distance (GED) is a versatile and widely used measure of similarity between two graphs. It quantifies the dissimilarity by computing the minimum number of edit operations required to transform one graph into another. These operations typically include the insertion, deletion, or substitution of vertices and edges. GED is particularly useful in applications where structural differences and similarities need to be rigorously analyzed, such as in pattern recognition, bioinformatics, and cheminformatics.

8.4.1 Edit Operations

- **Vertex Insertion**: Adding a new vertex to the graph.
- **Vertex Deletion**: Removing an existing vertex from the graph.
- **Vertex Substitution**: Replacing a vertex with another vertex.
- **Edge Insertion**: Adding a new edge between two vertices.
- **Edge Deletion**: Removing an existing edge.
- **Edge Substitution**: Replacing an edge with another edge connecting different vertices.

Each of these operations is associated with a specific cost, and the total cost of transforming one graph into another is the sum of the costs of the individual operations. The general formal definition of GED, based on the paper that introduced it [77], can be seen in 8.4.1. The Graph edit Distance between the graph pair G_1 and G_2 is denoted as $GED(G_1, G_2)$, the edit operations are e_i, their costs are $c(e_i)$ and $P(G_1, G_2)$ denotes the set of edit paths transforming the first graph to an isomorphic of the second.

$$GED(G_1, G_2) = \min_{(e_1,\ldots,e_k)\in P(G_1,G_2)} \sum_{i=1}^{k} c(e_i) \tag{8.4.1}$$

8.4.2 Cost Function

The cost function c defines the cost of each edit operation. This function can be adjusted depending on the specific requirements of the application. For instance, in some cases, substituting a vertex might be less costly than deleting and then inserting a new vertex.

8.4.3 Computation of GED

The process of calculating GED involves finding the sequence of edit operations that has the minimum total cost. This problem can be formulated as an optimization problem and is known to be NP-complete [32]. As a result, exact computation is feasible only for small graphs, and approximate methods are often used for larger graphs.

Accurate algorithms for GED computation typically try to reduce the cost of the edit path from one graph to the other. Pathfinding search or shortest paths are the methods used for this computation and they frequently make use of the A* search algorithm. GED belongs to the APX-hard complexity class since its approximation is likewise in a challenging class. Numerous graph edit distance approximation techniques have been proposed, most of which achieve cubic complexity. Among the most well-known ones are Hungarian [52], Hausdorff [31] and BP-Beam [66].

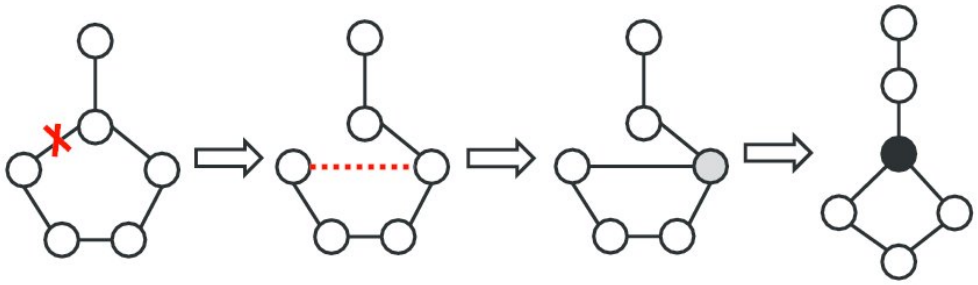


Figure 8.4.1: Graph Edit Distance Between Two Graphs. [6]

Minimum GED requires 3 edit operations and if they were all equally weighted its value would be 3.

An approximation based on converting graphs into sets and bipartite graph matching is examined in this thesis [20]. To manage the complexity of GED computation, the paper proposes simplifying the problem by converting graphs into sets of sets of concepts. The connected components of exemplars on the ABox graph are converted into sets by rolling up the roles into concepts. Specifically, new concepts of the form $\exists r.C$ are defined for each pair of role name r and concept name C. These are then added to the labels of nodes in the ABox graph. An exemplar with a connected component such as $\{Exemplar(e), depicts(e,a), depicts(e,b), depicts(e,c), Cat(a), eating(a,b), Fish(b), in(b,c), Water(c)\}$ would be represented as sets of labels like $\{\{Cat, \exists eating.Fish\}, \{Fish, \exists in.Water\}, \{Water\}\}$. After converting the connected components into sets, the problem of computing counterfactual explanations reduces to solving a set edit distance problem. This problem is solved using the framework for generating counterfactual explanations using conceptual edits described in the previous chapter.

Chapter 9

Virality Prediction of YouTube videos

The phenomenon of virality, where content rapidly spreads across digital platforms, plays a significant role in the modern digital landscape. YouTube, since its inception in 2005, has been a pivotal platform for such viral content, enabling videos to reach millions of viewers within days. This chapter delves into the mechanisms of virality on YouTube, exploring how certain videos achieve widespread popularity and the various factors influencing this process. It reviews pioneering studies that examine video virality and discusses advanced methodologies for predicting viral success. By leveraging a combination of early viewing patterns, visual attributes, and cross-platform dynamics, researchers have developed models that provide insights into what makes a video viral, offering valuable tools for content creators and marketers alike. This comprehensive exploration highlights the evolution of predictive models and the latest innovations aimed at understanding and forecasting the viral potential of YouTube videos.

Contents

9.1 Virality

Virality refers to the phenomenon where content, such as videos, images, or articles, spreads rapidly and widely across digital platforms, particularly through social media sharing and other online channels. This rapid dissemination is often driven by users sharing the content within their own networks, thereby creating a chain reaction that amplifies its reach. Virality leverages the power of social networks to generate significant engagement and visibility, often without the need for traditional advertising, making it a cost-effective strategy for increasing brand exposure and customer engagement [85].

9.2 YouTube

YouTube is a prominent online video-sharing platform that was founded in February 2005 by Chad Hurley, Steve Chen, and Jawed Karim, three former employees of PayPal. The initial idea was to create a video dating site named "Tune In, Hook Up," but it evolved into a general video-sharing platform where users could upload, share, and view videos. The domain name "YouTube.com" was activated on February 14, 2005, and the first video, titled "Me at the zoo," was uploaded by Karim on April 23, 2005 [101, 17].

The YouTube platform quickly grew in popularity, reaching over 100 million video views per day by the summer of 2006. In November 2006, Google acquired YouTube for $1.65 billion in stock, significantly boosting its resources and infrastructure [101].

In 2007, YouTube introduced the YouTube Partner Program, enabling content creators to monetize their videos, which transformed the platform into a lucrative venue for many users. Over the years, YouTube has continued to innovate with features like live streaming and subscription services such as YouTube Premium and YouTube Music [42]. Today, it serves billions of users globally, remaining a central hub for entertainment, education, and information sharing [42].

9.3 Viral videos

In the context of YouTube videos, virality refers to the rapid and widespread sharing of video content across the internet, particularly on social media platforms. This phenomenon is characterized by a significant and sudden increase in views, shares, likes, comments, and overall engagement within a short period. For a YouTube video to be considered viral, it often needs to reach hundreds of thousands or even millions of views in a matter of days or weeks [98].

Virality leverages the network effect, where the value and reach of a video expand exponentially as more people view and share it. The YouTube algorithm also plays a crucial role in promoting viral content by boosting videos that show high initial engagement, such as rapid increases in views, likes, and shares. This feedback loop can further amplify a video's reach, making it visible to an even larger audience [1].

9.4 Predicting video virality

The analysis of content popularity has garnered significant interest recently. One of the initial studies examining the factors influencing internet video popularity is presented in [96]. However, this study's limited sample size hinders its generalizability. Deza et al. [23] identified various visual attributes impacting image virality. Characteristics of viral videos were also examined in [45], where the authors examined viral videos, which gain popularity through extensive internet sharing. The study introduced the CMU Viral Video Dataset, a large open benchmark for researching viral videos. The dataset was analyzed to verify existing observations about viral videos and discover new characteristics. They also propose a model for predicting the peak view day of viral videos, which is valuable for planning advertising campaigns and viral marketing strategies. Broxton et al. [13] also analyzed YouTube viewing patterns, revealing that popular videos experience sharp spikes and declines in view counts, unlike less popular videos. More complex interdependencies between cross-platform content were explored in [89], where the authors analyzed factors that help popularize both Tweets and the YouTube videos they mention. The study collected data on tweets containing YouTube links and corresponding video metadata to identify properties that make videos both popular and viral. The authors

developed a prediction framework using features from both platforms, achieving high accuracy with minimal training data. They concluded that cross-system prediction of video popularity and virality is feasible and that features related to user engagement on YouTube are particularly effective for predicting virality.

While researchers have sought to understand content popularity factors, several methods have been developed to predict online content popularity. In [74], it was demonstrated that future views of a YouTube video could be forecasted using its early viewing patterns. This study trained a regression model to predict the total future views based on past patterns, although collecting view counts at different times posed a significant challenge. This method also doesn't consider video attributes, thus failing to correlate viewership with video quality. Using visual cues, a spatial transformer model was proposed in [25] to predict image virality. Additionally, [81] predicted online content popularity using only the title, and [15] introduced a multimodal prediction method for micro-videos, leveraging social, visual, acoustic, and textual modalities. However, the popularity definition in this model is ad-hoc, and it cannot predict view counts. A similar multimodal approach for images is discussed in [58]. Zhang et al. [103] proposed combining visual, textual, and user information in an attention model for predicting Flickr image popularity. In [86], a support vector regression model was proposed to predict online video popularity. Furthermore, a multimodal self-attention model for video popularity prediction is introduced in [10]. Finally, Kong et al. [51] introduced HIPie, an interactive visualization system designed to help users understand and predict the popularity of YouTube videos. Utilizing the Hawkes Intensity Process (HIP), a state-of-the-art model for online popularity, HIPie retrieves video metadata and historical popularity data to provide insights and forecasts. The tool allows users to identify potential viral videos, simulate the impact of promotions, and compare video popularity through various interactive visualizations. This system aims to empower both content creators and consumers by making data-driven predictions accessible and understandable.

More recent studies on predicting video virality have explored various innovative approaches. One such study by Dinko Bacic and Curt Gilstrap [4] used biometric data and machine learning to predict video viewer engagement and virality. They employed physiological data, including facial expressions and skin conductance, to create a predictive model with over 80% accuracy, highlighting the significant role of subconscious responses in understanding video popularity. Another study focused on a PyTorch implementation of ViViT to predict the virality of TikTok videos, showing the continuous advancements in deep learning applications for this purpose [39]. Another study by Wang et al. [93] presented a framework for predicting the virality and popularity of videos using minimal training data. The authors introduced a method that utilizes cross-platform dynamics, specifically between YouTube and Twitter, to accurately forecast which videos will become popular or viral. The key contribution was the ability to predict video virality with high accuracy using only a single day of training data, highlighting the framework's efficiency and effectiveness in real-world applications.

Chapter 10

Proposal

In this section, we propose the method with which we will tackle the problem of optimizing video metadata to help increase popularity and determining what makes a YouTube video go viral.

We first highlight the main contributions of this thesis and then explain the proposed method in detail.

10.1 Contributions

The contributions of this dissertation are multiple and can be summarized as follows:

- The most notable contribution of this thesis is the development of a comprehensive framework that integrates multiple advanced analytical techniques to understand and predict the virality of YouTube videos. The framework combines graph theory, sentiment analysis, image captioning, embeddings, and counterfactual explanations, each of which plays a crucial role in the holistic analysis of video content.
- Graphs for Relationship Modeling: The use of graph theory to model the relationships between various video metadata elements (such as tags, thumbnails) is innovative. Graphs provide a visual and mathematical means to understand the complex interactions and dependencies that contribute to a video's virality.
- Actionable Insights: The inclusion of counterfactual explanations provides actionable insights by showing how slight modifications to video elements (e.g., title, tags) can significantly impact viewership and engagement. This aspect enhances the practical utility of the framework for content creators.

10.2 Proposed Method

The process of transforming ordinary YouTube videos into viral sensations remains a complex and ever evolving challenge. The core of our suggested approach is the development of a customized dataset made up of YouTube video thumbnails and the textual metadata that goes with them. To enable a thorough analysis, we maintain a dataset that encompasses a range of content elements, engagement indicators, and video attributes. Then, graph representations are created from these video-related data. When it comes to expressing the complex relationships and dependencies present in video content, viewer interactions, and external factors that impact virality, graphs provide an organized and adaptable framework. For this transformation, nodes and edges that encode different video properties, viewer actions, and contextual data must be defined. The study then moves into the field of graph counterfactual algorithms, where the main goal is to identify the crucial elements that cause a video to go viral from non-viral to viral. The created graph structures are manipulated and examined using graph-based algorithms and semantic counterfactual methods. To do this, it is necessary to investigate speculative scenarios, alter graph elements, and pinpoint crucial interventions that enhance the virality of videos.

10.2.1 YouTube Trending Video Dataset

Dataset Description

In this thesis, I utilize Kaggle's YouTube Trending Video Dataset[1], which is an extensive dataset updated daily, ensuring that it remains current and reflective of the latest trends on YouTube. This dataset is curated to provide insights into trending videos on YouTube across multiple regions. It includes a wealth of information that is valuable for analyzing trends, popularity metrics, and content characteristics on the platform.

Dataset Overview

The YouTube Trending Video Dataset contains records of trending videos from various countries, including the United States, Canada, Great Britain, Germany, France, India, Japan, South Korea, Mexico, and Russia, with up to 200 listed trending videos per day. Each entry in the dataset represents a video that has appeared on the trending page of YouTube, providing detailed metadata about the video and its performance metrics. The data is collected utilizing the YouTube API and the dataset is the structurally improved version of the Trending YouTube Video Statistics dataset[2].

Attributes and Features

The dataset comprises several attributes that encapsulate various aspects of each trending video. Key features include:

- **Video ID**: A unique identifier for each video.
- **Title**: The title of the video.
- **Published Date**: The date and time when the video was published on YouTube.
- **Channel ID**: The unique ID of the channel that published the video.
- **Channel Title**: The name of the channel that published the video.
- **Category ID**: An identifier for the category to which the video belongs.
- **Trending Date**: The date when the video was recorded as trending.
- **Tags**: Tags associated with the video.
- **View Count**: The number of views the video has garnered.
- **Likes**: The number of likes the video has received.
- **Dislikes**: The number of dislikes the video has received.
- **Comment Count**: The number of comments on the video.
- **Thumbnail Link**: A link to the video's thumbnail image.
- **Description**: A brief description of the video content.
- **Comments Disabled**: Indicates whether comments are disabled for the video.
- **Ratings Disabled**: Indicates whether ratings are disabled for the video.

10.2.2 Our Dataset

After making the decision on which dataset was to be used as a base to construct our own viral video dataset, the challenge was to decide which characteristics and metadata should be kept, analyzed further, or ignored. These vital choices were made while considering the impact they would have on our goal. Below we explain the reasons why we preserved or not each feature:

[1] *https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data?select=US_youtube_trending_data.csv*

[2] *https://www.kaggle.com/datasets/datasnaek/youtube-new*

- **Video ID**: The column containing the video ID is preserved in order to identify each video uniquely.
- **Title**: The column containing the title of the video is preserved, as the title is a significant contributor to the choice of whether to click on a video. It is one of the first things one can see when encountering a new video on the platform. In addition, intuitively one can comprehend that a good title, with good syntax and appropriate use of punctuation and capital letters, can help persuade a viewer to click on a certain video.
- **Published Date**: The date when a video was published is relevant to this study as virality is often judged not only by the number of views a video has accumulated but also by the days in which it achieved such results.
- **Channel ID**: The ID used to uniquely identify a channel is also of no use to us, as the channel's title is a unique identifier alone.
- **Channel Title**: We did not preserve the title of the channel that shared the specific video, as, even though it is a characteristic that is visible to the viewer before the video is clicked and could possibly sway someone toward watching the video if, for example, the channel is well-known, it can not help turn another video viral. More specifically, in comparing a random video to a viral video in order to give advice for valuable changes in the random video's data to increase its popularity, it would be quite useless to suggest the title of the random video's channel be changed to the title of the popular video's channel. In addition, for unknown channels, it is reasonable to assume that a channel's title plays little to no role in whether a video will achieve virality. As a result, it would be impractical to try to extract valuable information from a video's channel title.
- **Category ID**: The category is not something visible to the viewer, however, it can easily be deduced from other information available. Moreover, it will be useful in order to perform experiments focused on specific categories. Therefore, we decided to preserve it.
- **Trending Date**: As with the Published Date, we decided to keep the feature so as to be able to measure the number of views a video had in a specific moment in time.
- **Tags**: The tags help the users come across the video when their search contains certain keywords. As a result, they greatly influence a video's chance to go viral and therefore we chose to include them.
- **View Count**: The view count is one of the main determinants of virality so this feature is maintained. In this way, we are able to measure just how viral a video is.
- **Likes & Dislikes**: The number of likes and dislikes a video has. These features were calculated after the video went viral and they are therefore impertinent to our study.
- **Comment Count**: This is not visible to the user before clicking, so it is not used in our study. Moreover, it is not a feature that could be influenced by the content creator in any way.
- **Thumbnail Link**: The thumbnail is the first thing anyone sees when encountering a new video. As a result, it is imperative that we include it in our research.
- **Comments Disabled & Ratings Disabled**: Whether or not the users can interact with the video by commenting and rating it is expected to be quite relevant to its popularity. In many viral videos the comments are overflowing, and many passionate or even heated conversations take place in the comment section. In other videos, comments and ratings are turned off and this is something possibly preferred for some cases. Moreover, turning on or off the ratings and comments in a video is managed by the content creator and is therefore something that can be changed if deemed helpful.
- **Description**: The description of the video is partially visible before clicking the video only on the search page (not on the start page, or the recommended columns). We decided against keeping it since its size and details are not generally something a user will turn to in order to choose whether they will watch a video.

It should also be mentioned that by simply examining this dataset, we perceived that there are no missing (null) values, except for the column that contains the videos' description and tags, as there are some videos

that do not have one or both of them. We also noticed that several videos are included more than once, with a different number of views, likes and dislikes.

This dataset contains over 240000 entries, including duplicates. The unique entries are around 43000.

The numerical data was not altered, however, the categorical data (title and tags) as well as the thumbnail images needed to be analyzed and processed so that meaningful and mostly comparable information could be extracted. Below we elaborate on the analysis of each one of these features:

Title

The title of a video is one of its most important characteristics and a lot of information can be extracted from it. More specifically, our analysis includes:

- **Keyword Extraction:** We use a specific function to erase all stop words as well as punctuation marks from the title. The remaining words are turned into lowercase ones and assembled in a list. This is done to facilitate theme comparison between titles.
- **Sentiment analysis:** We use the VADER model introduced in chapter 4 to get the sentiment analysis of the full title. From the results, we only keep the compound score in order to have a single descriptive score. The title's sentiment is undeniably important and affects how the user perceives it.
- **Punctuation marks:** We count the number of specific punctuation marks included in the title. Specifically, we search for six punctuation marks: full stop, exclamation mark, question mark, quotation mark, hyphen, vertical line. These six marks were chosen among all punctuation marks firstly because it would be too computationally intensive to account for all existing punctuation marks and secondly because they are considered the most eye-catching ones that alter the effect of the title and influence human perception of it.
- **Length:** We count the number of words the title consists of as the length of the title could be a factor affecting its virality status.
- **First letter capitalization:** We check whether the first letter for each sentence included in the title is capitalized. A feature is added corresponding to 1 if the first letter of each sentence is capitalized and 0 if not.
- **Capital words ratio:** We calculate the number of words written in capital letters in the title and divide it by the total number of words in the title.

Thumbnail

The thumbnail of a video is probably the first feature anyone sees upon encountering it. In order to get useful information from the thumbnail link provided in the dataset we have to load each image and use an image captioner to inspect them. Consequently, we use the GIT model to provide a caption for each image in the dataset. Afterward, we erase all stop words from the caption as well as punctuation marks and gather the remaining words in a list. The list therefore contains words describing objects, people and actions apparent in a thumbnail image.

Tags

Some videos contain a lot of tags while others contain none. We can not modify the ones that have no tags in any way, but we can adjust the ones that do. We observed that in many cases tags are repeated, or very similar tags are involved with the same video. Some tags are also subsets or subwords of other tags. It is also common that a video has a very large number of tags. Upon realizing the above, we opted to extract a set number of keywords using the Term Frequency – Inverse Document Frequency technique from the union of all tags and also keep the number of different tags included. This was imperative as the very large number of tags would make the comparison between videos overly and needlessly computationally intensive without actually having a lot to offer. Repetitive tags would not actually help in matching the videos and it was deemed better to just maintain the most common themes among tags and their total number.

10.2.3 Transformation to Knowledge Graphs

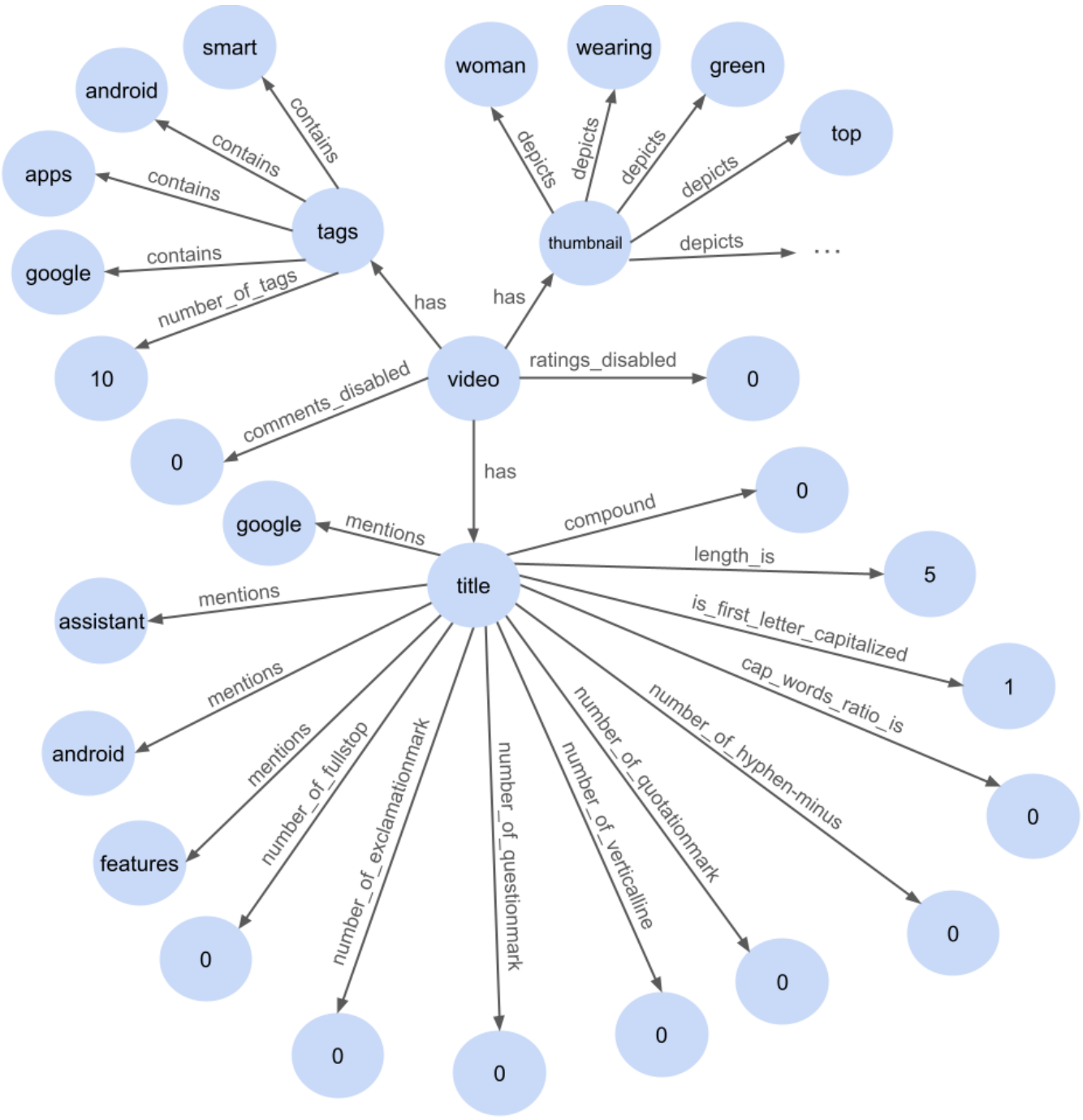


Figure 10.2.1: An example YouTube video knowledge graph

We continued by transforming the unique rows of the above dataset into knowledge graphs. We did not include the view count, the published date and the trending date in the graphs, as the graphs will be used for the comparison of videos and there is no point in comparing published and trending dates or view count. These features were only retained in order to divide the dataset into smaller datasets later on. For the representation of graphs, we used the Resource Description Framework (RDF) through python's rdflib library. Each one of the main elements of a video is a separate entity. A different URI was created for the relationships between the nodes named "properties". Keeping only the names of the relationships and the

nodes (not the whole URI) an example graph can be seen in 10.2.1.

10.2.4 Comparison

In order to produce viable suggestions for content optimization our goal is to compare a given random YouTube video graph with each and every graph in our dataset. The cost to transform this random graph into each and every graph in our dataset is calculated and the transformation with the lower cost is selected. This graph will be for the video that is the most similar one to the random input video. By examining their similarities and differences we can understand what made the video in our dataset go viral instead of the random one.

In order to compare the video graphs with one another we will use the cece library created by Filandrianos et al. [30] and analyzed in chapter 7. For this, we created our own object distance, object addition and object removal functions. The cost of exchanging categorical nodes equals the semantic distance of the words they contain, as measured using AnglE embeddings to represent the words and a scaled version of cosine similarity to calculate their distance. The cost of exchanging numerical nodes equals their numerical difference. Since all graphs contain a fixed number of nodes in addition to a volatile number of nodes, which consist of the thumbnail description nodes whose multitude differs among videos, we prohibit the exchange of different types of edges, as with the numerical nodes it could cause mistakes in the cost calculation. An example of this can be seen in 10.2.2. In this case, if trading different types of edges was allowed, the cheapest way to go from the first graph to the second would be to change the edge (title, length_is^5) into (title, number_of_exclamationmark^4) and the edge (title, number_of_exclamationmark^2) into (title, length_is^2) which would have a total cost of 1+0=1. However, the actual cost to transform the first subgraph to the other would be 3+2=5 by turning the edge (title, length_is^5) into (title, length_is^2) and the edge (title, number_of_exclamationmark^2) into (title, number_of_exclamationmark^4).

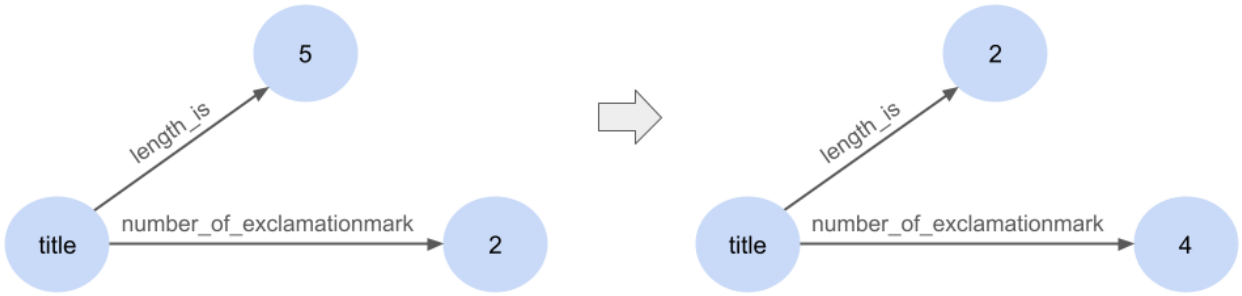


Figure 10.2.2: An example of bad cost calculation when trading of different edge types is allowed

Our first thought was to compare entire graphs, however, this approach proved overly computationally intensive. This was because the algorithm compares every node with every other node and every edge with every other edge when contrasting two graphs. This resulted in an immense number of comparisons, most of which were pointless, since we are not interested in exchanging edges of different types. The decision was then made to divide the graphs into smaller ones. Specifically, each video was now comprised of three graphs, one for its title, one for its thumbnail and one for its tags. In this way, we compare tags with tags, thumbnails with thumbnails and titles with titles. Moreover, the choice was made to not include all nodes that do not contain information about the video's theme in the search for the pair with the lowest transformation cost. This is because we aim to find the most similar pair thematically and understand why their different presentations caused one to succeed and the other not. As a result, the difference in presentations will be examined only for the most similar pair at the end of its computation. The three graphs that represent each YouTube video and take part in the search for the most similar pair in the dataset can be seen in 10.2.3, 10.2.4, 10.2.5.

Even though we originally used the cece graph interface to compute the graph edit distance for all graphs, after dividing the graphs into smaller ones, we no longer needed it. Instead, we could treat the graphs as queries instead. This is possible due to the structure of the graphs. More specifically, all three graphs that

we will compare have a node at the center with whom all other nodes are connected. In addition, there is no edge between the rest of the nodes. In other words, the graphs resemble star schemas and the center node is the same for all videos. Therefore, it makes more sense to simply create three queries using all the nodes from each graph, except the center ones and compute their semantic distance. As a result, the three graphs are finally transformed into:

title: ['google', 'assistant', 'features', 'android']

thumbnail: ['woman', 'wearing', 'green', 'striped', 'top', 'showing', 'computer', ...]

tags: ['android', 'smart', 'google', 'apps']

We create 3 datasets using cece. One for all the title queries in our dataset, one for all the thumbnail queries and one for all the tags queries. Each random input video is converted to the above form and the cost to transform its title query to every title query in the dataset is computed. The same is done for its thumbnail and tags. In the end, the costs are added for each transformation pair and the total costs to transform the random video's queries to each query in the dataset are computed. We should note here that it was engineered so the title is the most important feature to match, followed by the tags and finally the thumbnail, by adjusting the costs in the total addition. From these total costs the lowest one is selected and the corresponding video in the dataset is returned. Afterward, we use the cece library again along with our own functions to compute the differences between the two videos and suggest changes for the random input video to increase its potential for virality. The changes include changes in tags, in what the thumbnail shows, in the title's words, punctuation, capitalization etc, as we use all the information extracted so far.

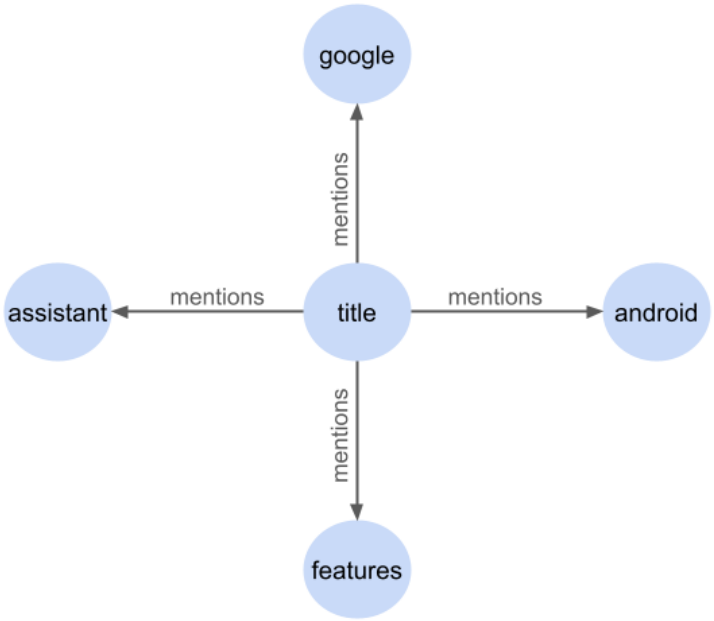


Figure 10.2.3: An example of the title's graph

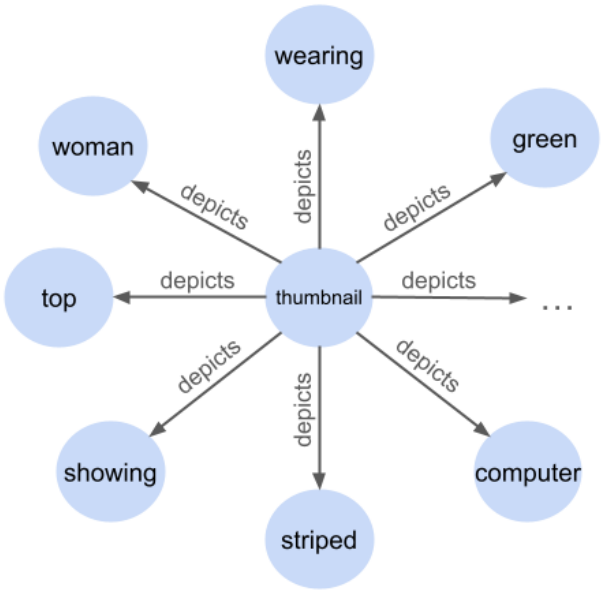


Figure 10.2.4: An example of the thumbnail's graph

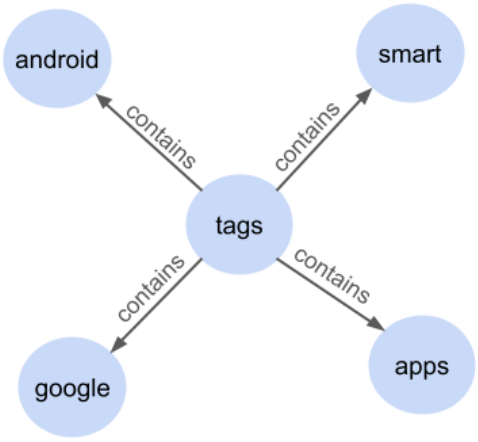


Figure 10.2.5: An example of the tags' graph

Chapter 11

Experiments

This chapter describes the tests that were carried out to determine what makes viral YouTube videos unique from those with far lower view counts. We aim to identify common patterns and characteristics that enhance video virality in different categories by utilizing a mixed dataset. Using a dataset of 4500 films with at least 5,000,000 views in a week, paired with an equal number of videos with the lowest view counts, the trials start off with a generic methodology. This significant variation in viewership offered a solid foundation for analyzing the differences and similarities in video elements. Then, in order to guarantee meaningful results, we expand our study to include particular categories including Music, Sports, Gaming, People & Blogs, and Entertainment. Each of these categories had a sizable number of samples. The results gathered from these studies are presented in comprehensive tables that include information on interaction metrics, thumbnail preferences, tag usage, and title formatting.

Contents

11.1 General Experiments

We began our experiments with a more general approach, using our mixed dataset, without dividing into different categories. This was done in an effort to identify universal tendencies and characteristics of viral videos. Since the number of views among our samples have quite big differences it was possible to use our dataset to create the testing data as well. More specifically, we extracted all videos with at least 5.000.000 views within a week to use as our dataset. This was about 4500 videos. We also extracted the 4500 videos with the least amount of views within a week to use as our testing data. For the first dataset the minimum number of views is 5.002.539, while for the second the maximum number of views is 325.458. The difference is quite big, so it makes sense to evaluate their dissimilarities.

The statistical results from our experiment our shown in tables 11.1, 11.2, 11.3.

In general, it should be highlighted that there are good and bad matches between videos made by our method. The good ones involve test videos that are concerned with common issues or popular subjects. For these videos, it is easier to find a match in our dataset, as it is more likely such subjects were discussed by viral videos too. For videos, however, that have very specific subjects it is much more difficult. Therefore, in most cases, these videos get paired up with a sample in the dataset that is not a very good parallel or close to them thematically.

Feature	**Increased**	**Decreased**	**Remains the same**
Title length	32.5%	59.9%	7.6%
Title's capital words ratio	30.2%	48.2%	21.6%
Title's compound score	36.2%	34.6%	29.2%
Number of exclamation marks in title	11.9%	20.8%	67.4%
Number of question marks in title	2%	4.7%	93.3%
Number of full stops in title	9.5%	15%	75.5%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	17.5%	21.7%	60.8%
Number of hyphens in title	27.9%	14.9%	57.2%
Number of different tags	37.3%	60.9%	1.8%

Table 11.1: Statistical results for numerical data

Feature	**True->False**	**False->True**	**True->True**	**False->False**
Ratings disabled	1%	0.8%	0%	98.2%
Comments disabled	2.7%	1.4%	0.1%	95.8%
First letter capitalization in title	12.5%	18%	66.3%	3.1%

Table 11.2: Statistical results for categorical data

It is obvious from the statistical data above that, in comparison to the videos with fewer views, viral videos have shorter titles, fewer capital words and fewer and more focused tags. When it comes to whether or not their titles have positive or negative connotation, the results are equally divided. In most cases, punctuation marks are not really used and the advice is to decrease their occurrences.

From the table 11.2 we can understand that most videos keep both comments and ratings enabled, regardless of their number of views. The same goes for the capitalization of the first letter of each sentence. It is

practiced in the majority of cases, especially in our viral videos dataset. When this is not the case, it is most likely that our algorithm will advise against it.

Top 5 thumbnail edits/additions	**Top 5 tags edits/additions**
man	video
person	game
game	trailer
movie	music
poster	diy

Table 11.3: Statistical results for thumbnails and tags

By analyzing the thumbnail edits it was found that the most common additions were "man" and "person", and also the majority of the transformations were focused on words like these as well. This indicates that showing people in the thumbnail increases a video's popularity and its chances of being clicked on. The rest of the most popular edits sourced from the largest categories of videos in the data, which we will analyze later on.

When it came to tags there were no edits often and meaningful enough to be worth mentioning. The popular edits once again originated from the abundance of samples from certain categories.

11.2 Experiments based on category

In total, there are 15 categories of videos in our dataset. These are:

- **Category ID = 1:** Film & Animation (1703 samples)
- **Category ID = 2:** Autos & Vehicles (876 samples)
- **Category ID = 10:** Music (6899 samples)
- **Category ID = 15:** Pets & Animals (190 samples)
- **Category ID = 17:** Sports (5614 samples)
- **Category ID = 19:** Travel & Events (255 samples)
- **Category ID = 20:** Gaming (8787 samples)
- **Category ID = 22:** People & Blogs (3769 samples)
- **Category ID = 23:** Comedy (2121 samples)
- **Category ID = 24:** Entertainment (8539 samples)
- **Category ID = 25:** News & Politics (1574 samples)
- **Category ID = 26:** Howto & Style (1122 samples)
- **Category ID = 27:** Education (1032 samples)
- **Category ID = 28:** Science & Technology (1319 samples)
- **Category ID = 29:** Nonprofits & Activism (19 samples)

We will test only for 5 of them: Music, Sports, Gaming, People & Blogs and Entertainment, as the rest have too few samples to be able to use for both the dataset and the tests and have a significant difference in the number of views, for the experiments to be meaningful.

11.2.1 Music

This category consists of music videos, official music video clips and lyric videos. Since it has 6899 samples, we create a dataset containing the 2000 ones with the most views. Our test dataset is comprised of the 2000 ones with the fewest views. The statistical results from this experiment are shown in the tables.

Feature	Increased	Decreased	Remains the same
Title length	44.4%	46%	9.6%
Title's capital words ratio	45.3%	24.1%	30.5%
Title's compound score	27.6%	30.1%	42.3%
Number of exclamation marks in title	1.2%	2.9%	95.9%
Number of question marks in title	0.9%	1.1%	98%
Number of full stops in title	13.3%	16.9%	69.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	5.5%	4.5%	90%
Number of hyphens in title	19.4%	19.9%	60.7%
Number of different tags	60.4%	35.8%	3.8%

Table 11.4: Statistical results for numerical data

Feature	True->False	False->True	True->True	False->False
Ratings disabled	0.6%	0.7%	0%	98.7%
Comments disabled	0.4%	0.3%	0%	99.3%
First letter capitalization in title	8.5%	8.6%	81.5%	1.4%

Table 11.5: Statistical results for categorical data

In tables 11.4, 11.5 we can see that in our viral videos' titles there is a higher ratio of words in capital to lower case in comparison to the test dataset. Again, when it comes to punctuation marks the two sets do not appear to be very different. In the majority of matches it is also advised that the number of different tags be increased. The tendency to capitalize the first letter of every sentence in the title is very high in both sets. Moreover, just like in the mixed categories scenario, comments and ratings are enabled in both sets.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
person	music
man	video
group	new
standing	bts
woman	entertainment

Table 11.6: Statistical results for thumbnails and tags

For the thumbnail, it seems apparent from the most popular edits in table 11.6 that there is a preference for

it to depict people, probably the artist or artists who created the song. In the tags column of the table, we can see the most popular and general ones contain the words "music", "video", "new" and "entertainment". However, tags containing the name of a popular boy band "BTS" are also commonly proposed. This is because there was no way to list the names of all artists and bands in order to prevent the algorithm from transforming the input names of the artists to those. This is a general problem in this category as there is really no point in suggesting that the title of a video that is a song be changed, or that the artist's name be changed. Therefore, finding meaningful matches between music videos is tricky, in the sense that the title, the thumbnail or the tags in most cases do not describe the essence of the song, its genre and overall energy and "vibe".

11.2.2 Sports

This category consists of videos concerning sports. Since it has 5614 samples, we create a dataset containing the 2000 ones with the most views. Our test dataset is comprised of the 2000 ones with the fewest views. The statistical results from this experiment are shown in the tables.

Feature	**Increased**	**Decreased**	**Remains the same**
Title length	40.6%	51.5%	7.9%
Title's capital words ratio	49.8%	31.8%	18.4%
Title's compound score	30.9%	30.2%	38.9%
Number of exclamation marks in title	7.7%	14.9%	77.3%
Number of question marks in title	1.2%	4.5%	94.2%
Number of full stops in title	20.2%	19.0%	60.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	48.5%	21.9%	29.6%
Number of hyphens in title	16.1%	16.6%	67.2%
Number of different tags	42.7%	53.9%	3.4%

Table 11.7: Statistical results for numerical data

Feature	**True->False**	**False->True**	**True->True**	**False->False**
Ratings disabled	0.3%	0.8%	0%	98.9%
Comments disabled	0.4%	0%	0%	99.6%
First letter capitalization in title	18.8%	12.8%	64.4%	4.1%

Table 11.8: Statistical results for categorical data

In tables 11.7, 11.8 we can see that in our viral videos' titles there is a higher ratio of words in capital to lower case in comparison to the test dataset. Again, when it comes to punctuation marks the two sets do not appear to be very different, apart from the vertical lines, which the answers suggest should be increase in half of the test videos. In a slightly higher percentage of matches it is also advised that the number of different tags be decreased in comparison to being increased. The tendency to capitalize the first letter of every sentence in the title is again high in both sets, though with a higher number of exceptions in this category. Moreover, just like in the mixed categories scenario, comments and ratings are enabled in both sets.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
player	game
players	league
game	basketball
football	highlights
basketball	nba

Table 11.9: Statistical results for thumbnails and tags

Concerning the thumbnail, it is again preferred to depict people, and specifically in this category, players. Images of the sport in question are also recommended, with some of the most popular transformations being toward "football" and "basketball", which are arguably the most popular sports in USA, where the data was collected. This is expected as the majority of both the viral video dataset and the test dataset will be about these two sports. For videos that discuss more obscure sports there exists the possibility they won't find a good match to be compared to. The most popular tag theme transformations are towards general sports terms, but also types of sports.

11.2.3 Gaming

This category consists of content related to video games, including but not limited to gameplay, reviews, tutorials, and industry news. It is the category with the largest number of samples (8787 samples). We create a dataset containing the 2500 ones with the most views. Our test dataset is comprised of the 2500 ones with the fewest views. The statistical results from this experiment are shown in the tables.

Feature	Increased	Decreased	Remains the same
Title length	34.8%	55.0%	10.2%
Title's capital words ratio	30.6%	44.7%	24.7%
Title's compound score	38.4%	33.9%	27.7%
Number of exclamation marks in title	15.3%	22.6%	62.1%
Number of question marks in title	2.0%	5.4%	92.6%
Number of full stops in title	14.3%	13.9%	71.8%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	8.8%	13.0%	78.3%
Number of hyphens in title	14.0%	18.0%	67.9%
Number of different tags	38.3%	56.9%	4.8%

Table 11.10: Statistical results for numerical data

Once again, shorter titles are preferred and fewer capitalized words. Punctuation marks exist in a small percentage of samples and it is advised that their number decreases in most cases. The number of tags is also generally smaller in comparison to the number of tags in videos of the test dataset. In every video in our dataset it appears that both ratings and comments are enabled and it is always recommended that they are enabled in the test dataset too. The capitalization of the first letter in each sentence in the title is also common practice for both datasets.

Feature	True->False	False->True	True->True	False->False
Ratings disabled	1.8%	0%	0%	98.2%
Comments disabled	2.6%	0%	0%	97.4%
First letter capitalization in title	15.9%	19.7%	60.3%	4%

Table 11.11: Statistical results for categorical data

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
game	minecraft
video	game
series	battle
man	clash
screenshot	fortnite

Table 11.12: Statistical results for thumbnails and tags

When it comes to thumbnail images, the most popular edits are screenshot images from video games and also depictions of men, possibly the content creators or the avatars the use. Tags containing the words "game" and "battle" are also popular, along with several commercial video games, namely Minecraft, Fortnite. This means that many samples in this category, especially the ones with the most views, concern the most popular video games, which is expected.

11.2.4 People & Blogs

This category encompasses a wide range of content that revolves around personal experiences, storytelling, lifestyle, and community interaction. Since this category has 3769 unique samples, we create a dataset containing the 1000 ones with the most views. Our test dataset is comprised of the 1000 ones with the fewest views. The statistical results from this experiment are shown in the tables.

Feature	Increased	Decreased	Remains the same
Title length	39.0%	52.3%	8.7%
Title's capital words ratio	26.8%	45.6%	27.5%
Title's compound score	40.4%	26.5%	33.0%
Number of exclamation marks in title	13.7%	31.3%	55.0%
Number of question marks in title	3.2%	5.4%	91.4%
Number of full stops in title	11.5%	16.6%	72.0%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	10.7%	11.4%	77.9%
Number of hyphens in title	9.9%	9.1%	81.0%
Number of different tags	30.5%	50.9%	18.6%

Table 11.13: Statistical results for numerical data

Feature	True->False	False->True	True->True	False->False
Ratings disabled	0.3%	2.3%	0.1%	97.3%
Comments disabled	2.3%	1.2%	0%	96.5%
First letter capitalization in title	14.7%	22%	58%	5.4%

Table 11.14: Statistical results for categorical data

Once again shorter titles than those of the testing data are preferred as well as fewer capital words. In addition, the testing data seem to contain more exclamation marks than the samples from our dataset. The number of tags also tends to be smaller for our dataset. Like all other datasets, ratings and comments are enabled. First letter capitalization in the title is also prominent in both datasets, but there exists a small portion (22%) of the testing dataset where this is violated.

The thumbnails of our dataset in contrast to the ones of the testing dataset depict men and women, sitting or holding something, or a picture perhaps of a family or a person on vacation. In other words, the thumbnail images tend to show people in everyday situations. The most popular tag transformation is to tags with family and life themes, vlogs and tags that contain the word funny.

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
man	family
woman	real
holding	life
picture	funny
sitting	vlogs

Table 11.15: Statistical results for thumbnails and tags

11.2.5 Entertainment

This category includes a variety of content aimed at amusing, entertaining, and engaging viewers. This can span multiple subgenres, from comedy and drama to celebrity news. It contains 8539 samples and is the second largest category. We create a dataset using the 2500 samples with the highest view count. We also generate a test dataset consisting of the 2500 samples with the lowest view count. The statistical results from the experiments are shown in the tables.

Apart from the title's length, which our results recommend should be generally shorter than it is in the test samples, there are no other strong general directions in the first two tables. This can be explained by the type of the category we are examining. In this category, since it deals with celebrity news, comedy and drama, videos that get the biggest exposure are either ones that discuss current events and gossip, or are very unique and entertaining in their own way. This means that there isn't really a secret formula that makes them succeed, but really it's most about timing and individuality. Therefore, it is very hard to match non-viral videos with the viral ones in a way that will help increase their popularity.

For this category, the thumbnail images tend to again depict people and people's faces. The tags contain in general the words "funny", "diy" (do it yourself), "tips" etc and these edits are what the counterfactual explanations suggest.

Feature	Increased	Decreased	Remains the same
Title length	36.0%	55.3%	8.6%
Title's capital words ratio	36.2%	37.4%	26.4%
Title's compound score	36.4%	36.8%	26.8%
Number of exclamation marks in title	13.0%	19.4%	67.6%
Number of question marks in title	3.3%	5.8%	90.9%%
Number of full stops in title	9.5%	12.4%	78.1%
Number of quotation marks in title	0%	0%	100%
Number of vertical lines in title	22.1%	19.0%	58.9%%
Number of hyphens in title	14.9%	14.1%	71.0%
Number of different tags	43.2%	50.3%	6.5%

Table 11.16: Statistical results for numerical data

Feature	True->False	False->True	True->True	False->False
Ratings disabled	0.4%	0.4%	0%	99.2%
Comments disabled	1%	1%	0%	97.9%
First letter capitalization in title	14.1%	16%	66%	3.9%

Table 11.17: Statistical results for categorical data

Top 5 thumbnail edits/additions	Top 5 tags edits/additions
man	funny
person	diy
woman	tips
movie	activities
face	challenge

Table 11.18: Statistical results for thumbnails and tags

11.3 Cumulative Statistical Results

Our thorough examination of viral vs low-view YouTube videos in a variety of genres revealed a number of common patterns and traits that set successful material apart. Shorter titles, frequently with fewer capitalized words and little punctuation, are a common feature of viral videos, reflecting a preference for clarity and brevity that likely appeals to a broader audience. Punctuation marks such as exclamation points, question marks, and full stops are used sparingly, and the number of tags tends to be smaller but more focused. Both ratings and comments are consistently enabled in viral videos, suggesting that engagement and interaction are important factors in video success. Additionally, it is common practice to capitalize titles' sentences' first letter, as it adds professionalism and appeal.

Furthermore, the popularity of viral videos is largely dependent on their thumbnails, and one recurring theme

is the portrayal of humans, particularly in everyday or relatable situations. It is possible that this visual approach draws viewers in by creating a sense of curiosity or intimacy. Broad, well-liked subjects like "music," "game," "funny," "family," and "life" are frequently used in successful films' tags because they are likely to pique the interest of a large audience.

These patterns held true for a number of areas, such as Entertainment, People & Blogs, Sports, Music, and Gaming. The dynamic and varied character of music videos is reflected in the larger ratio of capitalized words and more tags in the Music category. Sports films prioritized names that emphasized important terms like "player" and "game," with a higher proportion of capitalized words and fewer tags. Similar trends could be seen in the gaming videos, which mostly featured well-known titles like Fortnite and Minecraft. Thumbnails and tags pertaining to People & Blogs videos tended to favor daily life and family themes. The wide variety of content seen in entertainment videos—from comedy to celebrity news—highlights the value of timing and uniqueness, making it more difficult to identify a single recipe for success.

Overall, our research indicates that, regardless of the content category of a video, carefully considered tag selection, clever title structuring, and captivating thumbnails are essential components that boost video virality. Simplicity, relatability, and engagement are key components of this insightful methodology that helps content creators expand the audience and impact of their videos.

11.4 Qualitative Results

In this section we will present examples of matching between video graphs made by our framework. As we mentioned earlier, there exist some "good" and some "bad" matchings. This distinction is based on their quality or in other words the actual similarity of the two videos and the value of their comparison. We will explore one "good" example and one "bad" example.

11.4.1 Example of a "good" matching

We will first explore a "good" case of video comparison. Table 11.19 contains the characteristics of a video included in the test dataset from the experiment carried out for the category "People & Blogs". These are its features exactly as the "YouTube Trending Video Dataset" incorporates them in its columns. Table 11.20 contains the characteristics of the video included in our dataset from the experiment carried out for the category "People & Blogs" with which the previous video is matched by our framework. Images 11.4.1, 11.4.2 depict the knowledge graphs constructed by our framework for these two videos.

Simply by looking at the characteristics of these two videos it is quite easy to pick up on why they were matched by our framework and why we would consider this match a "good" one. More specifically, the videos share the same exact theme: A new baby's gender reveal. Their difference in views is large, with the test video having 462.406 total views, while the video from our dataset has 2.194.817 views.

video_id	**title**	**publishedAt**	**channelId**
s5xEL3UTxE0	Finding Out The Gender Of BABY #2!	2022-09-30T23:16:33Z	UCtPqR1gumPRaMLh STlLn5nw
channelTitle	**categoryId**	**trending_date**	**tags**
The Herberts	22	2022-10-05T00:00:00Z	[None]
view_count	**likes**	**dislikes**	**comment_count**
462406	13618	0	695
thumbnail_link	**comments_disabled**	**ratings_disabled**	**description**
https://i.ytimg.com/vi/s5xEL3UTxE0/default.jpg	False	False	We are so excited! Come along with us to find out the gender of baby #2!

Table 11.19: Metadata for the video from the test data

video_id	title	publishedAt	channelId
J1TshEz8yq4	Finding out the gender of our baby	2023-04-21T16:00:17Z	UC3zZX4ttC52HlqVis7eQ-Hg
channelTitle	**categoryId**	**trending_date**	**tags**
Matt & Abby	22	2023-04-27T00:00:00Z	finding out the gender|gender reveal|pregnancy|doctors appointment|vlog|finding out the gender of our baby vlog|family|pregnant| surprise pregnancy|matt howard|abby howard|matt and abby
view_count	**likes**	**dislikes**	**comment_count**
2194817	73112	0	4456
thumbnail_link	**comments_disabled**	**ratings_disabled**	**description**
https://i.ytimg.com/vi/J1TshEz8yq4/default.jpg	False	False	Follow us on Instagram: @abbyelizabethoward @_matt_howard_

Table 11.20: Metadata for the video from our data

As we explained before, we do not compare entire graphs with one another, but rather we break them up into 3 smaller graphs: one for the title, one for the thumbnail and one for the tags. Afterwards, these graphs are turned into 3 separate queries. The queries created from our graphs are the following:

For the video from the test data:

title: ['finding', 'gender', 'baby', '2']

thumbnail: ['hand', 'woman', 'sitting', 'hospital', 'bed', 'tv']

tags: []

And for its match from our dataset:

title: ['finding', 'gender', 'baby']

thumbnail: ['book', 'couple', 'car']

tags: ['finding', 'gender', 'pregnancy', 'matt']

To calculate the edit distance between the title queries, we notice that one is a subset of the other. In other words, all that needs to be done is remove the word '2' from the test title. The test tags are an empty query, therefore all words from the match should be added. For the thumbnail queries our framework calculates the four words with the shorter semantic distance from the four words that the match thumbnail contains and the cost to delete the rest. Adding the costs calculated (with the appropriate scaling explained before), we get the total cost of this transformation, which was the minimum from the entire dataset. After uncovering this minimum cost transformation, our framework outputs the suggestions for the test video to elevate its view count:

Features advised to change	Initial value	Suggested final value
number_of_different_tags	0	12
cap_words_ratio_is	1.0	0.0
number_of_exclamationmark	1	0
title keyword	'2'	-
thumbnail depiction	"woman"	"couple"
thumbnail depiction	"tv"	"book"
thumbnail depiction	"hand"	"car"
thumbnail depiction	"woman"	"couple"
thumbnail depictions	"bed"	-
thumbnail depictions	"hospital"	-
thumbnail depictions	"sitting"	-
tag	-	finding
tag	-	gender
tag	-	pregnancy
tag	-	matt

Table 11.21: Suggested changes

Based on the above table, it is suggested that the thumbnail picture is altered to depict a couple with a book in a car, instead of a woman in a hospital bed. It is also suggested that 12 different tags be added, with themes surrounding pregnancy, finding the gender and matt. Last but not least, it is recommended that all capital words in the title become lowercase words and that the exclamation mark be deleted.

Our framework was evidently successful in pairing up these two videos and comparing their differences. The owner of the test video could benefit from the suggestions made and modify their video's characteristics in order to increase its chances of becoming viral. Clearly, not all of these suggestions are valuable and some may not even affect the video in any significant way. This is for the user to decide while operating with critical thinking and staying true to what their video and channel represent.

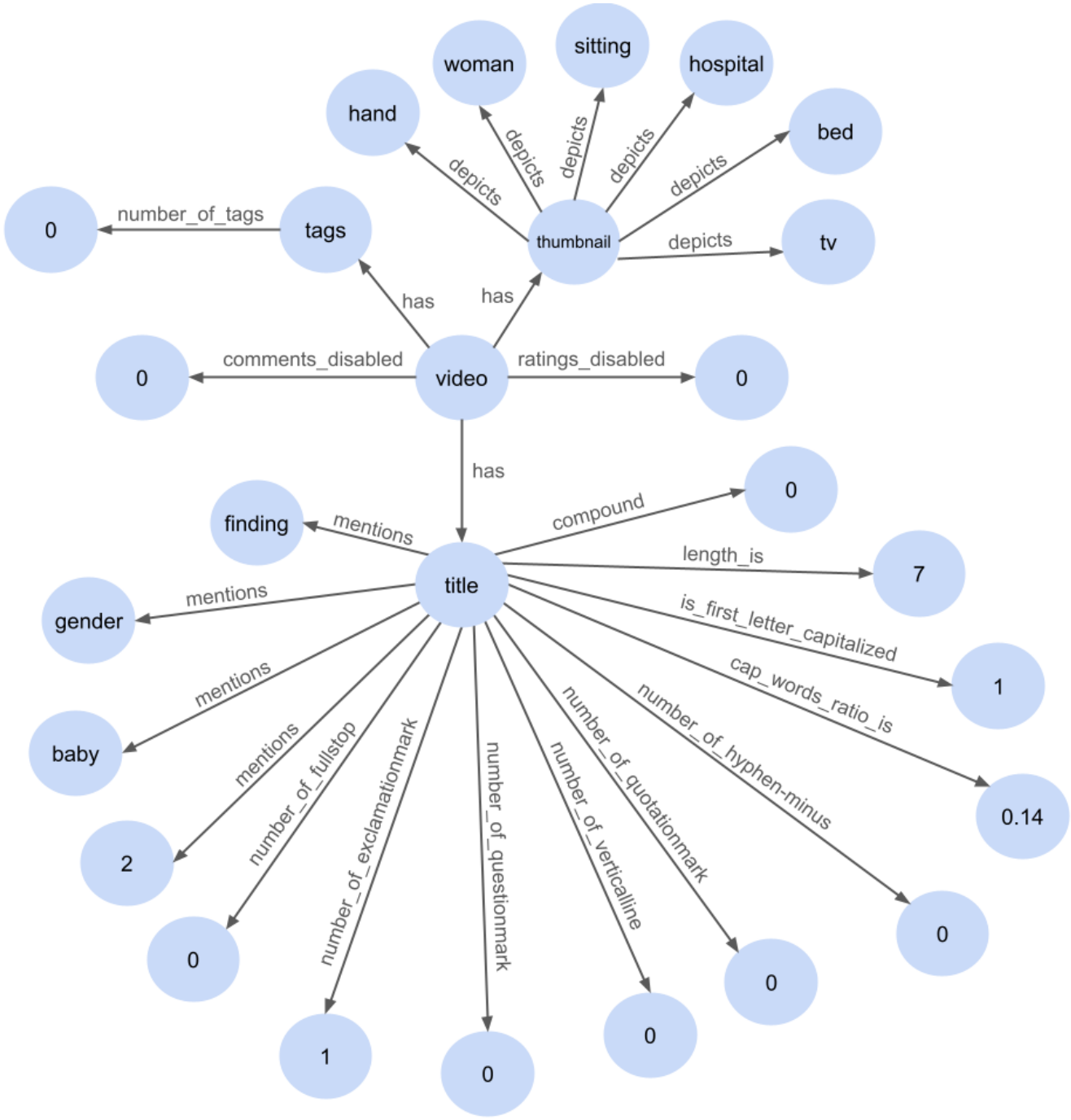


Figure 11.4.1: The complete graph of the video from the test dataset

Figure 11.4.2: The complete graph of the video from our dataset

11.4.2 Example of a "bad" matching

We will now explore a "bad" case of video matching. Table 11.22 contains the characteristics of a video included in the test dataset from the experiment carried out for the category "People & Blogs". These are its features exactly as the "YouTube Trending Video Dataset" incorporates them in its columns. Table 11.23 contains the characteristics of the video included in our dataset from the experiment carried out for the category "People & Blogs" with which the previous video is matched by our framework. Images 11.4.3, 11.4.4 depict the knowledge graphs constructed by our framework for these two videos.

video_id	title	publishedAt	channelId
v3MkXz2S5Ww	Building a 5 Acre Pond! (Baby Ducks Hatched)	2022-05-29T23:31:11Z	UC8ha6SsRNvDGkwcPTCXkW3g
channelTitle	**categoryId**	**trending_date**	**tags**
BamaBass	22	2022-06-04T00:00:00Z	[None]
view_count	**likes**	**dislikes**	**comment_count**
420151	22920	0	1329
thumbnail_link	**comments_disabled**	**ratings_disabled**	**description**
https://i.ytimg.com/vi/v3MkXz2S5Ww/default.jpg	False	False	Building a 5 acre pond for our two pet bass! Subscribe for weekly pond build videos: http://bit.ly/Bama_Bass Check out the Butcherbox grilling bundle here: https://bchrbox.co/BamabassGrillBBBuilding a 5 acre pond for our two pet bass! Day 1...

Table 11.22: Metadata for the video from the test data

video_id	title	publishedAt	channelId
SXQ90lqYEuI	WHAT'S THE CAPITAL OF USA?!	2021-12-26T18:36:15Z	UCWeqRSPFvAxebBj24PpDbaw
channelTitle	**categoryId**	**trending_date**	**tags**
Cloutom	22	2022-01-04T00:00:00Z	[None]
view_count	**likes**	**dislikes**	**comment_count**
1895054	28647	0	663
thumbnail_link	**comments_disabled**	**ratings_disabled**	**description**
https://i.ytimg.com/vi/SXQ90lqYEuI/default.jpg	False	False	NaN

Table 11.23: Metadata for the video from our data

A simple and quick overview of the two tables indicates that this is not a good matching. From the title alone it is obvious that the two videos are very different thematically, as one discusses the creation of a pond for some duck hatchlings, while the other presents itself with the simple question of what the capital of the USA is. By inspecting further and searching the videos on YouTube, we realize that the video with the question is actually a humoristic video of a public survey in a mall where an interviewer asks passersby to answer this question. On the other hand, the test video tells the story of how a five-acre pond was created for ducks. At first glance, this matching puzzles us, however, by examining the graphs corresponding to these videos, while also considering our knowledge of our dataset's characteristics, the reasons behind it become apparent. To begin with, the duck video is extremely specific and through a keyword search in the dataset we built for

the category "People & Blogs" where the experiments took place, we realized that there are no other videos concerning ducks, or ponds. Therefore, our framework had to look further away for a match. The video with the interviewer asking people what the capital of USA is has no tags, just like the duck pond video. Moreover, they both have 3 thumbnail keywords in total, which means there will not be any extra cost to add or subtract a keyword, only to transform them. Additionally, both videos' titles contain exactly one exclamation mark and no vertical lines, quotation marks or hyphens. Their compound scores are also equal to each other. All in all, while these two videos do not actually resemble one another, the cost to transform one into the other is lower than other videos, especially since the test video is quite specific. If our dataset was expanded, more similar videos would be included and cases like this one could get eliminated.

Figure 11.4.3: The complete graph of the video from the test dataset

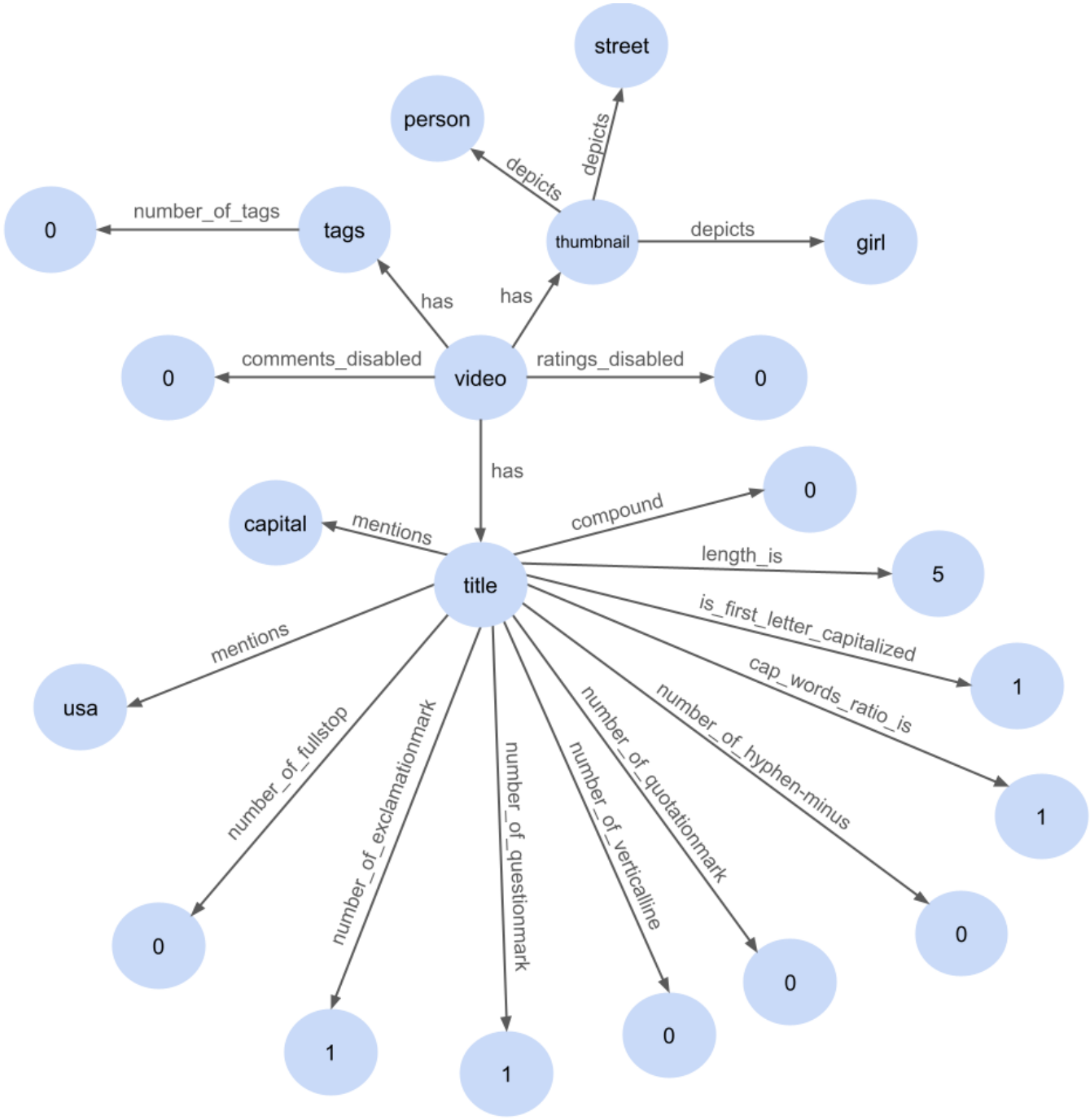


Figure 11.4.4: The complete graph of the video from our dataset

Features advised to change	**Initial value**	**Suggested final value**
cap_words_ratio_is	0.0	1.0
length_is	8	5
number_of_questionmark	0	1
title keyword	'building'	'capital'
title keyword	'acre'	'usa'
title keyword	'baby'	-
title keyword	'hatched'	-
title keyword	'5'	-
title keyword	'pond'	-
title keyword	'ducks'	-
thumbnail depiction	"view"	"person"
thumbnail depiction	"lake"	"street"
thumbnail depiction	"aerial"	"girl"

Table 11.24: Suggested changes

In the table above, the changes suggested by our framework are visible. It is obvious that these changes have no value to us, apart from participating in gathering accumulative statistical results regarding the differences between viral and non-viral videos.

Chapter 12

Conclusion

12.1 Discussion

This thesis explores the intricate mechanisms that influence video virality on the YouTube platform. We have leveraged advanced analytical tools, such as graph theory, sentiment analysis and counterfactual explanations and identified key characteristics that increase the chances for video success. More specifically, our research highlights the importance of concise and compelling titles, thumbnails depicting people and focused tags. Giving the ability to users to comment and rate videos has also been shown to enhance viewer engagement and video popularity.

We conducted a series of general and category-specific experiments to uncover distance patterns that set highly viewed videos apart from those with fewer views. Specifically, in the category titled "Music" videos with more tags and capitalized words tend to perform better. In the category titled "Sports" tags tend to be more precise and fewer. In the category titled "Gaming" popular video game names took over the video tags and their thumbnails frequently portray screenshots from said games. Videos found in the "People & Blogs" category are inclined to showcase everyday situations in their thumbnails and have tags related to family and life themes. Videos in the "Entertainment" category, have diverse content and highlight the value of timing and uniqueness in achieving virality.

Our research underscores how critical the use of thumbnails and tags is in order to attract viewers. More specifically, it was established that thumbnails showing people, particularly in relatable situations, tend to draw more attention. Moreover, popular and general tags appeal to a wide audience, thus increasing the chances that a video will be shared and recommended.

While this study offers valuable insights, it is not without limitations. Developing and running our code in Google Collab entails that we did not have the ability to run experiments with datasets larger than 5000 samples, as the results would take too long to be produced. Moreover, our dataset only contained data about the videos themselves and not about concurrent events, which could help pinpoint the reasons that a specific video was made popular at a specific time. Lastly, we limited our study to data from the USA.

In conclusion, this thesis contributes to a deeper understanding of what makes a video go viral. The insights gained from our study interest not only academic researchers but also have practical implications for anyone involved in digital content creation and marketing, who can enhance their ability to produce highly viewed videos by considering our findings and recommendations. As the digital media landscape continues to evolve, the principles outlined in this research will remain critical for achieving success in the competitive world of YouTube videos.

12.2 Future Work

In closing this thesis, we would like to give a few suggestions for future improvements on our work or different research pathways. To begin with, the analysis of the thumbnail images of all YouTube videos can be greatly

improved. More specifically, instead of simply generating a simple caption describing the image, a lot can be done regarding color analysis, object detection or even sentiment analysis of the faces of humans present in the picture. It would be of great interest to examine whether or not such details have an actual impact on virality, as one can intuitively think that brighter colors could get viewers' attention more easily and fast, or that shocked expressions of people in the image intrigue viewers, compelling them to explore certain videos. Another way images can be incorporated in this field would be to not only include the thumbnail image, but also screenshots from the entirety of the video, especially the first seconds. The analysis of said screenshots could give many interesting details. For instance, if the video consists of a person talking about a certain subject, the analysis of their expression throughout their speech might give us hints about why this particular video achieved such high popularity. In other words, whether a person is enthusiastic and lively throughout the video, or calm and reserved could indicate why they accumulated a large number of views or not, as the energy of a person speaking certainly impacts the way the speech is perceived. Another path our research can be expanded towards is taking into account real-world events taking place at approximately the same time as the upload date of videos. This is of great significance as, for instance, the airing of a popular TV series episode featuring a specific song as a soundtrack can cause an increase in the number of views of the official video of said song on YouTube. Other examples of the influence of real-world events on YouTube video popularity are the many cases of celebrity news videos. The popularity of such videos does not source simply from the video itself, but also from the popularity and relevance of the celebrities they discuss. In addition, timing plays a big role in such scenarios as, in most cases, the videos that get uploaded first get the lion's share of the views. Therefore, not only should major social events be considered, but also the proximity of the videos to them is quite important. Lastly, our research was limited to examining data only from the USA. In the future, data from other countries should be examined as well.

Chapter 13

Bibliography

[1] Advisor, T. R. *How Many Views is Viral? The Algorithm Factor*. Accessed: 2024-06-24. n.d. URL:

[2] Auer, S. et al. "Dbpedia: A nucleus for a web of open data". In: *international semantic web conference*. Springer. 2007, pp. 722–735.

[3] Babai, L. "Graph isomorphism in quasipolynomial time". In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, pp. 684–697.

[4] Bačić, D. and Gilstrap, C. "Predicting video virality and viewer engagement: a biometric data and machine learning approach". In: *Behaviour & Information Technology* (2023), pp. 1–27.

[5] Backstrom, L. et al. "Group formation in large social networks: membership, growth, and evolution". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 44–54.

[6] Bai, Y. et al. "Simgnn: A neural network approach to fast graph similarity computation". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019, pp. 384–392.

[7] Berberich, K. et al. "A language modeling approach for temporal information needs". In: *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings 32*. Springer. 2010, pp. 13–25.

[8] Bertsekas, D. P. "The auction algorithm: A distributed relaxation method for the assignment problem". In: *Annals of Operations Research* 14.1 (1988), pp. 105–123. DOI: 10.1007/BF02288302.

[9] Bickle, A. *Fundamentals of Graph Theory*. Pure and Applied Undergraduate Texts. American Mathematical Society, 2020.

[10] Bielski, A. and Trzcinski, T. "Understanding multimodal popularity prediction of social media videos with self-attention". In: *IEEE Access* 6 (2018), pp. 74277–74287.

[11] Bollacker, K. et al. "Freebase: a collaboratively created graph database for structuring human knowledge". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 1247–1250.

[12] Bradley, M. M. and Lang, P. J. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology ..., 1999.

[13] Broxton, T. et al. "Catching a viral video". In: *Journal of Intelligent Information Systems* 40 (2013), pp. 241–259.

[14] Cambria, E. et al. "New Avenues in Opinion Mining and Sentiment Analysis". In: *IEEE Intelligent Systems* 28.2 (2013), pp. 15–21.

[15] Chen, J. et al. "Micro tells macro: Predicting the popularity of micro-videos via a transductive model". In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016, pp. 898–907.

[16] Chung, F. R. *Spectral graph theory*. Vol. 92. American Mathematical Soc., 1997.

[17] Collins, J. *History of YouTube: Timeline, Growth & Future*. Accessed: 2024-06-24. 2024. URL:

[18] Dantzig, G. B. "Maximization of a linear function of variables subject to linear inequalities". In: *Activity Analysis of Production and Allocation*. Ed. by T. C. Koopmans. New York: Wiley, 1951, pp. 339–347.

[19] Dervakos, E., Filandrianos, G., and Stamou, G. "Heuristics for evaluation of AI generated music". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9164–9171.

[20] Dervakos, E. et al. "Choose your data wisely: a framework for semantic counterfactuals". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023, pp. 382–390.
[21] Dervakosa, E. et al. "Semantic Enrichment of Pretrained Embedding Output for Unsupervised IR". In: *AAAI-MAKE* (2021).
[22] Devlin, J. et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).
[23] Deza, A. and Parikh, D. "Understanding image virality". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1818–1826.
[24] Dimitriou, A. et al. "Structure Your Data: Towards Semantic Graph Counterfactuals". In: *Forty-first International Conference on Machine Learning*. 2024.
[25] Dubey, A. and Agarwal, S. "Modeling image virality with pairwise spatial transformer networks". In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 663–671.
[26] Edition, T. "Reinhard Diestel". In: ().
[27] Egami, S. et al. "RDF-star2Vec: RDF-star Graph Embeddings for Data Mining". In: *IEEE Access* 11 (2023), pp. 142030–142042. ISSN: 2169-3536.
[28] Ehrlinger, L. and Wöß, W. "Towards a defnition of knowledge graphs". In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4 (2016), p. 2.
[29] Filandrianos, G. et al. "Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9507–9525. DOI: 10.18653/v1/2023.findings-acl.606. URL:
[30] Filandrianos, G. et al. "Conceptual Edits as Counterfactual Explanations." In: *AAAI Spring Symposium: MAKE*. 2022.
[31] Fischer, A. et al. "Approximation of graph edit distance based on Hausdorff matching". In: *Pattern Recognition* 48.2 (2015), pp. 331–343.
[32] Gao, X. et al. "A survey of graph edit distance". In: *Pattern Analysis and applications* 13 (2010), pp. 113–129.
[33] Gelling, E., Fletcher, G., and Schmidt, M. "Bridging graph data models: RDF, RDF-star, and property graphs as directed acyclic graphs". In: *arXiv preprint arXiv:2304.13097* (2023).
[34] Ghosh, A. and Kandasamy, D. "Interpretable artificial intelligence: why and when". In: *American Journal of Roentgenology* 214.5 (2020), pp. 1137–1138.
[35] Giadikiaroglou, P. et al. "Puzzle Solving using Reasoning of Large Language Models: A Survey". In: *arXiv preprint arXiv:2402.11291* (2024).
[36] Grath, R. M. et al. "Interpretable credit application predictions with counterfactual explanations". In: *arXiv preprint arXiv:1811.05245* (2018).
[37] Grigoriadou, N. et al. "AILS-NTUA at SemEval-2024 Task 6: Efficient model tuning for hallucination detection and analysis". In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by A. K. Ojha et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1549–1560. URL:
[38] Grover, A. and Leskovec, J. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.
[39] Harbarex. *Predicting the virality of TikTok videos*. GitHub repository. 2023.
[40] Hochreiter, S. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.
[41] Hogan, A. et al. "Knowledge Graphs". In: *ACM Computing Surveys* 54.4 (2021).
[42] Hosch, W. L. *YouTube | History, Founders, & Facts*. Accessed: 2024-06-24. 2024. URL:
[43] Hussein, R., Yang, D., and Cudré-Mauroux, P. "Are meta-paths necessary? Revisiting heterogeneous graph embeddings". In: *Proceedings of the 27th ACM international conference on information and knowledge management*. 2018, pp. 437–446.
[44] Hutto, C. and Gilbert, E. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[45] Jiang, L. et al. “Viral video style: A closer look at viral videos on youtube”. In: *Proceedings of International Conference on Multimedia Retrieval*. 2014, pp. 193–200.
[46] Karp, R. “An algorithm to solve the m×n assignment problem in expected time O(mnlogn)”. In: *Networks* 10 (1980), pp. 143–152.
[47] Kazani, A. et al. “Semantic Integration of Data: From Theory to Social Research Practice”. In: *Quantitative Demography and Health Estimates: Healthy Life Expectancy, Templates for Direct Estimates from Life Tables and other Applications*. Springer, 2023, pp. 303–314.
[48] King, J. et al. “Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions”. In: *AWS Whitepaper* (2021).
[49] Kipf, T. N. and Welling, M. “Semi-supervised classification with graph convolutional networks”. In: *arXiv preprint arXiv:1609.02907* (2016).
[50] Ko, H. et al. “Machine learning and knowledge graph based design rule construction for additive manufacturing”. In: *Additive Manufacturing* 37 (2021), p. 101620.
[51] Kong, Q. et al. “Will this video go viral: Explaining and predicting the popularity of youtube videos”. In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 175–178.
[52] Kuhn, H. W. “The Hungarian method for the assignment problem”. In: *Naval Research Logistics Quarterly* 2.1–2 (1955), pp. 83–97.
[53] Li, J. et al. “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation”. In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.
[54] Li, X. and Li, J. “Angle-optimized text embeddings”. In: *arXiv preprint arXiv:2309.12871* (2023).
[55] Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2020.
[56] Lymperaiou, M. et al. “Towards Explainable Evaluation of Language Models on the Semantic Similarity of Visual Concepts”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3639–3658. URL:
[57] Manola, F., Miller, E., McBride, B., et al. “RDF primer”. In: *W3C recommendation* 10.1-107 (2004), p. 6.
[58] Meghawat, M. et al. “A multimodal approach to predict social media popularity”. In: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE. 2018, pp. 190–195.
[59] Mikolov, T. et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
[60] Miller, G. A. et al. “Introduction to WordNet: An on-line lexical database”. In: *International journal of lexicography* 3.4 (1990), pp. 235–244.
[61] Mishra, S. et al. “Image Caption Generation using Vision Transformer and GPT Architecture”. In: *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE. 2024, pp. 1–6.
[62] Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL:
[63] Mondal, B. and De, K. “An Overview Applications of Graph Theory in Real Field”. In: *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 2 (2017).
[64] MonkeyLearn. *Sentiment Analysis: A Definitive Guide*. Accessed: 2024-06-08. 2022. URL:
[65] Munkres, J. “Algorithms for the assignment and transportation problems”. In: *Journal of the Society for Industrial and Applied Mathematics* 5.1 (1957), pp. 32–38.
[66] Neuhaus, M., Riesen, K., and Bunke, H. “Fast suboptimal algorithms for the computation of graph edit distance”. In: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer. 2006, pp. 163–172.
[67] Noy, N. et al. “Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it’s done”. In: *Queue* 17.2 (Apr. 2019), pp. 48–75.
[68] Panagiotopoulos, I. et al. “AILS-NTUA at SemEval-2024 Task 9: Cracking Brain Teasers: Transformer Models for Lateral Thinking Puzzles”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by A. K. Ojha et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1733–1746. URL:
[69] Papadimitriou, C. et al. “Masked Generative Story Transformer with Character Guidance and Caption Augmentation”. In: *arXiv preprint arXiv:2403.08502* (2024).

[70] Parsanoglou, D. et al. *Combining innovative methodological tools to approach digital transformations in leisure among children and young people*. Tech. rep. DigiGen Working Paper Series, 2022.

[71] Peng, C. et al. "Knowledge Graphs: Opportunities and Challenges". In: *Artificial Intelligence Review* 56.1 (2023), pp. 13071–13102.

[72] Pennebaker, J. W., Francis, M. E., and Booth, R. J. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[73] Pennington, J., Socher, R., and Manning, C. D. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[74] Pinto, H., Almeida, J. M., and Gonçalves, M. A. "Using early view patterns to predict the popularity of youtube videos". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 365–374.

[75] Raymond, J. W., Gardiner, E. J., and Willett, P. "Rascal: Calculation of graph similarity using maximum common edge subgraphs". In: *The Computer Journal* 45.6 (2002), pp. 631–644.

[76] Resnik, P. "Using information content to evaluate semantic similarity in a taxonomy". In: *arXiv preprint cmp-lg/9511007* (1995).

[77] Sanfeliu, A. and Fu, K.-S. "A distance measure between attributed relational graphs for pattern recognition". In: *IEEE transactions on systems, man, and cybernetics* 3 (1983), pp. 353–362.

[78] Sharan, R. et al. "Conserved patterns of protein interaction in multiple species". In: *Proceedings of the National Academy of Sciences* 102.6 (2005), pp. 1974–1979.

[79] Sharma, H. D. and Goyal, P. "An Analysis of Sentiment: Methods, Applications, and Challenges". In: *Engineering Proceedings* 59.1 (2023).

[80] Sharma, P. et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2556–2565.

[81] Stokowiec, W. et al. "Shallow reading with deep learning: Predicting popularity of online content using only its title". In: *Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings 23*. Springer. 2017, pp. 136–145.

[82] Stone, P. J., Dunphy, D. C., and Smith, M. S. "The general inquirer: A computer approach to content analysis." In: (1966).

[83] Sun, R. et al. "Multi-modal knowledge graphs for recommender systems". In: *Proceedings of the 29th ACM international conference on information & knowledge management*. 2020, pp. 1405–1414.

[84] Symeonaki, M., Filandrianos, G., and Stamou, G. "Visualising key information and communication technologies (ICT) indicators for children and young individuals in Europe". In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–12.

[85] Techopedia. *What is Viral Content? - Definition from Techopedia*. Accessed: 2024-06-24. n.d. URL:

[86] Trzciński, T. and Rokita, P. "Predicting popularity of online videos using support vector regression". In: *IEEE Transactions on Multimedia* 19.11 (2017), pp. 2561–2570.

[87] Ugander, J. et al. "The anatomy of the facebook social graph". In: *arXiv preprint arXiv:1111.4503* (2011).

[88] Ullman, J. R. et al. "An algorithm for subgraph isomorphism". In: *Journal of the ACM* 23.1 (1976), pp. 31–42.

[89] Vallet, D. et al. "Characterizing and predicting viral-and-popular video content". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 2015, pp. 1591–1600.

[90] Verma, S., Dickerson, J., and Hines, K. "Counterfactual explanations for machine learning: Challenges revisited". In: *arXiv preprint arXiv:2106.07756* (2021).

[91] Vinyals, O. et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

[92] Vishwanathan, S. V. N. et al. "Graph kernels". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1201–1242.

[93] Wang, J. et al. "Will you ever become popular? Learning to predict virality of dance clips". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18.2 (2022), pp. 1–24.

[94] Wang, J. et al. "Git: A generative image-to-text transformer for vision and language". In: *arXiv preprint arXiv:2205.14100* (2022).
[95] Wang, X. et al. "Heterogeneous graph attention network". In: *The world wide web conference*. 2019, pp. 2022–2032.
[96] West, T. "Going viral: Factors that lead videos to become internet phenomena". In: *The Elon Journal of Undergraduate Research in Communications* 2.1 (2011), pp. 76–84.
[97] Wikipedia contributors. *Knowledge graph — Wikipedia, The Free Encyclopedia*. [Online; accessed 8-June-2024]. 2024. URL:
[98] Wikipedia contributors. *Viral video — Wikipedia, The Free Encyclopedia*. [Online; accessed 8-June-2024]. 2024. URL:
[99] Xu, J. *Theory and Application of Graphs*. Vol. 10. Network Theory and Applications. Springer US, 2013.
[100] Yang, L. et al. "Dynamic heterogeneous graph embedding using hierarchical attentions". In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer. 2020, pp. 425–432.
[101] YouTube, H. of. *YouTube - Wikipedia*. Accessed: 2024-06-24. 2024. URL:
[102] Zarkogianni, K., Dervakos, E., Filandrianos, G., et al. "The smarty4covid dataset and knowledge base as a framework for interpretable physiological audio data analysis". In: *Sci Data* 10 (2023), p. 770. DOI: 10.1038/s41597-023-02646-6. URL:
[103] Zhang, W. et al. "User-guided hierarchical attention network for multi-modal social image popularity prediction". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 1277–1286.